2021

Investigation Of Rna-Protein Interactions In Prc2 Function

Robert Warneford-Thomson
University of Pennsylvania

Investigation Of Rna-Protein Interactions In Prc2 Function

Abstract

Chromatin regulation contributes to control of gene expression and what identity a cell will adopt. In the last decade the role that RNA plays in chromatin regulation has become increasingly clear. RNA mediates protein recruitment and eviction from chromatin, forms nuclear condensates with proteins and DNA, and contributes to proper chromatin organization. Yet our knowledge of the mechanisms that govern RNA activity on chromatin lags significantly and limits our ability to understand nuclear function. To effectively answer some of the questions of RNA function in the nucleus we need a comprehensive atlas of RNA-protein interactions, which would enable generation of protein mutants defective in RNA-binding. The goal of my thesis was to develop an unbiased method to profile RNA-binding proteins in the nucleus and apply it to Polycomb repressive complex 2 (PRC2). PRC2 is an epigenetic regulatory complex that deposits mono, di- and tri- methyl lysine onto histone H3 (H3K27me3) and maintains gene silencing during development. PRC2 shows extensive contacts with RNA but their function remains unclear. In the first chapter, we present a novel method, dubbed RBR-ID, for the identification of RNA-protein interactions, which usees UV-crosslinking of photosensitive nucleotide analogs to proteins followed by high resolution mass spectrometry (LC- MS/MS). We identified over 800 RNA-binding proteins, of which 427 were novel and enriched for chromatin-related functions. In the second chapter we adapted RBR-ID to study PRC2, identifying RNA-binding-regions (RBRs) on every subunit of the complex. An RBR identified on EED fell near the regulatory center of PRC2, and we showed that RNA-mediated inhibition of PRC2 can be reversed by stimulatory peptides that bind in the regulatory center, reflecting the antagonistic relationship between RNA and PRC2. In the final chapter we present a testing method we developed for the SARS-CoV-2 virus. Our method, COV-ID, uses reverse transcription and loop-mediated isothermal amplification (RT-LAMP) from patient saliva paired with high-throughput sequencing. Using this method we can detect as little as 5-10 SARS-CoV-2 virions/µL, and we successfully replicate classification of saliva samples (10/10) from clinical COVID-19 patients. We show that COV-ID can be multiplexed to detect influenza as well as SARS-CoV-2. Finally we demonstrate thatCOV-ID can process saliva samples collected on filter paper with sensitivity as low as 50 virions/µL.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Biochemistry & Molecular Biophysics

First Advisor
Roberto Bonasio

Second Advisor
Kristen W. Lynch

Keywords
Polycomb, Protein-RNA interactions, SARS-CoV-2 Diagnostics

Subject Categories
Biochemistry | Genetics | Molecular Biology

INVESTIGATION OF RNA-PROTEIN INTERACTIONS IN PRC2 FUNCTION

Robert R. Warneford-Thomson

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

____

Roberto Bonasio, Ph.D., Associate Professor of Cell and Developmental Biology

Graduate Group Chairperson

____

Kim A. Sharp, Ph.D., Associate Professor of Biochemistry and Biophysics

Dissertation Committee

Kristen W. Lynch, Ph.D., Benjamin Rush Professor of Biochemistry (Chair)

John Isaac Murray, Ph.D., Associate Professor of Genetics

Kavitha Sarma, Assistant Professor, The Wistar Institute Cancer Center

Richard Jenner, M.A., Ph.D., Reader in Molecular Biology, UCL Cancer Institute

INVESTIGATION OF RNA-PROTEIN INTERACTIONS IN PRC2 FUNCTION

DEDICATION

To my loving wife, Cecilia...

ACKNOWLEDGEMENT

This work was supported by NIH training grant T32-GM008216.

ABSTRACT

INVESTIGATION OF RNA-PROTEIN INTERACTIONS IN PRC2 FUNCTION

Robert R. Warneford-Thomson

Roberto Bonasio

Chromatin regulation contributes to control of gene expression and what identity a cell will adopt. In the last decade the role that RNA plays in chromatin regulation has become increasingly clear. RNA mediates protein recruitment and eviction from chromatin, forms nuclear condensates with proteins and DNA, and contributes to proper chromatin organization. Yet our knowledge of the mechanisms that govern RNA activity on chromatin lags significantly and limits our ability to understand nuclear function. To effectively answer some of the questions of RNA function in the nucleus we need a comprehensive atlas of RNA-protein interactions, which would enable generation of protein mutants defective in RNA-binding. The goal of my thesis was to develop an unbiased method to profile RNA-binding proteins in the nucleus and apply it to Polycomb repressive complex 2 (PRC2). PRC2 is an epigenetic regulatory complex that deposits mono, di- and tri- methyl lysine onto histone H3 (H3K27me3) and maintains gene silencing during development. PRC2 shows extensive contacts with RNA but their function remains unclear. In the first chapter, we present a novel method, dubbed RBR-ID, for the identification of RNA-protein interactions, which usees UV-crosslinking of photosensitive nucleotide analogs to proteins followed by high resolution mass spectrometry (LC-MS/MS). We identified over 800 RNA-binding proteins, of which 427 were novel and enriched for chromatin-related functions. In the second chapter we adapted RBR-ID to study PRC2, identifying RNA-binding-regions (RBRs) on every subunit of the complex. An RBR identified on EED fell near the regulatory center of PRC2, and we showed that RNA-mediated inhibition of PRC2 can be reversed by stimulatory peptides that bind in the regulatory center, reflecting the antagonistic relationship between RNA and PRC2. In the final chapter we present a testing method we developed for the SARS-CoV-2 virus. Our method, COV-ID, uses reverse transcription and loop-mediated isothermal amplification (RT-LAMP) from patient saliva paired with high-throughput sequencing. Using this method we can detect as little as 5-10 SARS-CoV-2 virions/µL, and we successfully replicate classification of saliva samples (10/10) from clinical COVID-19 patients. We show that COV-ID can be multiplexed to detect influenza as well as SARS-CoV-2. Finally we demonstrate that COV-ID can process saliva samples collected on filter paper with sensitivity as low as 50 virions/µL.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF DIGITAL SUPPLEMENTAL TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 RNA-protein interactions and chromatin

1.1.1 Introduction to chromatin

Chromatin describes the complement of DNA, RNA, and associated proteins that are packaged inside the nucleus of the cell. The ordered packaging of DNA with protein components dictates the arrangement of genetic material in the nucleus and dynamically changes to allow RNA transcription. These complexes therefore serve both a structural and functional purpose. Chromatin is remodeled and chemically modified by other proteins to change the pattern of gene expression, and changes in chromatin state allow a single genome to sustain the many cell types and developmental stages required for multicellular life.

The basic unit of chromatin structure is the nucleosome, an octamer of histone proteins around which 147 bp of DNA is wrapped. The four core histone proteins (H3, H4, H2A, H2B) that make up the nucleosome can be chemically modified or substituted with different histone variants in ways that frequently correlate with different nuclear processes such as gene transcription or silencing. Based on these connections, a 'histone code' hypothesis emerged that proposed that combinations of histone modifications convey information that can be read by other proteins to carry out their activities [1]. An example of this is the tri-methylation of histone H3 at lysine 27 (H3K27me3) found in transcriptionally silent heterochromatin. In mammals *Polycomb* repressive complex (PRC2) can deposit H3K27me3, and can also recognize its catalytic output in a manner that allosterically stimulates the complex (Figure 1.1). This read-write activity is required for maintenance of H3K27me3 domains across cell divisions, qualifying the mark as a true epigenetic modification [2]. However, in the time since the 'histone code' hypothesis was first proposed, subsequent research has found no evidence that distinct combinations of histone modifications directly cause different chromatin states [3, 4]. Additionally, the majority of histone modifications that associate with different types of transcriptional activity, such as H4K16ac found in transcriptionally active chromatin, do not persist across cell divisions or display any type of memory [2]. Therefore the 'histone code' has limited use in describing epigenetic inheritance, but instead can serve as a means to understand dynamic chromatin regulation.

1.1.2 Chromatin and transcription regulation

Many changes on chromatin relate to transcriptional regulation, and affect gene expression in ways that define cell type and developmental state. Much of the effort in chromatin biology has been directed towards uncovering the relationship between chromatin structure and transcription, which operates at several scales. In the most narrow context, chemical modifications on DNA and histones can act to both inhibit (DNA methylation, H3K9me3, H3K27me3) or stimulate transcription (H3K4me3, H3K27ac). Moving up to the gene scale, cis-regulatory DNA sequences can form loops with gene promoters to activate transcription. At the scale of chromosomes, megabase (10^8 nucleotide) regions of DNA are folded with architectural proteins into spatially constricted topologically associated domains (TADs) [5] and active or inactive compartments [6]. These mechanisms are but some of the examples of the multiple levels of chromatin structure, and indicate the scope of the challenge in understanding transcriptional regulation.

Transcription itself involves the synthesis of RNA from a DNA template via a tightly-regulated process involving conserved molecular machinery. Transcribed RNA is further spliced and processed by additional chromatin-associated protein complexes, such as the CPSF complex that cleaves elongating mRNA and facilitates addition of a poly(A) tail [7]. While RNA involved in protein translation (mRNA, rRNA, and tRNA) is ultimately exported from the nucleus, there are several types of RNA that are retained and carry out their functions inside the nucleus. These nuclear-retained RNAs include long non-coding RNA (lncRNAs) and small nucleolar RNAs (snoRNAs) among others. The expansive role of RNA in regulating chromatin structure is beginning to become clear, such as mediating interactions between proteins and chromatin, regulating the activity of chromatin-modifying enzymes, or facilitating changes in chromatin architecture. In certain contexts RNA can also interact with proteins and/or DNA to form liquid-liquid phase separated nuclear bodies, which have been associated with processes such as transcription [8] and nuclear pore formation [9]. These observations have gradually updated the view that RNA is primarily an information carrier in eukaryotic cells to include its own contributions in nuclear function.

1.1.3 Nuclear bodies

Nuclear bodies form by liquid-liquid phase separation among DNA, RNA and proteins that share mutual affinity, resulting in chemically distinct sub-compartments [10]. These nuclear bodies, also known as nuclear condensates, display liquid-like properties and are immiscible with the surrounding nucleoplasm. There are several examples of nuclear condensates including Cajal bodies, histone locus bodies and nucleoli—these function in splicing assembly [11], histone mRNA processing [12], and rRNA processing [13], respectively. The mechanisms surrounding the nucleation and behavior of nuclear condensates are currently an area of major interest, and there is an ongoing debate over the role such condensates play in regulating chromatin organization and transcription [14, 15]. While the function of and mechanisms involving nuclear bodies remain controversial, they all seem to involve a nucleation step when a combination of RNA, DNA, and proteins reach a critical concentration and segregate into a biomolecule-rich phase. The concentration at which condensates form is affected by the affinity between the constituent biomolecules, which can be regulated by biochemical modifications of proteins and RNA [10]. This indicates that RNA-protein interactions are a key part of understanding the role of liquid condensates in chromatin biology.

1.1.4 Approaches to study RNA-protein interactions

RNA-protein interactions are emerging as increasingly relevant players in chromatin regulation. Classically, RNA-binding proteins (RBPs) were thought to bind to specific RNA targets via a conserved RNA-binding domains (RBDs) such as the RNA recognition motif (RRM) [16] or hnRNP K homology domain (KH) [17]. These interactions were thought to dictate the processing and function of the bound RNA. However, emerging evidence demonstrating the prevalence of RNA condensates in multiple cellular compartments, reviewed here [10, 18]–along with the cataloguing of many lncRNAs with evident functions in gene regulation [19]–have shown that RNA can also regulate its protein partners, termed riboregulation [20]. Given the importance of RNA-protein interactions to the structure of nuclear condensates and chromatin regulation more generally, improved mapping of RNA-protein links can enable design of separation-of-function mutants that can reveal the function of condensates in chromatin biology. Given that many chromatin-associated RBPs interact with a wide range of RNA transcripts [21], efforts to probe RNA-protein interactions in chromatin regulation have tended to use protein-centric approaches. For example some groups have identified novel chromatin-associated RBPs, using candidate-based domain mapping to identify and remove RNA binding regions (RBRs) and determine their effect on chromatin dynamics [22–25]. While showing promise, systematic cataloguing of RBPs requires a more unbiased approach, particularly given that many novel RBPs lack classical RBDs [26].

When I began my dissertation work there were already several approaches to identify RBPs in an unbiased manner. These include RNA interactome capture where mRNA-bound proteins are enriched using poly-adenylated (polyA) RNA-based selection and identified via mass spectrometry [27, 28]. These approaches have revealed many novel RBPs and expanded our understanding of RBP characteristics, in particular that many RBPs lack classical RBDs, underscoring the shortcomings of identifying novel RBPs based on similarity to existing proteins. However, because of their reliance on polyA-based selection, these interactome capture methods miss proteins that bind non poly-adenylated RNAs such as eRNAs or nascent RNA. Another approach from Kramer *et al.* used UV crosslinking of cells and analysis of protein-RNA adducts via mass spectrometry to identify sites of crosslinking [29]. Using this method the authors identified 257 peptides with nucleotide adducts from only 124 RBPs, likely reflecting the challenge of unambiguously assigning unpredictable nucleotide adducts to peptide spectra [29]. We reasoned that developing a truly unbiased approach could identify novel RBPs and generate new hypotheses regarding chromatin biology.

1.2 *Polycomb* Repressive Complex 2 (PRC2) and RNA

1.2.1 Overview of PRC2

Proteins of the *Polycomb* Group (PcG) ensure maintenance of transcriptional repression during cell differentiation and are essential for proper development. Mutations in PcG genes cause homeotic transformations in body patterning and are associated with aberrant cell proliferation in multiple cancer types [30]. PcG proteins commonly function as part of multi-subunit complexes, which in metazoans chiefly comprise PRC1 and PRC2, which are recruited to chromatin where they modify histones to deposit H2AK119 ubiquitination and H3K27 mono, di-, and tri-methyl modifications, respectively (Figure 1.1 and [31]). Presence of these chromatin modifications are associated with chromatin compaction and repression of transcription [32]. PRC2 is a histone methyltransferase complex with a core complex consisting of a catalytic subunit (EZH1 or EZH2); a regulatory subunit (EED); a chromatin-binding subunit RBBP4 or RBBP7, and a structural subunit SUZ12. In addition to these four subunits several additional subunits have been discovered (AEBP2 [33], JARID2 [34], PCL1-3 [35], EPOP [36, 37], and PALI1/2 [38]) that associate into two mutually exclusive holo-complexes, denoted as PRC2.1 and PRC2.2, are are required for proper PRC2 localization and repression, reviewed in [39, 40].

The mechanisms governing the recruitment and activity of PRC1 and PRC2 are multi-faceted and have attracted significant attention in recent years. As mentioned above, PRC2-deposited H3K27me3 can recruit PRC1 onto chromatin at *Polycomb* target genes [33, 41, 42]. PRC2 can also bind to the histone mark H2AK119Ub and is dependent on the mark for proper chromatin localization [43–45]. This suggests that there is cross-talk between the two complexes that reinforce their proper localization. In addition to histone marks, PRC2 also recognizes un-methylated CpG islands on genomic DNA in order to bind gene promoters. [46, 47].

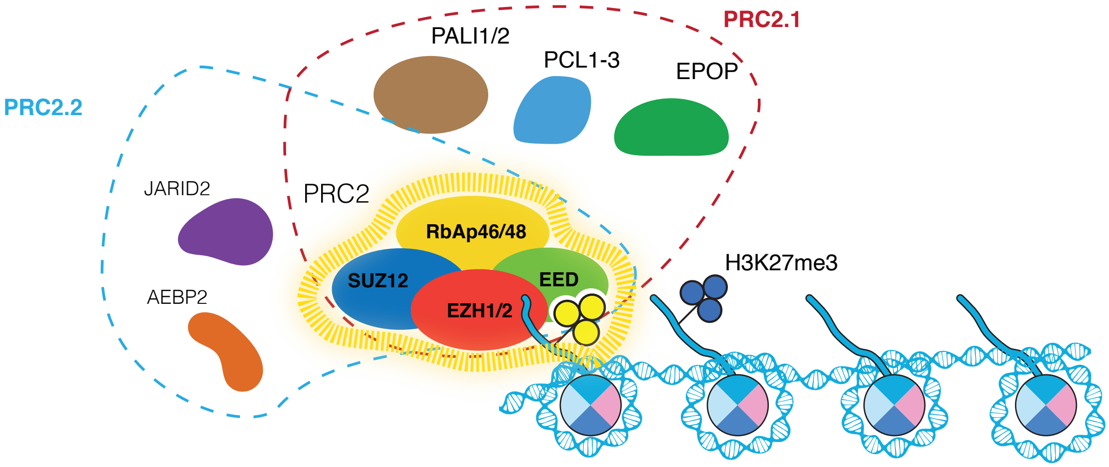


FIGURE 1.1 PRC2 is a histone methyltransferase complex

PRC2 consists of 4 core subunits: SUZ12, RBBP6/8, EED, and catalytic subunit EZH1/2. PRC2 methylates histone H3 to deposit the H3K27me3 modification that is required for maintenance of transcriptional repression. EED can bind to existing H3K27me3 modifications to stimulate PRC2 methyl-transferase activity in a positive feedback mechanism [48]. Additional PRC2 subunits associate with the core complex into two mutually exclusive holo-complexes, PRC2.1 and PRC2.2 [49]. These subunits have been shown to mediate PRC2 recruitment onto chromatin and to stimulate PRC2 activity [31]

Both PRC1 and PRC2 have also been found to interact with RNA [22, 50–53], suggesting that RNA-protein interactions also affect *Polycomb* function in chromatin. While we still know very little about PRC1 interaction with RNA, there has been a concerted effort to dissect the interactions of PRC2 with RNA and their function. [22, 23, 50, 54–59].

1.2.2 RNA-mediated PRC2 recruitment

RNA has been linked to PRC2 recruitment to chromatin through several lines of evidence. The first observations found that PRC2 interacted specifically with lncRNAs that led to recruitment of PRC2 onto chromatin [52, 53], such as the lncRNA Xist-RepA. Xist was proposed to recruit PRC2 onto the X chromosome during X inactivation through the *RepA* region [52]. However, subsequent work has undermined this observation, finding that the Xist RepA fragment was not required for recruitment of PRC2 to the inactive X chromosome [60]. Biochemical characterization of PRC2 affinity for RNA showed that PRC2 has promiscuous affinity for many RNA sequences at can be found at many actively transcribing genes, suggesting that a broader set of RNA transcripts outside of a few specific lncRNAs interact with PRC2 [55, 59, 61]. Careful biochemical experiments have revealed that PRC2 binds G-quadruplex RNA structures [62] with higher affinity than other RNA ligands [21, 56, 63]. Attempts to abolish PRC2-RNA interactions through site-directed mutagenesis of EZH2 resulted in reduced but still relevant nM scale affinity for G-quadruplex RNA [64], suggesting there are other regions of the complex that contribute to RNA recognition. Yet recent evidence contradicts the idea that RNA recruits PRC2 to chromatin, such as a recent study showing that upon application of RNAse to cell nuclei PRC2 was enriched onto chromatin, implying that RNA competes with chromatin for binding to PRC2 [61, 63].

1.2.3 RNA regulation of PRC2

In addition to affecting PRC2 localization on chromatin, RNA can inhibit the methyltransferase activity of PRC2 [56, 57]. This observation formed the basis of a ’poised inhibition’ model where nascent RNA from actively transcribed genes can inhibit PRC2 activity and prevent inappropriate silencing [56]. The mechanism by which RNA inhibits PRC2 has not been fully elucidated. It has been shown that G-quadruplex RNA can compete PRC2 away from nucleosomal substrates and inhibits histone methyltransferase activity without interfering with the automethylation activity of EZH2 [63]. This suggested that RNA competes with DNA for binding PRC2. However in an earlier study in Drosophila, over-expression of a *Polycomb* responsive element (PRE) anti-sense RNA led to significantly lower H3K27me3 levels at the corresponding PRE without major loss of PRC2 occupancy [65]. This suggests that there may be other mechanisms of RNA-inhibition of PRC2 besides competition for initial chromatin binding, such as competition for binding to the catalytic site, or yet to be identified allosteric effects. It is also plausible that one or more of these mechanisms could both contribute to RNA inhibition of PRC2.

1.3 SARS-COV-2 diagnostics

1.3.1 History of COVID-19 pandemic

At the end of 2019, unusually severe cases of respiratory infection were reported in Wuhan, China [66]. Infected individuals presented with cough, fever and flu-like symptoms that could progress to pneumonia and respiratory insufficiency. RNA isolation and sequencing of samples from the affected patients combined with phylogenetic analysis led to the identification of a novel Betacoronavirus strain in the subgenus Sarbecovirus [66–68]. Based on these studies, the novel virus was officially classified as SARS-CoV-2 (severe acute respiratory syndrome coronavirus-2) and its associated disease was termed COVID-19 [69]. Similar to two other viruses of the same family, SARS-CoV and MERS-CoV (Middle East Respiratory Syndrome), SARS-CoV was suspected to be zoonotic [70], with the SARS-CoV-2 nsp7 protein sharing 100% amino acid identity with an existing bat coronavirus [68]. Quickly after its identification in China, a number of similar cases were identified worldwide. By March 2020, confirmed SARS-CoV-2 cases were present in

114 countries and the World Health Organization (WHO) declared the global COVID-19 outbreak as a pandemic [71].

1.3.2 Biology of SARS-CoV-2

SARS-CoV-2 is a Betacoronavirus of the *Coronaviridae* family, which is named for the resemblance of the protein spikes on their viral surface to the solar corona [72]. SARS-CoV-2 is an enveloped single-stranded positive-sense RNA virus with a predicted broad range of hosts [73]. Its genome consists of 6 functional open reading frames (ORFs): ORF1a/ORF1b, S, E, M and N. The ORFs are highly conserved across members of the Betacoronavirus genus. ORF1a and ORF1b make up the initial two-thirds of the genome sequence and encode the non-structural proteins, including those required for viral replication. The 3' final one-third end of the genome encodes the main structural proteins, which are the spike glycoprotein (S), the envelope glycoprotein (E), the membrane protein (M) and the nucleocapsid protein (N) [74]. The spike protein S interacts with the host receptor angiotensin-converting enzyme 2 (ACE2), a protein expressed by various cell types of the lower respiratory tract [74]. Binding of ACE2 with the SARS-CoV-2 spike protein leads to fusion of the virion with the host cell and release of its genome into the cytoplasm. A viral-encoded RNA-dependent RNA polymerase (RdRp) then amplifies the genome to generate full length positive-strand copies, as well using discontinuous RNA extension to generate a set of nested subgenomic RNAs that serve as the templates for translation of the structural proteins [75]. The resulting full-length viral genomes are packaged with structural proteins into viral particles and exit the host cell via the lysosomal trafficking pathway [74].

1.3.3 Clinical features and transmission of SARS-CoV-2 disease

Estimate of Lethality

SARS-CoV-2 infection is associated with a broad spectrum of symptoms, ranging from asymptomatic cases to severe and potentially lethal respiratory illness [76]. Respiratory damage is due to damage to tissues of the respiratory system as well as secondary damage caused by pro-inflammatory cytokine responses [76]. A study of SARS-CoV-2 cases during the spring of 2020 in New York City used epidemiological modeling augmented by detailed population-level data to estimate the infection-fatality risk, i.e. the likelihood of death among infected individuals. They estimated an overall infection-fatality risk of 1.39%, with a rate as high as 14.2% among individuals aged 75 years and older [77].

Viral Transmission

A major factor behind the severity of the SARS-CoV-2 pandemic is the high transmission rate of the virus. A common measure of contagiousness is the basic reproduction number R_0, defined as the average number of infections caused by one infected person in a population with no immunity. A R_0 below 1 indicates that the epidemic will eventually subside. Estimates of R_0 across countries show high variability, likely reflecting the different policy responses and socioeconomic characteristics of each country. A survey of 15 estimates of R_0 in different countries from December 2019 to May 2020 showed that $R_0 > 1$ in all countries at the time of publication [78]. The high R_0 observed for SARS-CoV-2 may be linked to the fact that a large fraction–48 to 62%–of the transmission may involve pre-symptomatic carriers [76]. The combined potential of rapid SARS-CoV-2 transmission and clinical complications highlight the need for effective public health responses.

1.3.4 SARS-CoV-2 detection in public health and clinical medicine

In response to the increasing number of infections, several countries have invested unprecedented efforts in the development and conduction of testing. In the United States, the number of daily new tests performed increased from 32,687 on May 1st 2020 to 1,661,491 as of Jan 31st 2021, showing a positive-tested ratio of 8.4% and 8.1%, respectively [79].

Early and accurate detection of SARS-CoV-2 is key for both effective clinical treatment as well as limiting disease transmission. Diagnostic tests are used to confirm the presence of disease, and to inform clinical care. To these ends, effective diagnostic tests must be accurate, sensitive and cost-effective. At the level of public health interventions, screening tests are used to detect the presence of a pathogen in a population of subjects, regardless of whether they display symptoms. Screening tests are designed for the frequent monitoring of disease in a larger population sample, especially in populations at higher risk [80]. In contrast to diagnostic testing, screening tests are not intended to guide individual clinical care, but instead to detect disease early . In most cases, positive screening test results are subsequently confirmed with a diagnostic test [80]. Because SARS-CoV-2 is a highly transmissible disease with prevalence in asymptomatic subjects, early detection of cases through screening can provide individual-specific quarantine recommendations to limit transmission. Systematic application of screening tests to an at-risk population is considered a part of population surveillance, which can help identify novel disease clusters and refine understanding of transmission and reveal at-risk behaviors or populations. In addition to testing, surveillance can encompass other interventions, such as contact-tracing [81]. The broader purpose of population surveillance tests mean they do not need to be as sensitive as diagnostic tests, and should instead minimize costs and turn-around times while maintaining an acceptable sensitivity and accuracy profile [82].

1.3.5 Diagnostic testing terminology

Current testing approaches focus on the detection of SARS-CoV-2 RNA or protein as well as on the detection of markers of the host immune response (antibodies against SARS-CoV-2). For the purpose of this dissertation, I will focus on methods aimed at detecting active SARS-CoV-2 infections and refer to the following review for information on serological testing methods for COVID-19 [83]. Before introducing the different testing methods I want to clarify some commonly confused terminology. The *limit of detection* (LoD) refers to the lowest amount of virus that yields an analytical response. Typically, the lowest analytical response is defined as the lowest analyte signal that can be identified from noise. The FDA defines the LoD for SARS-CoV-2 molecular tests as the level at which 19/20 replicates are positive [84]. The *limit of quantification* (LoQ) applies to methods that aim to quantify an analyte, and is the lowest concentration of analyte for which a quantitative assessment can be provided [85]. For example some testing methods may not show a linear response at very low analyte concentrations and therefore their LoD would be lower than their LoQ. *Accuracy* typically refers to the level of agreement between the results from a test in question with a reference test. As the first COVID-19 tests developed used reverse-transcription polymerase chain reaction (RT-PCR) to detect viral RNA [67], and since RT-PCR is still the most common diagnostic method [86] this is typically the test chosen as reference. Since the negative consequences of false positive (unnecessary isolation) and false negative results (risk of transmission) are considerable, the FDA recommends that clinical testing should always be paired with clinical evaluation [84]. *Analytical sensitivity* of testing is defined as the likelihood to obtaining a positive result for any sample harboring the virus. As a corrolary, *Analytic specificity* refers to the likelihood of obtaining a negative result for samples not containing the analyte in study but containing other pathogens.

1.3.6 Sample collection and processing

Several types of biological samples have been considered as potential candidates for SARS-CoV-2 testing. The most frequently used are nasal swab, saliva, blood and serum but other sources such as fecal samples have also been proposed as useful alternatives [87]. Sample collection and processing should preserve integrity of the viral analyte, while protecting laboratory personnel from hazardous exposures. Great effort has been made in developing sample collection protocols that minimize interpersonal contact, such as point of care (POC) appointment-only collection models, drive-thru collection models, and self-collection or self-testing models [87]. Most of the testing techniques require a series of post-collection processing steps (most commonly RNA extraction). Depending on the downstream analytical approach, these steps can be costly and labor-intensive. During the early phases of the pandemic, the shortage of testing reagents posed a significant challenge to increasing testing capacity. Alternative strategies have focused on automated extraction protocols and extraction free testing protocols [88]. Although the latter approach could provide faster and more inexpensive alternatives, it is typically associated with lower accuracy and sensitivity [89]. These features make them more suitable for screening and surveillance purposes compared to diagnostic testing.

1.3.7 Survey of testing methods

Current SARS-CoV-2 molecular testing approaches can be divided into two broad categories: methods to detect viral RNA and methods to detect viral proteins.

Detection of SARS-COV-2 proteins

Methods to detect viral proteins are primarily immune-based assays capable of recognizing specific antigens. Low antibody-antigen specificity is often the limiting step in the development of these assays, since there is a risk of antibodies cross-reacting with unrelated pathogens. Sensitivity largely varies across different tests and is highly affected by sample collection procedures. In addition, SARS-CoV-2 viral protein levels may be below the LoD in the early phases of infection. For these reasons, antigen tests are typically used for screening purposes and not for diagnostic purposes [87].

Detection of SARS-CoV-2 RNA

Methods to detect SARS-CoV-2 RNA are the most widely-used form of testing for diagnostic purposes, as they are able to detect the earliest stages of infection. RT-PCR based approaches are the current standard for diagnostic testing and are the most widely used method. They involve processing collected samples with an optional RNA purification step followed by reverse transcription of RNA to cDNA and PCR amplification with specific primer sequences. Amplified DNA can then be detected using a fluorescent dye [87]. RT-PCR is a well validated approach, but an important factor to its success is the design of effective primer sets. Different studies have shown that the selection of primers can affect the sensitivity of the method, likely due to differences in abundance of coronavirus sgRNAs [90]. Studies comparing primer sensitivity for SARS-CoV-2 have found primers targeting the N, E, and Orf1ab genes demonstrate the highest sensitivity [91–93].

Clustered regularly interspaced short palindromic repeats (CRISPR) based diagnostic methods have recently emerged as an alternative to RT-PCR [94]. One approach, dubbed SHERLOCK, first amplifies RNA via recombinase polymerase amplification (see paragraph below), then transcribes the DNA back to RNA. The amplified RNA sequence can then be recognized by a complementary RNA nucleotide (guide RNA, gRNA) and an RNA targeting CRISPR-associated nuclease (Cas) such as Cas13. Upon binding of the gRNA to target sequence, the Cas nuclease is activated and induces cleavage of surrounding reporter RNA probes [95]. An alternative method, DETECTR, instead uses a DNA-recognizing nuclease Cas12a, thus omitting the need for the reverse transcription step [96]. These approaches can be used in isothermal conditions, and therefore do not require thermocyclers and are amenable to point-of-care settings, and have acceptable sensitivity to detect viral RNA concentrations as low as 100 copies/µL in under an hour [96].

In addition to the CRISPR-based diagnostics, other isothermal amplification-based methods have been developed to detect SARS-CoV-2 RNA, such as recombinase polymerase amplification (RPA) and reverse transcription coupled with loop-mediated isothermal amplification (RT-LAMP) [97, 98]. RPA uses recombinases to insert primers into dsDNA and single-stranded binding proteins to stabilize the resulting evicted ssDNA, followed by DNA extension using a strand-displacing polymerase [97, 99]. Similar to RPA, RT-LAMP also takes advantage of a strand-displacing polymerase to amplify from RNA or DNA templates; however, it also uses 4-6 unique primers to bind the target sequence, leading to rapid amplification via a series of ’cauliflower-like’ intermediates [100, 101]. Typical RT-LAMP reactions can generate microgram quantities of DNA in less than 30 minutes, convenient for point of care testing settings. As the RT-LAMP reaction proceeds, insoluble magnesium pyrophosphate accumulates and increases the turbidity of the solution, allowing for a convenient visual indicator of amplification [102]. Other readouts include colorimetric pH indicators to monitor reaction progress [103] or fluorescent or colorimetric intercalating dyes to detect amplified DNA [104]. Due to concerns that these generic readouts are susceptible to artifacts, sequence-specific fluorescent readouts have been developed using quenched fluorescent primers or molecular beacons [105–107].

High throughput sequencing-based detection of SARS-CoV-2 RNA

In addition to typical 96 or 384-well plate-based testing methods, other groups have explored using high-throughput DNA sequencing to allow for massively parallel testing [108–114]. These methods used modified RT-PCR or RT-LAMP protocols by incorporating a uniquely identifiable nucleotide barcode into each sample. Following amplification, up to tens of thousands of barcoded DNA samples can be combined and processed for sequencing on a next-generation sequencing platform. Barcodes that appear in conjunction with viral sequences can then be deconvoluted to provide a positive result. While these approaches require more sophisticated equipment and analysis than standard tests, the wide and inexpensive availability of DNA sequencing capacity in academic and private settings make this an attractive option for large-scale testing programs. Testing protocols that can harness this powerful capacity coupled with unsophisticated sample collection and processing could enable a huge increase in testing capacity and more effective surveillance.

1.4 Goals

In the following chapters, I present experiments aimed at developing tools to probe RNA-protein interactions in chromatin biology, and then adapt those tools to answer questions about how RNA regulates PRC2 activity on chromatin. Following the emergence of COVID-19, I switched focus to work on a diagnostic methods for SARS-COV-2, and here I present the results of that work as well.

In the first chapter, I present data on the development of a novel method to identify RNA-protein interactions using UV crosslinking and high resolution mass spectrometry. We developed a method to identify RNA-binding proteins and RNA-binding regions (RBRs) using crosslinking and mass spectrometry. We identify 8̃00 RBPs in mouse embyronic stem cells, including over 400 that had not been previously found to bind RNA. These novel RBPs included many with functions related to chromatin, suggesting a greater role for RNA in chromatin regulation.

In the second chapter, I relate an additional set of experiments continuing to examine RNA-protein interactions in chromatin function. Using a targeted version of RBR-ID on the chromatin regulatory complex PRC2, we generated a map of multiples sites of RNA interaction on the complex, demonstrating that all known PRC2 variants show extensive RNA-binding activity. One of these sites fell inside an allosteric regulatory center of PRC2, and we found that stimulatory peptides binding in this site were able to rescue RNA-mediated inhibition of PRC2 methyltransferase activity. This study revealed new aspects of RNA regulation in PRC2 and suggested how stimulatory and inhibitory stimuli are integrated on PRC2.

In the third chapter I present a testing method we developed to detect SARS-CoV-2 and influenza RNA from saliva. The method, which we called COV-ID, is based on RT-LAMP coupled with high-throughput sequencing and is able to detect as few as 10 copies of viral RNA per microliter. COV-ID can be multiplexed to test for both SARS-CoV-2 and influenza in the same assay. We performed COV-ID on ten COVID-19 clinical samples, finding 100% agreement with clinical RT-PCR data. The method requires minimal sample processing and handling, and up to tens of thousands of samples can be processed in parallel with results delivered within 24 hours.

In the final chapter I summarized the results from the preceding chapters and present my findings with respect to subsequent advances in the field. I then discuss further experiments to build on the current state of the field of RNA regulation of PRC2, as well as the state of viral diagnostics in the current pandemic and in the future.

1.5 References

1. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403,** 41–5 (2000).
2. Reinberg, D. & Vales, L. D. Chromatin domains rich in inheritance. *Science* **361,** 33–34 (2018).
3. Henikoff, S. & Shilatifard, A. Histone modification: cause or cog? *Trends Genet* **27,** 389–96 (2011).
4. Henikoff, S. & Greally, J. M. Epigenetics, cellular memory and gene regulation. *Curr Biol* **26,** R644–8 (2016).
5. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Sci Adv* **5,** eaaw1668 (2019).
6. Hildebrand, E. M. & Dekker, J. Mechanisms and Functions of Chromosome Compartmentalization. *Trends Biochem Sci* **45,** 385–396 (2020).
7. Manley, J. L. & Di Giammartino, D. C. in *Encyclopedia of Biological Chemistry* (eds Lennarz, W. J. & Lane, M. D.) 188–193 (Academic Press, Waltham, 2013).
8. Wei, M. T. *et al.* Nucleated transcriptional condensates amplify gene expression. *Nat Cell Biol* **22,** 1187–1196 (2020).
9. Hampoelz, B. *et al.* Nuclear Pores Assemble from Nucleoporin Condensates During Oogenesis. *Cell* **179,** 671–686 e17 (2019).
10. Strom, A. R. & Brangwynne, C. P. The liquid nucleome - phase transitions in the nucleus at a glance. *J Cell Sci* **132** (2019).
11. Sawyer, I. A., Hager, G. L. & Dundr, M. Specific genomic cues regulate Cajal body assembly. *RNA Biol* **14,** 791–803 (2017).
12. Duronio, R. J. & Marzluff, W. F. Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. *RNA Biol* **14,** 726–738 (2017).

13. Brangwynne, C. P., Mitchison, T. J. & Hyman, A. A. Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proc Natl Acad Sci U S A* **108,** 4334–9 (2011).

14. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169,** 13–23 (2017).

15. McSwiggen, D. T., Mir, M., Darzacq, X. & Tjian, R. Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Genes Dev* **33,** 1619–1634 (2019).

16. Clery, A., Blatter, M. & Allain, F. H. RNA recognition motifs: boring? Not quite. *Curr Opin Struct Biol* **18,** 290–8 (2008).

17. Valverde, R., Edwards, L. & Regan, L. Structure and function of KH domains. *FEBS J* **275,** 2712–26 (2008).

18. Sabari, B. R., Dall'Agnese, A. & Young, R. A. Biomolecular Condensates in the Nucleus. *Trends Biochem Sci* **45,** 961–977 (2020).

19. Shields, E. J., Petracovici, A. F. & Bonasio, R. lncRedibly versatile: biochemical and biological functions of long noncoding RNAs. *Biochem J* **476,** 1083–1104 (2019).

20. Horos, R. *et al.* The Small Non-coding Vault RNA1-1 Acts as a Riboregulator of Autophagy. *Cell* **176,** 1054–1067 e12 (2019).

21. Hendrickson, D., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biology* **17** (2016).

22. Kaneko, S. *et al.* Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. *Genes and Development* **24,** 2615–2620 (2010).

23. Kaneko, S. *et al.* Interactions between JARID2 and Noncoding RNAs Regulate PRC2 Recruitment to Chromatin. *Molecular Cell* **53,** 290–300 (2014).

24. Bonasio, R. *et al.* Interactions with RNA direct the Polycomb group protein SCML2 to chromatin where it represses target genes. *Elife* **3,** e02637 (2014).

25. Saldaña-Meyer, R. *et al.* CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. *Genes and Development* **28,** 723–734 (2014).

26. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol* **19,** 327–341 (2018).

27. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* **46,** 674–90 (2012).

28. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149,** 1393–1406 (2012).

29. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods* **11,** 1064–70 (2014).

30. Pasini, D., Bracken, A. P. & Helin, K. Polycomb group proteins in cell cycle progression and cancer. *Cell Cycle* **3,** 396–400 (2004).

31. Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* **171,** 34–57 (2017).

32. Grau, D. J. *et al.* Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. *Genes Dev* **25,** 2210–21 (2011).

33. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298,** 1039–43 (2002).

34. Peng, J. C. *et al.* Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139,** 1290–302 (2009).

35. Wang, S., Robertson, G. P. & Zhu, J. A novel human homologue of Drosophila polycomblike gene is up-regulated in multiple cancers. *Gene* **343,** 69–78 (2004).

36. Beringer, M. *et al.* EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. *Mol Cell* **64,** 645–658 (2016).

37. Liefke, R., Karwacki-Neisius, V. & Shi, Y. EPOP Interacts with Elongin BC and USP7 to Modulate the Chromatin Landscape. *Mol Cell* **64,** 659–672 (2016).

38. Conway, E. *et al.* A Family of Vertebrate-Specific Polycombs Encoded by the LCOR/LCORL Genes Balance PRC2 Subtype Activities. *Mol Cell* **70,** 408–421 e8 (2018).

39. Laugesen, A., Hojfeldt, J. W. & Helin, K. Molecular Mechanisms Directing PRC2 Recruitment and H3K27 Methylation. *Mol Cell* **74,** 8–18 (2019).

40. Van Mierlo, G., Veenstra, G. J. C., Vermeulen, M. & Marks, H. The Complexity of PRC2 Subcomplexes. *Trends Cell Biol* **29,** 660–671 (2019).

41. Min, J., Zhang, Y. & Xu, R. M. Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. *Genes Dev* **17,** 1823–8 (2003).

42. Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431,** 873–8 (2004).

43. Blackledge, N. P. *et al.* Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell* **157,** 1445–1459 (2014).

44. Cooper, S. *et al.* Jarid2 binds mono-ubiquitylated H2A lysine 119 to mediate crosstalk between Polycomb complexes PRC1 and PRC2. *Nat Commun* **7,** 13661 (2016).

45. Kasinath, V. *et al.* JARID2 and AEBP2 regulate PRC2 in the presence of H2AK119ub1 and other histone modifications. *Science* **371** (2021).

46. Mendenhall, E. M. *et al.* GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet* **6,** e1001244 (2010).

47. Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature* **549,** 287–291 (2017).

48. Margueron, R. *et al.* Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461,** 762–7 (2009).

49. Hauri, S. *et al.* A High-Density Map for Navigating the Human Polycomb Complexome. *Cell Rep* **17,** 583–595 (2016).

50. Kanhere, A. *et al.* Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell* **38,** 675–88 (2010).

51. Bonasio, R. & Shiekhattar, R. Regulation of transcription by long noncoding RNAs. *Annu Rev Genet* **48,** 433–55 (2014).

52. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–6 (2008).

53. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–23 (2007).

54. Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40,** 939–53 (2010).

55. Kaneko, S., Son, J., Shen, S. S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20,** 1258–64 (2013).

56. Kaneko, S., Son, J., Bonasio, R., Shen, S. S. & Reinberg, D. Nascent RNA interaction keeps PRC2 activity poised and in check. *Genes Dev* **28,** 1983–8 (2014).

57. Cifuentes-Rojas, C., Hernandez, A. J., Sarma, K. & Lee, J. T. Regulatory Interactions between RNA and Polycomb Repressive Complex 2. *Molecular Cell* **55,** 171–185 (2014).

58. Sarma, K. *et al.* ATRX directs binding of PRC2 to Xist RNA and Polycomb targets. *Cell* **159,** 869–883 (2014).

59. Davidovich, C., Zheng, L., Goodrich, K. J. & Cech, T. R. Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* **20,** 1250–7 (2013).

60. Da Rocha, S. T. *et al.* Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. *Mol Cell* **53,** 301–16 (2014).

61. Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res* **26,** 896–907 (2016).

62. Bugaut, A. & Balasubramanian, S. 5'-UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic Acids Res* **40,** 4727–41 (2012).

63. Wang, X. *et al.* Targeting of Polycomb Repressive Complex 2 to RNA by Short Repeats of Consecutive Guanines. *Mol Cell* **65,** 1056–1067 e5 (2017).

64. Long, Y. *et al.* Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *eLife* **6** (2017).

65. Herzog, V. A. *et al.* A strand-specific switch in noncoding transcription switches the function of a Polycomb/Trithorax response element. *Nat Genet* **46,** 973–981 (2014).

66. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382,** 727–733 (2020).

67. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (2020).

68. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579,** 265–269 (2020).

69. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5,** 536–544 (2020).

70. Cui, J., Li, F. & Shi, Z. L. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* **17,** 181–192 (2019).

71. World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (2020).

72. Virology: Coronaviruses. *Nature* **220,** 650–650 (1968).

73. Damas, J. *et al.* Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc Natl Acad Sci U S A* **117,** 22311–22322 (2020).

74. V'Kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol* (2020).

75. Sawicki, S. G. & Sawicki, D. L. Coronaviruses use discontinuous extension for synthesis of subgenome-length negative strands. *Adv Exp Med Biol* **380,** 499–506 (1995).

76. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA* **324,** 782–793 (2020).

77. Yang, W. *et al.* Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis. *Lancet Infect Dis* **21,** 203–212 (2021).

78. Barber, A. *et al.* The basic reproduction number of SARS-CoV-2: a scoping review of available evidence. *medRxiv,* 2020.07.28.20163535 (2020).

79. Johns Hopkins University Coronavirus Resource Center. Daily State-By-State Testing Trends. https://coronavirus.jhu.edu/testing/individual-states (2021).

80. Prinzi, A. Screening Versus Diagnostic Tests for COVID-19, What's the Difference? https://asm.org/Articles/2020/December/Screening-Versus-Diagnostic-Tests-for-COVID-19,-Wh (2020).

81. Oleske, D. M. Screening and Surveillance for Promoting Population Health. *Epidemiology and the Delivery of Health Care Services: Methods and Applications,* 131–150 (2009).

82. Mina, M. J., Parker, R. & Larremore, D. B. Rethinking Covid-19 Test Sensitivity - A Strategy for Containment. *N Engl J Med* **383,** e120 (2020).

83. Espejo, A. P. *et al.* Review of Current Advances in Serologic Testing for COVID-19. *Am J Clin Pathol* **154,** 293–304 (2020).

84. United States Food and Drug Administration. Policy for Coronavirus Disease-2019 Tests During the Public Health Emergency (Revised). https://www.fda.gov/media/135659/download (2020).

85. Armbruster, D. A. & Pry, T. Limit of blank, limit of detection and limit of quantitation. *Clin Biochem Rev* **29 Suppl 1,** S49–52 (2008).

86. MacKay, M. J. *et al.* The COVID-19 XPRIZE and the need for scalable, fast, and widespread testing. *Nat Biotechnol* **38,** 1021–1024 (2020).

87. Jayamohan, H. *et al.* SARS-CoV-2 pandemic: a review of molecular diagnostic tools including sample collection and commercial response with associated advantages and limitations. *Anal Bioanal Chem* **413,** 49–71 (2021).

88. Esbin, M. N. *et al.* Overcoming the bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for COVID-19 detection. *RNA* **26,** 771–783 (2020).

89. Israeli, O. *et al.* Evaluating the efficacy of RT-qPCR SARS-CoV-2 direct approaches in comparison to RNA extraction. *Int J Infect Dis* **99,** 352–354 (2020).

90. Weiss, S. R. & Leibowitz, J. L. Characterization of murine coronavirus RNA by hybridization with virus-specific cDNA probes. *J Gen Virol* **64 (Pt 1),** 127–33 (1983).

91. Nalla, A. K. *et al.* Comparative Performance of SARS-CoV-2 Detection Assays Using Seven Different Primer-Probe Sets and One Assay Kit. *J Clin Microbiol* **58** (2020).

92. Corman, V. M. *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill* **25** (2020).

93. Jung, Y. *et al.* Comparative Analysis of Primer-Probe Sets for RT-qPCR of COVID-19 Causative Virus (SARS-CoV-2). *ACS Infect Dis* **6,** 2513–2523 (2020).

94. Gootenberg, J. S. *et al.* Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356,** 438–442 (2017).

95. Joung, J. *et al.* Point-of-care testing for COVID-19 using SHERLOCK diagnostics. *medRxiv* (2020).

96. Fozouni, P. *et al.* Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. *Cell* **184,** 323–333 e9 (2021).

97. Daher, R. K., Stewart, G., Boissinot, M. & Bergeron, M. G. Recombinase Polymerase Amplification for Diagnostic Applications. *Clin Chem* **62,** 947–58 (2016).

98. Augustine, R. *et al.* Loop-Mediated Isothermal Amplification (LAMP): A Rapid, Sensitive, Specific, and Cost-Effective Point-of-Care Test for Coronaviruses in the Context of COVID-19 Pandemic. *Biology (Basel)* **9** (2020).

99. Xia, S. & Chen, X. Single-copy sensitive, field-deployable, and simultaneous dual-gene detection of SARS-CoV-2 RNA via modified RT-RPA. *Cell Discov* **6,** 37 (2020).

100. Notomi, T. *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res* **28,** E63 (2000).

101. Nagamine, K., Hase, T. & Notomi, T. Accelerated reaction by loop-mediated isothermal amplification using loop primers. *Mol Cell Probes* **16,** 223–9 (2002).

102. Mori, Y., Kitao, M., Tomita, N. & Notomi, T. Real-time turbidimetry of LAMP reaction for quantifying template DNA. *J Biochem Biophys Methods* **59,** 145–57 (2004).

103. Calvert, A. E., Biggerstaff, B. J., Tanner, N. A., Lauterbach, M. & Lanciotti, R. S. Rapid colorimetric detection of Zika virus from serum and urine specimens by reverse transcription loop-mediated isothermal amplification (RT-LAMP). *PLoS One* **12,** e0185340 (2017).

104. Rabe, B. A. & Cepko, C. SARS-CoV-2 detection using isothermal amplification and a rapid, inexpensive protocol for sample inactivation and purification. *Proc Natl Acad Sci U S A* **117,** 24450–24458 (2020).

105. Sherrill-Mix, S. *et al.* LAMP-BEAC: Detection of SARS-CoV-2 RNA Using RT-LAMP and Molecular Beacons. *medRxiv,* 2020.08.13.20173757 (2020).

106. Tanner, N. A., Zhang, Y. & Evans T. C., J. Simultaneous multiple target detection in real-time loop-mediated isothermal amplification. *Biotechniques* **53,** 81–9 (2012).

107. Ball, C. S. *et al.* Quenching of Unincorporated Amplification Signal Reporters in Reverse-Transcription Loop-Mediated Isothermal Amplification Enabling Bright, Single-Step, Closed-Tube, and Multiplexed Detection of RNA Viruses. *Anal Chem* **88,** 3562–8 (2016).

108. Bloom, J. S. *et al.* Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. *medRxiv,* 2020.08.04.20167874 (2020).

109. Yelagandula, R. *et al.* SARSeq, a robust and highly multiplexed NGS assay for parallel detection of SARS-CoV2 and other respiratory infections. *medRxiv,* 2020.10.28.20217778 (2020).

110. Aynaud, M.-M. *et al.* A Multiplexed, Next Generation Sequencing Platform for High-Throughput Detection of SARS-CoV-2. *medRxiv,* 2020.10.15.20212712 (2020).

111. Chappleboim, A. *et al.* ApharSeq: An Extraction-free Early-Pooling Protocol for Massively Multiplexed SARS-CoV-2 Detection. *medRxiv,* 2020.08.08.20170746 (2020).

112. James, P. *et al.* LamPORE: rapid, accurate and highly scalable molecular screening for SARS-CoV-2 infection, based on nanopore sequencing. *medRxiv,* 2020.08.07.20161737 (2020).

113. Dao Thi, V. L. *et al.* A colorimetric RT-LAMP assay and LAMP-sequencing for detecting SARS-CoV-2 RNA in clinical samples. *Sci Transl Med* **12** (2020).

114. Schmid-Burgk, J. L. *et al.* LAMP-Seq: Population-Scale COVID-19 Diagnostics Using a Compressed Barcode Space. *bioRxiv,* 2020.04.06.025635 (2020).

2. HIGH-RESOLUTION MAPPING OF RNA-BINDING REGIONS IN THE NUCLEAR PROTEOME OF EMBRYONIC STEM CELLS

This chapter is adapted from a published manuscript:

Chongsheng He, Simone Sidoli, Robert Warneford-Thomson, Deirdre C. Tatomer, Jeremy E. Wilusz, Benjamin A. Garcia, and Roberto Bonasio.* High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells. *Molecular Cell.* 2016.
*corresponding author

2.1 Abstract

Interactions between non-coding RNAs and chromatin proteins play important roles in gene regulation, but the molecular details of most of these interactions are unknown. Using protein-RNA photo-crosslinking and mass spectrometry on embryonic stem cell nuclei, we identified and mapped, at peptide resolution, the RNA-binding regions in 800 known and previously unknown RNA-binding proteins, many of which are transcriptional regulators and chromatin modifiers. In addition to known RNA-binding motifs, we detected several protein domains previously unknown to function in RNA recognition, as well as non-annotated and/or disordered regions, suggesting that many functional protein-RNA contacts remain unexplored. We identified RNA-binding regions in several chromatin regulators, including TET2, and validated their ability to bind RNA. Thus, proteomic identification of RNA-binding regions (RBR-ID) is a powerful tool to map protein-RNA interactions and will allow rational design of mutants to dissect their function at a mechanistic level.

2.2 Introduction

In addition to their central roles as messengers and translators of genetic information, RNA molecules have key roles in gene regulation [1, 2]. Nowhere is this more evident than in the nucleus of mammalian cells, where many classes of poorly understood non-coding RNAs (ncRNAs) continue to be discovered [3–6].

Although some RNAs catalyze chemical reactions, they usually require association with proteins to function properly. Thus, it is reasonable to assume that the thousands of ncRNAs whose biochemical and biological roles are largely unknown exert their functions via protein-RNA interactions. Identifying proteins that interact with a given ncRNA has become a successful strategy to begin to dissect its biological roles; for example, the identification of Xist-associated proteins has provided important advances in understanding how this long ncRNA (lncRNA) controls X chromosome inactivation [7–9].

To identify RNA-binding proteins (RBPs) in a more general, unbiased way, multiple groups have used polyA+ selection followed by mass spectrometry (MS). These studies identified hundreds of previously unknown RBPs bound to mRNAs in human cell lines [10–13] and mouse embryonic stem cells (ESCs) [14]. However, most small RNAs and many lncRNAs are not polyadenylated, including abundant nuclear RNAs like MALAT1 [15, 16], enhancer-derived RNAs (eRNAs) [17], and circular RNAs [18]. Proteins interacting with these and other polyA− ncRNAs have thus been missed by existing approaches.

The identification of a protein as being RNA-associated is only the first step toward understanding the role of RNA interactions in its biochemistry. Mapping RNA-binding residues allows for the rational design of mutants to study the functional relationship between the protein and its cognate RNAs [19, 20]. Prediction of RNA-binding regions (RBRs) within RBPs is facilitated by the existence of well-characterized structural motifs that function as conserved RNA-binding domains (RBDs). The distinction between these two terms is important for this study: we refer to "RBRs" as minimal protein regions that make direct physical contacts with RNA [19], whereas "RBDs" are well-known, conserved domains that can be predicted from the primary sequence and typically function as RNA binders [21]. Examples of the latter category are the RNA-recognition motif (RRM) [22], the hnRNPK-homology domain (KH) [23], and the double-stranded RNA-binding domain (dsRBD) [24].

Until recently, it was widely believed that RBPs would contain one or more known RBDs and that the protein-RNA interaction could be assumed to take place within these domains. However, this simple concept has been challenged as the number of "non-canonical" RBPs (proteins that bind RNA without containing a classical RBD) continue to increase, especially for proteins that interface with chromatin and non-coding RNAs [25]. Because the RBRs of these proteins cannot be predicted a priori, we and others have resorted to various biochemical methods to identify them with candidate-based, low- throughput methods [19, 20, 26, 27].

Here, we report a high-throughput approach that exploits protein-RNA photocrosslinking and quantitative MS to identify proteins and protein regions interacting with RNA in vivo, regardless of the RNA polyadenylation status. As this approach not only identifies RNA- binding proteins, but also their respective RNA-binding regions, we named the technique RBR-ID. We applied RBR-ID to nuclei from mouse ESCs and identified RBRs within 803 proteins, more than half of which had not previously been reported as RBPs. We validated six RBRs, two in known RBPs whose mode of interaction with RNA was unknown and four in chromatin-associated proteins that had not been previously shown to bind RNA. Rational mutant design informed by RBR-ID nearly abolished RNA binding in vivo for these proteins, demonstrating the predictive power and practical utility of our technique for characterizing functional protein-RNA interactions.

2.3 Results

2.3.1 Development and optimization of RBR-ID

UV-mediated protein-RNA photocrosslinking generates adducts of RBPs with the covalent attachment localized at or near the site of physical interaction because of the short range of this type of crosslinking [28]. Thus, it should be possible to detect the RBR of a protein by MS, as an RNA-crosslinked peptide would have a different mass, causing the intensity of the signal for the non-crosslinked peptide to be lower in the irradiated sample (Figure 2.1 A).

Comparing mass spectra of UV-irradiated ESCs pulsed or not with 4-thiouridine (4SU), a uridine analog selectively activated by long-wavelength UV [29, 30], we observed that most peptides were unchanged in intensity (Figure 2.1 B, black lines) but some were depleted in the 4SU-treated samples; for example, peptide 74–89 from HNRNPC (Figure 2.1 B, red lines). We performed the experiment in three biological replicates, each acquired in duplicate MS runs, and noticed that the same HNRNPC peptide was consistently depleted by more than 50% (Figure 2.1 C), suggesting that 4SU incorporation and UVB-mediated crosslinking had caused a fraction of these peptides to change mass and thus not be counted toward the peak intensity of the non-crosslinked peptide. As HNRNPC is a well-known RBP [31] and the HRNPC74–89 peptide overlaps its RRM, we concluded that this analysis had the potential to reveal protein-RNA contacts in the entire proteome.

There are different ways to crosslink RNA to proteins. Conventional UV crosslinking exploits the excitation peak of natural nucleotides in the short-wavelength UVC range (254 nm) [32, 33], whereas incorporation of 4SU allows for more selective and less damaging crosslinking, typically using 365 nm UV (UVA). We previously showed that some protein-RNA interactions can only be captured with 4SU-aided crosslinking when an intermediate wavelength of 312 nm (UVB) is used [34]. We irradiated ESCs with the three different wavelengths and compared mass spectra obtained from isolated nuclei with those obtained from non-crosslinked samples (no 4SU treatment for 312 and 365 nm UV; no UV irradiation for 254 nm). Irradiation with 312 nm yielded the best compromise between sensitivity and specificity (Figure 2.1 D,E); 254 nm UVC yielded a large number of peptides with decreased intensities but with no preference for peptides overlapping known RRMs (Figure 2.1 D, blue dots); whereas, 365 nm UVA were too weak to consistently deplete a large number of peptides. For example, the RNA-binding peptide HNRNPC74–89 was significantly depleted only upon irradiation at 312 nm (Figure 2.1 D, red dot), similar to the RNA-binding peptides from SNRNP70, SPEN, and HNRNPM (Figure 2.7 A), three other well-known RBPs. Consistent with this, 254 nm UVC identified more candidates compared to 312 nm UVB (Figure 2.1 E), but a smaller fraction of them were annotated as RBPs, suggesting that the increased sensitivity came at the cost of decreased specificity. Crosslinking with 365 nm UVA resulted in more accurate identification (46% versus 40% of proteins identified were RBPs) than 312 nm UVB but with a considerable loss in sensitivity (Figure 2.1 E). Overall, 312 nm UVB crosslinking identified a larger fraction of all known RBPs (Figure 2.1 F), whether from previous empirically determined lists from HeLa, HEK293, or mouse ESCs, or from digital annotations such as the GO and Toronto RBP databases [35]. These proteome-wide observations were consistent with the higher efficiency of 4SU-dependent protein-RNA crosslinking, as measured by RNA pull-down followed by western blot for the U1SNRNP70 complex (Figure 2.7 B).

There was no correlation between the depletion of peptides by 4SU after UV crosslinking and depletion of peptides by 4SU alone (Figure 2.7 C), suggesting that changes in protein isoform representation or post-translational modification in response to the 4SU treatment could not explain the bulk of depletion observed after UV. Furthermore, although some peptides showed an increase in apparent abundance upon 4SU crosslinking (Figure 2.1 D), the majority of significant UV-induced changes were toward depletion in +4SU conditions (Figure 2.7 D).

We conclude that RBR-ID can identify known and unknown RBPs and that comparison of 4SU-treated versus untreated samples after irradiation with 312 nm UVB is the best compromise between sensitivity and specificity.

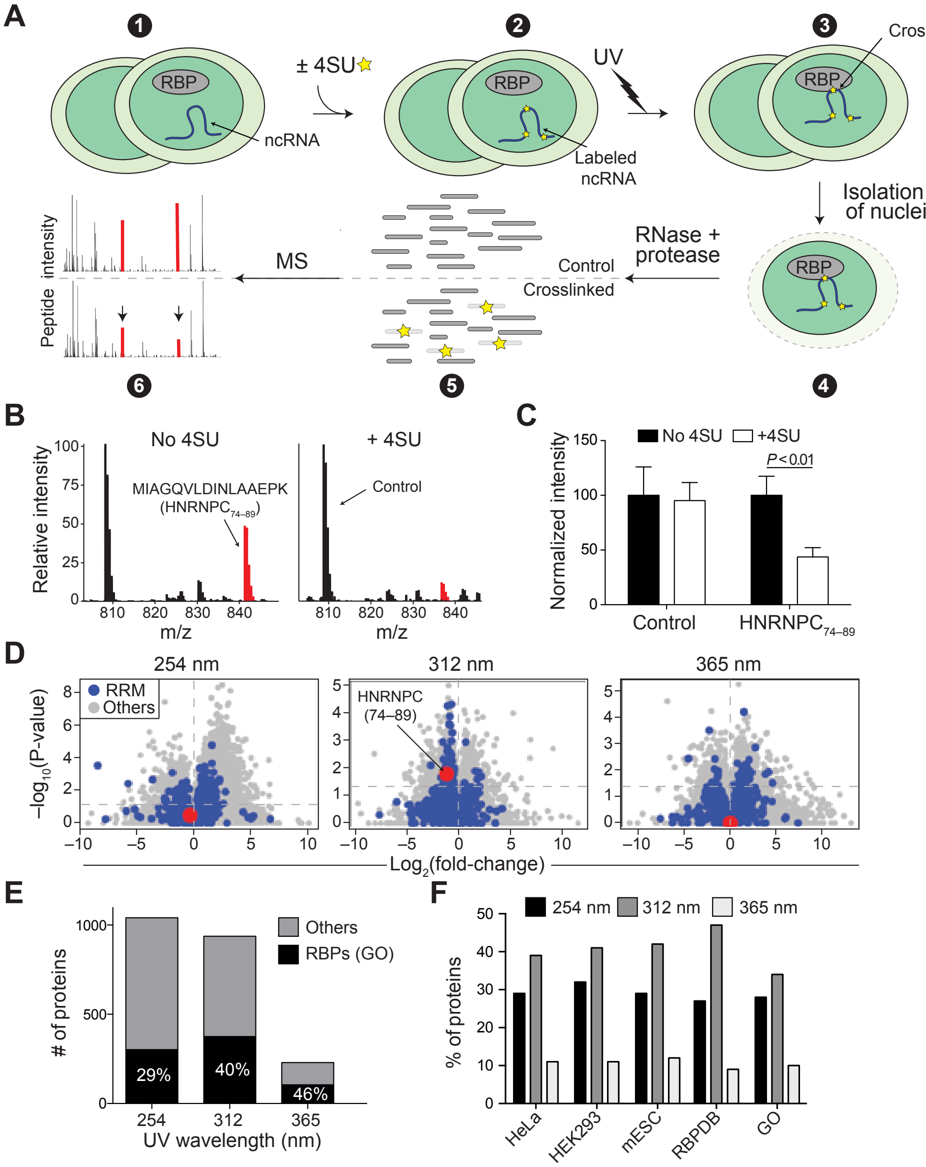


FIGURE 2.1 Development and Optimization of RBR-ID

(A) Mouse ESCs were pulsed with 4SU or not treated (1 and 2) and irradiated with different UV wavelengths (3). We isolated nuclei (4) and digested the crosslinked extracts with protease and RNase, producing a mixture of crosslinked and uncrosslinked peptides (5). Covalent crosslinks to RNA alters the peptide mass, and the mass spectrum of the corresponding uncrosslinked peptide decreases in intensity (6; red peaks). (B) Example averaged spectra from comparable retention time windows from untreated (left) and 4SU-treated ESCs (right). UV (312 nm) crosslinking caused decreased intensity of the highlighted spectrum in the 4SU sample. (C) Quantification of the extracted chromatogram for the control and HNRNPC peptides highlighted in (B). Bars indicate the average of the peak intensities normalized to the untreated sample (no 4SU) in six replicates + SEM. (D) Volcano plots showing log-fold changes in peptide intensities on the x axis and p values on the y axis for ±UV (254 nm) and ±4SU (312 and 365 nm). Peptides overlapping annotated RRM domains are in blue. The RNA-binding peptide from HNRNPC is highlighted in red. (E) Number of proteins with consistently ($p < 0.05$) depleted peptides and annotated as RBPs in the GO database (black) or not (gray). (F) Percentage of known RBPs according to the indicated studies and databases that were identified using different UV wavelengths for the crosslink. Experiments and Data generated by C.S.H. and S.S.

2.3.2 Protein-level analyses

To increase the confidence in peptide quantification, we acquired two technical replicates each for two additional biological replicates of 312 nm UVB irradiation ± 4SU. Despite the noise in each individual run, once aggregated, the first set (three replicates) and second set (two replicates) of RBR-ID results were consistent (Figure 2.2 A), suggesting that high replication could reduce artifactual identification of RBRs due to fluctuations in the MS signal.

In total, we detected 75,441 unique peptides from 4,929 proteins in mouse ESC nuclei; of these, 1,475 were consistently ($p < 0.05$) depleted by 4SU and UV, but not by 4SU alone (Table S1). These peptides belonged to 814 proteins (corresponding to 803 unique protein symbols), which we considered "primary hits" (Table S2). An additional set of 721 proteins identified with relaxed requirement ($0.05 < p < 0.1$) was used for some of the subsequent analyses and, along with the primary hits, constitutes our "extended" set (Table S2). GO annotations for the primary hits were enriched for functional terms related to RNA metabolism and function, including "RNA binding" (Figure 2.2 B; Table S3). Primary hits also showed large overlaps with a variety of existing RBP lists (Figure 2.8 A), either empirically determined [10–14] or digitally annotated [35, 36]. Based on these lists, 376 of the 803 primary candidates were previously known RBPs (Figure 2.2 C), a significant overlap ($p < 10^{-43}$, hypergeometric distribution). Among the previously known RBPs that were not recovered by RBR-ID, a large proportion (40%, Figure 2.8 A, compare bottom left with bottom right) could not be detected at all in ESC nuclei, likely because they were not expressed or were localized to the cytoplasm, as shown by their enrichment for ribosomal proteins and translation factors (Table S4, left). Nonetheless, 865 previously known RBPs were detectable in the ESC nuclear fraction and not recovered by RBR-ID (Figure 2.2 C). This set of proteins was also enriched for ribosomal biogenesis and translation-related GO terms (Table S4, right), suggesting that some might be present in the nucleus but only bind RNA in the cytosol. It is also possible that a substantial number of true nuclear RBPs cannot be crosslinked efficiently to 4SU.

We then turned our attention to the 427 previously unknown RBPs that were identified by RBR-ID. Even when considering only the detectable nuclear proteome as background, these non-canonical RBPs were enriched for GO terms related to gene regulation and chromatin biology

(Figure 2.2 D; Table S5), consistent with the notion that many chromatin-associated proteins bind RNA [25, 37]. Non-canonical RBPs identified by RBR-ID also contained different types of protein domains. The list of primary hits as a whole was enriched in known RBDs (RRM and KH) and RNA helicase domains, DEAD and DEAH (Figure 2.2 E; Table S6), whereas the 427 unknown RBPs were enriched for chromatin-related domains, such as bromodomain and chromodomain, (Figure 2.2 F; Table S7), which bind acetylated and methylated histones, respectively [38], and the SNF2-related domain found in ATP-dependent chromatin remodelers [39].

To estimate the confidence of identification for these proteins, we calculated an RBR-ID "score" for each peptide that captured both the extent of depletion (i.e., the log-converted fold-change between 4SU-treated and non-treated cells) and the consistency across replicates (i.e., the p value for the depletion; see Experimental Procedures). The previously unknown 427 RBPs had a distribution of RBR-ID scores comparable to that of the known 376 RBPs recovered RBR-ID (Figure 2.2 G).

We validated the unknown RBPs by performing photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) using conventional 365 nm UVA [30]. We tested five candidate RBPs from the set of 427 previously unknown primary hits and four of them (RARG, CDKN2AIPNL, PCED1B, and PCGF2/MEL18) showed 4SU-dependent radioactive labeling (Figure 2.8 B) indicative of RNA binding in vivo. A fifth one (CCDC115) was undetectable by PAR-CLIP, either because it was a false positive or because epitope tagging and overexpression interfered with its RNA binding activity. We also confirmed that NANOG, which was in the extended set, crosslinked to RNA in vivo (Figure 2.8 C), indicating that these additional candidates might also comprise previously unknown RBPs.

Thus, RBR-ID identified a considerable portion of previously known nuclear RBPs and at least 427 unknown, non-canonical RBPs enriched for GO terms and protein domains related to chromatin function.

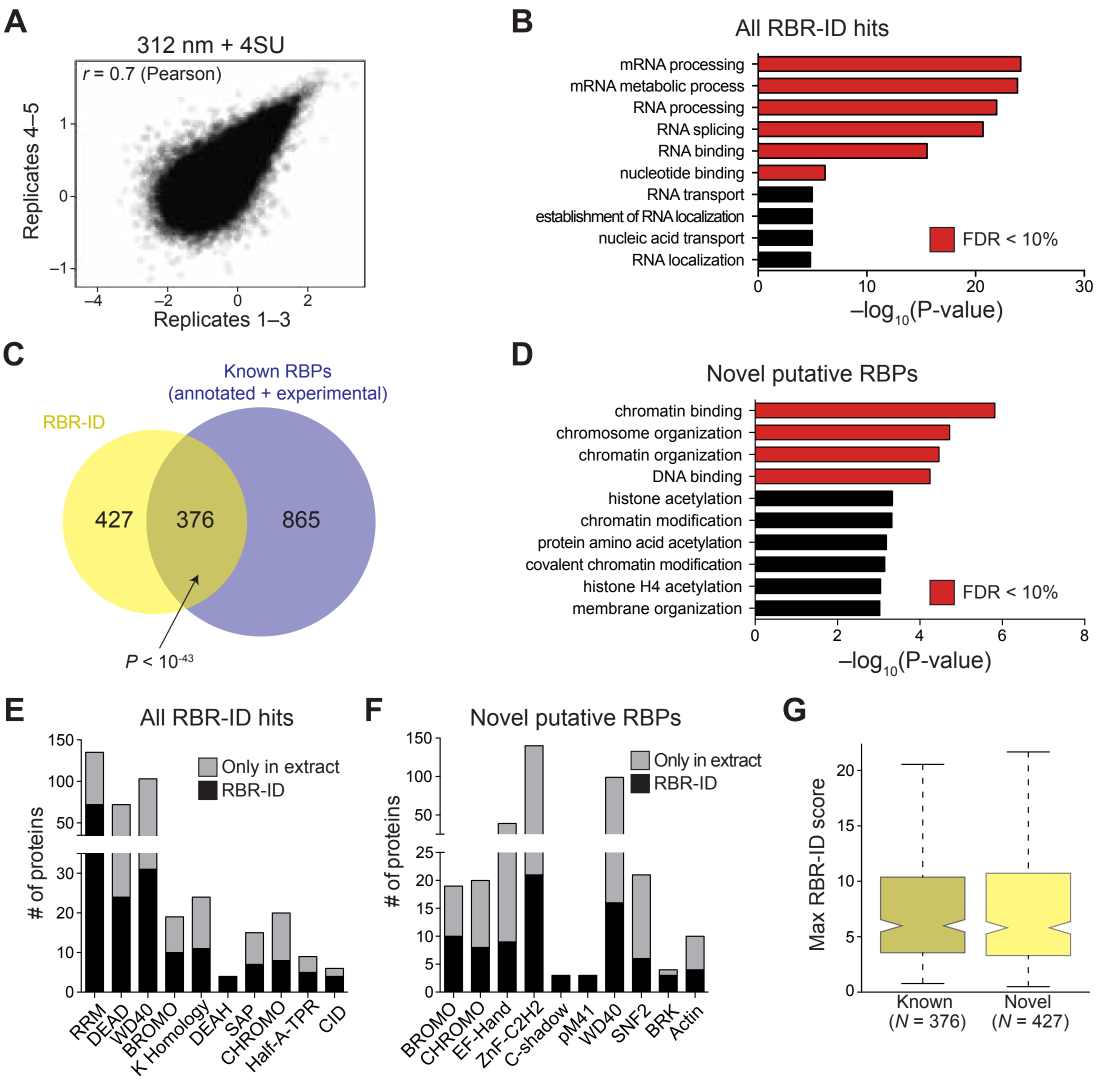


FIGURE 2.2 Protein-Level Analyses of Proteins Identified by RBR-ID

(A) Scatterplot showing log-converted and normalized average intensities for peptides from biological replicates 1–3 and the additional replicates 4 and 5. (B) Top ten enriched GO terms (biological process and molecular function) for primary RBR-ID protein hits. p values are plotted on the x axis, and terms with false discovery rate (FDR) < 10% are shown in red. (C) Overlap of RBR-ID protein hits and all known RBPs, both experimentally identified and annotated in databases. p value is from the hypergeometric distribution.(D) Top ten enriched GO terms as in (B) but only for the RBR-ID protein hits not found in the set of already known RBPs. (E and F) Top ten non-redundant protein domains enriched in the primary RBR-ID protein hits (E) or only in the unknown RBP set (F). The black section of the stacked bar plots indicates the number of proteins containing the domain and found in the primary RBR-ID candidate list; the gray section indicates the number of proteins not in the RBR-ID list but detected by MS in the ESC nuclear extract. (G) Tukey boxplot for the distribution of maximum RBR-ID scores per protein comparing known RBPs and unknown putative RBPs from (C). See also Digital Supplemental Tables S1, S2, S3, S4, S5, S6, and S7 and Figure 2.8. Experiments and Data generated by C.S.H. and S.S.

2.3.3 Known and unknown RBPs with chromatin-related function

Confirming the GO and domain enrichment analysis, visual inspection of the primary list of proteins identified by RBR-ID revealed many with chromatin-related functions whose moonlighting RNA-binding activities have been reported by candidate-based approaches (Table 1) but were missed by previous unbiased RBP identification endeavors. This included EZH2, the catalytic subunit of Polycomb repressive complex-2 (PRC2), which is responsible for formation of facultative heterochromatin and interacts with lncRNA [26, 40–42] and nascent transcripts [34]. We also recovered SUZ12, another subunit of PRC2 that binds RNA [43, 44], and HP1, a central component of constitutive heterochromatin and known RNA binder [45], as well as four chromatin factors whose binding to RNA was only recently reported: CTCF, ATRX, HDAC1, and DNMT3 [27, 46–50].

Among the many candidate chromatin proteins identified by RBR-ID that have not previously been reported to bind RNA, we noted TET1 and TET2, two methylcytosine oxidases required for the epigenetic process of DNA demethylation [51].

Name	**Functions**	**References**
ATRX	Chromatin remodeling	Sarma *et al.* [49]
CBX1/3/5 (HP1 α/β/γ)	Heterochromatin binding	Maison *et al.* [45] and Muchardt *et al.* [52]
CTCF	Chromatin organization	Saldaña-Meyer *et al.* [27], Kung *et al.* [48], and Sun *et al.* [50]
DNMT3A	DNA methylation	Holz-Schietinger & Reich [47]
EZH2	Histone methylation	Kaneko *et al.* [26], Rinn *et al.* [40], Tsai *et al.* [41], and Zhao *et al.* [42]
HDAC1	Histone deacetylation	Castellanos-Rubio *et al.* [46]
SUZ12	Histone methylation	Beltran *et al.* [43] and Kanhere *et al.* [44]
TET1	DNA demethylation	-
TET2	DNA demethylation	-

TABLE 2.1 Examples of Chromatin Factors Identified as RBPs

2.3.4 High-resolution mapping of RNA-interacting residues *in vitro*

We reasoned that the real power of RBR-ID would rely in its ability to identify not only RNA-binding proteins, but also their RNA-binding regions. We first sought to characterize a well-defined protein-RNA interaction in a fully reconstituted system. We chose the phage MS2 coat protein (MS2-CP) and its cognate stem-loop RNA (MS2-SL), a well-known protein-RNA pair with several high-resolution crystal structures available [53–55].

We incubated recombinant MS2-CP and MS2-SL RNA transcribed *in vitro* in the presence or absence of 4SU, subjected the complexes to UVB irradiation, and analyzed the crosslinks using MS (Figure 2.3 A). Incorporation of 4SU did not affect the ability of the coat protein to interact with the RNA (Figure 2.9 A). Similar to what we had observed in vivo (Figure 2.7 B), UVB were more efficient than UVA, although at the cost of some low-level background crosslinking even in absence of 4SU (Figure 2.3 B; Figure 2.9 B). Analysis of extracted ion chromatograms revealed a subset of peptides whose intensity was decreased in the 4SU sample (Figure 2.3 C). We generated three biological replicates for this *in vitro* RBR-ID assay and acquired them in technical duplicates. The most consistently depleted peptide corresponded to the 57–66 region of MS2-CP (Figure 2.3 D), which contains several residues known to form hydrogen bonds with RNA [54, 55].

Next, we calculated RBR-ID scores (combining extent and consistency of depletion) for each residue and plotted them along the primary sequence of MS2-CP. The RBR-ID score was a good metric for protein-RNA crosslinking, as positive scores precisely mapped to the known RBR of the protein (Figure 2.3 E), with the peak corresponding to glutamic acid 63, which forms hydrogen bonds with a uridine at position -5 in the stem loop [55].

The availability of crystal structures for the MS2-CP–MS2-SL complex allowed us to visualize the RBR-ID score in a more direct and powerful way. We converted the scores into a heat-map and used it to color the surface of the MS2 protein from the crystal structure [55], which revealed that the highest RBR-ID scores mapped to the pocket where most RNA contacts occur (Figure 2.4 A).

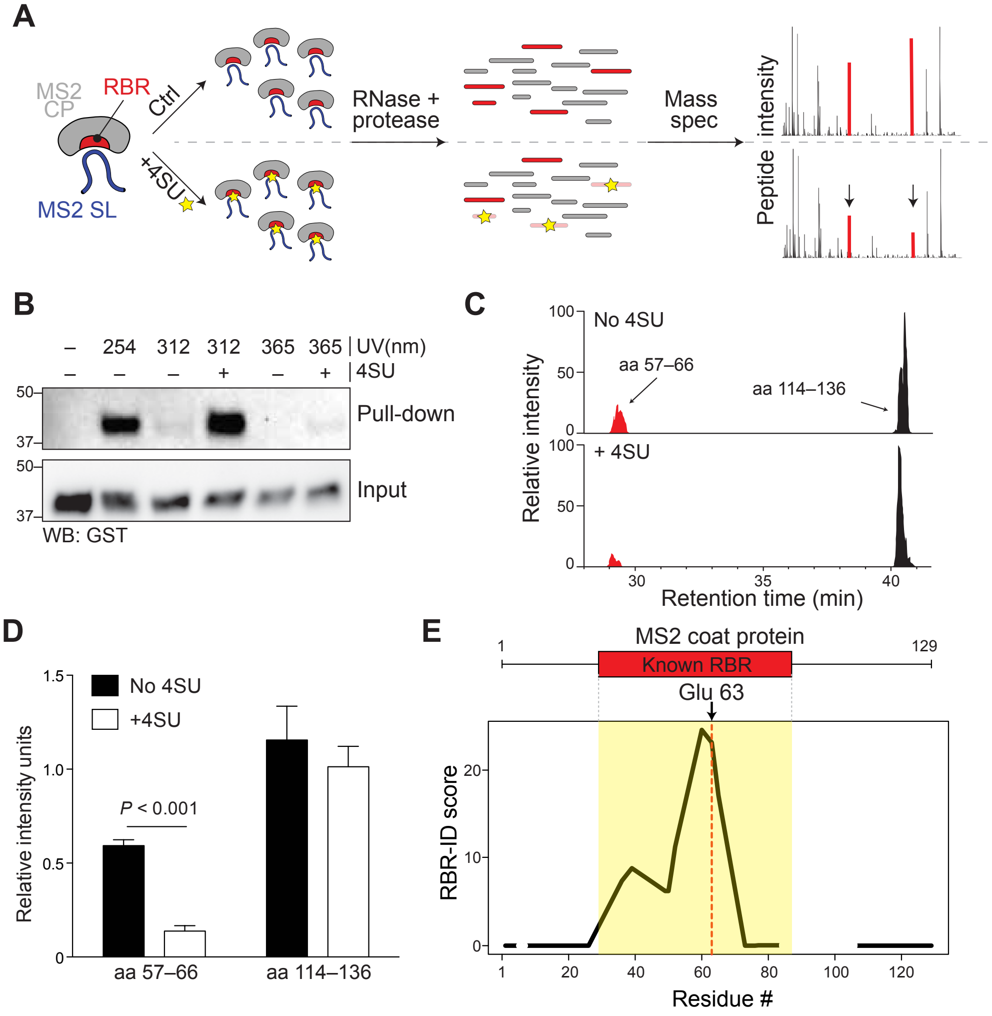


FIGURE 2.3 Mapping of the RBR for MS2-CP with RBR-ID In Vitro

(A) Recombinant MS2 coat protein and in vitro-transcribed stem-loop RNA with or without incorporated 4SU were allowed to form a complex, then crosslinked, digested, and analyzed by mass spectrometry. (B) Pull-down of MS2-SL RNA with or without 4SU and crosslinked to MS2-CP with different UV wavelengths. MS2-CP was detected via its fusion tag, GST. (C) Extracted ion chromatogram showing the elution profile of an RBR-overlapping peptide (red) and a peptide from an MS2-CP region that does not bind RNA (black) from MS2-CP crosslinked to natural (top) or 4SU-containing (bottom) MS2-SL RNA using 312 nm UV. (D) Quantification of peak intensities for the two peptides shown in (C) for three biological replicates each acquired in duplicates. Bars show average intensity + SEM. (E) Averaged and smoothed residue-level RBR-ID scores plotted along the primary sequence of MS2-CP. Regions with no peptide coverage are shown as gaps. Data are from three biological replicates each acquired in duplicates. Position of the known RBR and the uridine-interacting glu 63 residue are shown. See also Digital Supplemental Table S8 and Figure 2.9. Experiments and Data generated by C.S.H. and S.S.

2.3.5 Mapping of RBRs *in vivo*

We returned to our in vivo RBR-ID dataset from ESCs and assessed its precision in mapping RNA-interacting residues in known protein-RNA complexes in vivo at a proteome-wide level. We analyzed the subunits of the U1 small nuclear ribonucleoprotein (snRNP), a component of the spliceosome for which a high-resolution crystal structure was obtained [56]. The mouse U1 snRNP is composed of a polyA− ncRNA, U1, and ten protein subunits; we recovered four in the primary RBR-ID candidate list and four more in the extended list (Figure 2.4 B).

We calculated the single-residue RBR-ID scores for the identified subunits and used them to color the respective regions of the crystal structure. For the U1-70K subunit, our approach correctly identified two primary sites of RNA interactions, one within the conserved RRM that caps stem-loop I of the U1 RNA (Figure 2.10 A) and another within the stretch of residues that wraps around the ring formed by the Sm subunits to reach U1-C (Figure 2.4 C). Here, the highest RBR-ID scores were directly adjacent to uridine 137, which forms hydrogen bond contacts with the protein (Figure 2.4 C).

To determine whether spatial RBR mapping was also accurate for proteins in the extended RBR-ID candidate list, we analyzed the SmD2 subunit of the U1 snRNP particle (Figure 2.4 B), which contacts uridine 131 within U1 using H62 and N64. Even for this protein from the extended list, RBR-ID mapped the interacting region with great accuracy, with a peak in signal at the site of interactions as seen in the crystal structure (Figure 2.4 D). Similarly, the highest scores for SmB were near histidine 37, which interacts with uridine 129 (Figure 2.10 B).

To further validate the power of RBR-ID to identify known RBPs in vivo, we analyzed the subunits of RNA polymerase I and II (Figure 2.4 E,F), protein complexes responsible for transcription of rRNA and mRNA, respectively [57]. The two large subunits that form opposite sides of the active center cleft were recovered in both cases, as well as several of the smaller subunits. Interestingly, we recovered both subunits forming the “stalk” structure of RNA pol II (Figure 2.4 E), which were previously shown to crosslink to RNA *in vitro* [58, 59]. We also recovered the corresponding protein subunit from RNA pol I, RPA43 (Figure 2.4 F), suggesting that interactions of the polymerase stalk with nascent RNA might be a conserved feature in these related complexes.

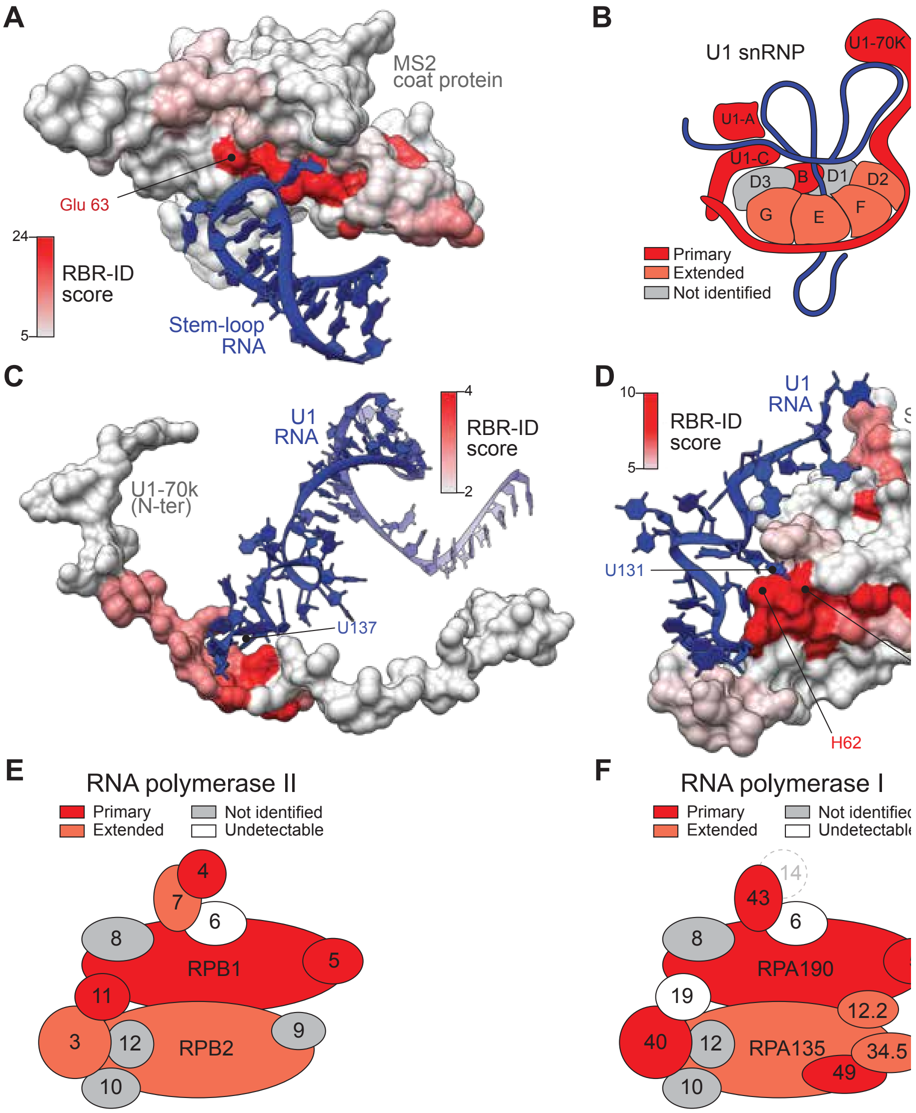


Figure 2.4 RBR-ID Maps the Sites of Protein-RNA Interactions In Vivo

(A) The surface rendering of the MS2 coat protein in complex with its cognate RNA (PDB: 1ZDI; Valegârd et al., 1997) was color coded according to the residue-level RBR-ID score from the experiment shown in Figure 2.3. (B) Schematic representation of the U1 snRNP particle (Kondo et al., 2015; Pomeranz Krummel et al., 2009). Subunits found in the primary list of RBR-ID candidates are in dark red; proteins in the extended list are in light red.(C and D) Zoomed-in regions of the crystal structure of U1 snRNP (PDB: 4PJO; Kondo et al., 2015) showing protein surfaces color coded according to their RBR-ID score and interacting RNAs for two regions of U1-70K (C) and SmD2 (D). (E and F) Schematic representation of the mammalian RNA pol II (E) and RNA pol I (F) complex according to Wild and Cramer (2012). Color coding is same as in (B). Subunits detected in the nuclear proteome, but not identified by RBR-ID, are in gray, undetectable subunits in white. The mammalian homolog for yeast RNA pol I subunit A14 (dashed circle) is unknown. See also Figure 2.10. Experiments and Data generated by C.S.H. and S.S.

2.3.6 Domain-level analyses on RBR-ID candidates

The mapping accuracy of RBR-ID was not restricted to the specific protein-RNA complexes discussed above, but extended to the entire proteome. Peptides overlapping RRM domains showed a strong bias for high RBR-ID scores compared to mock scores calculated from samples not irradiated with UV (Figure 2.5 A, left). Because the RRM domain is relatively frequent in mouse proteins, we also analyzed the distribution of RBR-ID scores for a control domain with similar frequency, the Ploop containing nucleoside triphosphate hydrolase (Interpro: IPR027417), which did not show the same bias (Figure 2.5 A, right). This demonstrated the selectivity of RBR-ID for a bona fide RBD in a proteome-wide manner.

Compared to all detected peptides, the primary list of RBR-ID peptides overlapped significantly with known RBDs but also contained a large proportion of peptides mapping to domains with no known RNA-related function (59%) or no domain annotations at all (23%, Figure 2.5 B). RBR-ID hits were enriched in peptides overlapping the three best- known RBDs—RRM, dsRBD, and KH [21]—as well as DEAD and DEAH RNA helicase domains (Figure 2.5 C). We also analyzed a list of non-classical RBDs [12] and found many of them enriched (Figure 2.5 D). In particular, the SAP domain, previously thought to mediate DNA binding [60], was enriched more than 5-fold compared to background. The reclassification of the SAP domain as a putative RBD was previously suggested based on its occurrence in empirically identified RBPs [12]. The enrichment of peptides overlapping the SAP domain in our RBR-ID candidate list provides direct evidence that this domain participates in RNA binding in vivo. Annotated domains enriched in the RBR-ID list, but not typically considered as possible RBDs (Figure 2.5 E), contained a few domains typically associated with chromatin-related functions, such as the high mobility group domain (HMG) and chromodomain, as well as domains known to participate in nuclear processes but whose function remain nebulous, such as DZF [61–63] and DUF1605 [64, 65].

Peptides that scored high in our RBR-ID screen, but could not be assigned to annotated domains, showed a slight tendency toward higher isoelectric points (Figure 2.5 F; Figure 2.11 A), consistent with a frequent role for positively charged amino acid in mediating direct interactions with RNA [66]. However, we saw no global correlation between the isoelectric point of a peptide and its RBR-ID score, excluding the possibility that crosslinking to RNA strongly favored patches of positive amino-acids in a non-specific fashion (Figure 2.11 B). Peptides identified by RBR-ID were also more likely to fall in protein regions predicted to be disordered (Figure 2.5 G; Figure 2.11 C), suggesting that in some cases the disordered regions might directly serve as RNA binding sites.

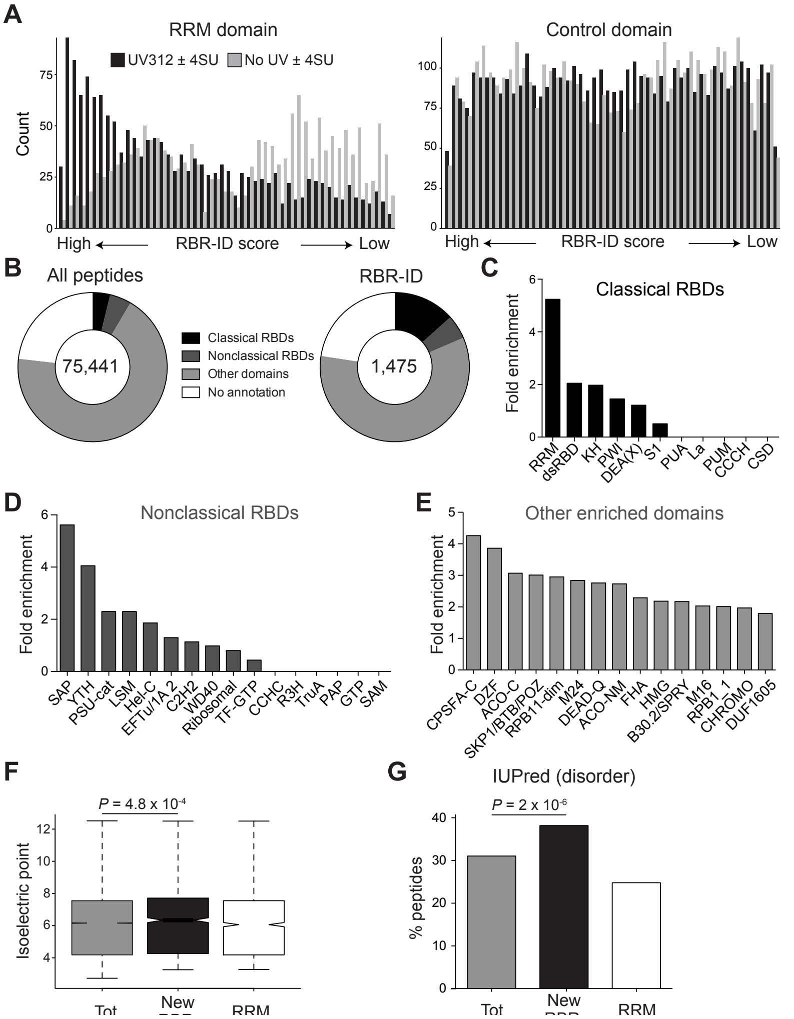


FIGURE 2.5 Known and Unknown RNA-Binding Regions in the ESC Proteome

(A) All detected peptides were sorted according to their RBR-ID score (UV312 ± 4SU) or a control score (no UV ± 4SU). The frequency of peptides overlapping the RRM domain (left) or a control, non-RNA binding domain (IPR027417, right) in these ranked lists is shown. (B) Categories of Interpro annotations for all peptides detected (left) or peptides in the primary list from RBR-ID (right).(C–E) Enrichment of selected domains in the top-tier RBR-ID peptides compared to the full list of detected peptides. Classical (C) and non-classical (D) RNA-binding domains are shown as well as enriched domains not previously reported to bind RNA (E). (F) Tukey boxplot of the isoelectric point for the indicated sets of peptides. p value is from a Student's t test. Tot, all detected peptides in the nuclear proteome; uRBRs, peptides in the primary candidate lists that did not overlap known RBDs; RRM, all detected peptides overlapping with the RRM domain. (G) Percentage of peptides overlapping with disordered regions from IUPred (Dosztányi et al., 2005; Oates et al., 2013). Values are shown for all detected peptides (tot), all top-tier RBR-ID peptides not mapping to a known RNA-binding domain (uRBRs), and all peptides overlapping RRM domains. p value is from a chi-square test. See also Figure 2.11. Experiments and Data generated by C.S.H. and S.S.

2.3.7 Validation of RBRs *in vivo*

To validate the RBRs predicted by RBR-ID, we selected proteins for which the RBR was previously unknown. We started with L1TD1, a protein whose RNA-binding activity had been previously reported but not mapped [14, 67]. The RBR- ID score plot pointed to a small region at residues 833–848 in the C terminus as a likely site for RNA interaction (Figure 2.6 A). We expressed epitope-tagged L1TD1 and a truncation mutant lacking the predicted RBR (ΔRBR) in HEK293 cells and performed PAR-CLIP using conventional 365 nm UVA [30]. We observed a radioactive signal that overlapped with the L1TD1 band (Figure 2.6 B) and could be assigned to protein-RNA crosslinks because its intensity was reduced after treatment with RNase A (Figure 2.6 C). Importantly, the mutant lacking the region predicted to interact with RNA by RBR-ID showed much lower PAR-CLIP signal despite equal expression levels and pull-down efficiencies for wild-type (WT) and mutant protein (Figure 2.6 B), suggesting that this region is a primary site of RNA interactions.

Next, we sought to validate a predicted RBR within a protein previously not known to interact with RNA. RBR-ID identified a nine-residue peptide adjacent to the catalytic domain of TET2 as the most likely site of RNA interaction (Figure 2.6 D). Indeed, a C- terminal fragment encompassing this predicted RBR was sufficient to bind to RNA *in vitro* (Figure 2.12 A) and in vivo (Figure 2.6 E,F), and the identified RBR was required for the interaction, as demonstrated by the drastically reduced PAR-CLIP signal in the ΔRBR mutant (Figure 2.6 E). We made similar observations for MYCN and its predicted RBR (Figure 2.12 B).

To validate additional candidate RBRs with a crosslinking-independent method, we switched to a native RNA immunoprecipitation assay [19]. Although lack of crosslinking renders this technique more prone to non-specific interactions, we reasoned that differences in RNA immunoprecipitation efficiency between WT and ΔRBR versions of the same protein would strongly suggest that the predicted RBR mediated binding to RNA.

Epitope-tagged versions of stem cell transcription factors POU5F1/OCT4 and NANOG as well as Polycomb protein MEL18 co-purified with RNA (Figure 2.12 C-E), and deletion of their predicted RBR impaired RNA binding (Figure 2.12 C-F), suggesting that the regions identified by RBR-ID were mainly responsible for RNA interactions.

Therefore, RBR-ID correctly identified six RBRs within two known (L1TD1 and OCT4) and four previously unknown (TET2, MYCN, MEL18, and NANOG) RBPs and guided the design of protein mutants that showed reduced RNA binding, demonstrating the validity of the predictions and the practical utility of RBR identification.

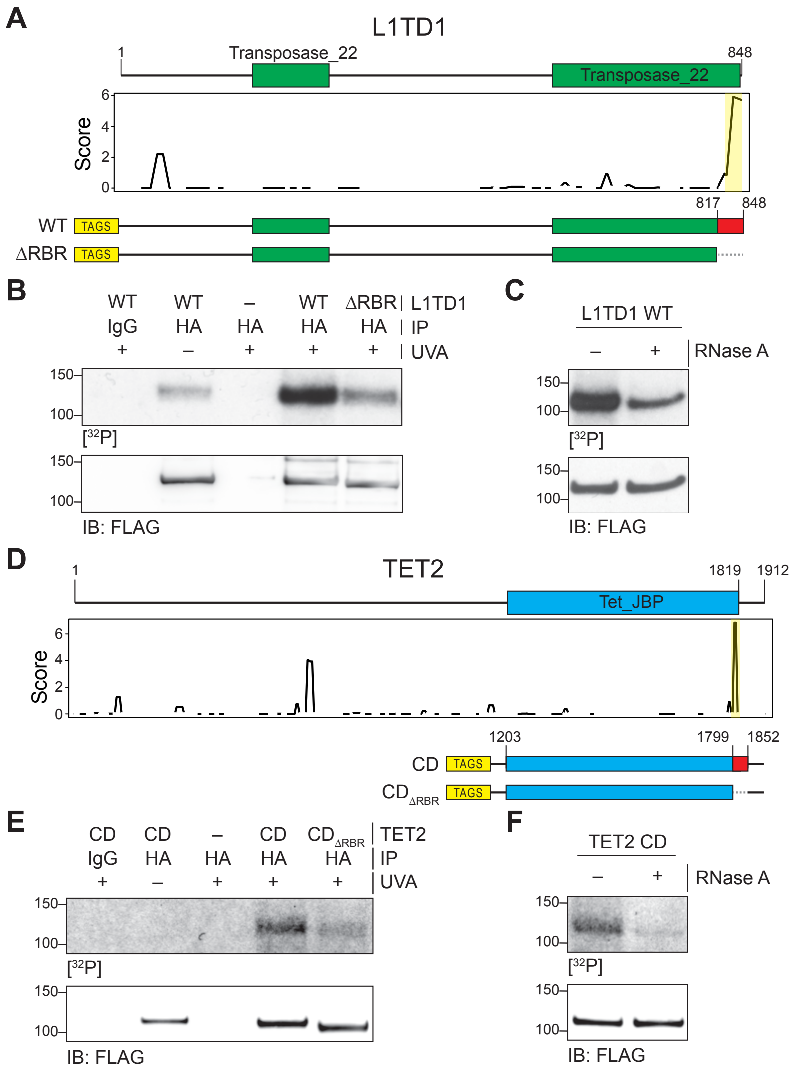


FIGURE 2.6 Validation of RBRs in L1TD1 and TET2

(A) Primary sequence and known domains for L1TD1 (top); smoothed residue-level RBR- ID score plotted along the primary sequence (middle); and scheme of epitope-tagged WT and RBR-deleted (ΔRBR) constructs used for validation (bottom). (B) PAR-CLIP of transiently expressed WT and ΔRBR L1TD1 in HEK293 cells. Autoradiography for 32P-labeled RNA (top) and control western blot (bottom). (C) PAR-CLIP for WT L1TD1 with and without treatment with RNase A (top) and control western blot (bottom). (D) Primary sequence and known domains for TET2 (top); smoothed residue-level RBR-ID score plotted along the primary sequence (middle); and scheme of epitope-tagged catalytic domain fragment (CD) and RBR-deleted (CDΔRBR) constructs used for validation (bottom). (E) PAR-CLIP of transiently expressed TET2-CD and TET2-CDΔRBR in HEK293 cells. Autoradiography for 32P-labeled RNA (top) and control western blot (bottom). (F) PAR-CLIP for TET2 CD with and without treatment with RNase A (top) and control western blot (bottom). See also Digital Supplemental Table S8 and Figure 2.12. Experiments and Data generated by C.S.H.

2.4 Discussion

Interactions with RNA constitute an important regulatory layer for the protein machinery that controls chromatin structure and gene expression. To obtain a mechanistic understanding of the biological and biochemical roles of these protein-RNA interactions, comprehensive lists of proteins bound to various classes of RNAs are needed, as well as detailed mapping of the protein regions involved. In vivo photocrosslinking followed by MS allows for the identification of hundreds of protein-RNA interactions in an unbiased manner and with peptide-level resolution.

2.4.1 Rationale for the development of RBR-ID

The identification of RBRs within non-canonical RBPs, such as Polycomb proteins SCML2 and JARID2, [19, 20], as well as CTCF [27], were important steps toward defining the biochemical roles of their interactions with RNA. Using ΔRBR mutants is particularly advantageous when the RBR of a given protein interacts with many RNAs so that depleting individual RNAs generally does not cause overt phenotypes. For example, a subset of the protein-RNA interactions within the PRC2 complex lack sequence specificity despite high affinities [68, 69], suggesting that the presence of any RNA, not a particular transcript, modulates the enzymatic activity of this complex [34, 70].

Mapping the RBRs of these non-canonical RBPs one at a time using recombinant protein fragments was a slow and labor-intensive strategy prone to *in vitro* artifacts. RBR-ID allowed us to identify the potential RBRs of 376 known and 427 unknown RBPs in ESC nuclei. These data will help focus future experiments on proteins and protein regions with the highest likelihood of forming protein-RNA contacts in vivo.

2.4.2 Advantages and limitations of RBR-ID

Previous endeavors to identify RBPs have relied on enrichment of complexes containing polyadenylated RNA [10, 12, 14]. Because of this experimental step, those approaches require up to 10^8–10^9 cells. RBR-ID can be performed with starting populations of 10^6 cells, making comparisons between cellular states (e.g., different differentiation trajectories) and studies in primary cells technically feasible.

Kramer et al. (2014) previously developed an MS pipeline capable, like RBR-ID, of assigning RNA binding sites within proteins based on UV crosslinking. They utilized their approach on human RBPs in a semi-artificial *in vitro* system, and even in those controlled conditions, crosslinks were identified in only 64 peptides from 49 proteins. This low number of RBPs was likely due to the difficulties in the positive identification of the complex mass spectra created by the heterogeneous products of protein-RNA crosslinking [71].

While our manuscript was being revised, Hentze and colleagues used a different technique to map RBRs in HeLa cells [72]. Their approach relies on two sequential oligo-dT pull-downs and therefore might have lower false positive rates than RBR-ID; however, it can only be used to identify RBPs that bind polyA+ RNA and requires 10–100 times more input material than RBR-ID.

The potential for false positives in RBR-ID should be curtailed by extensive replication, as was done for the experiments presented here. This is made possible by the low sample requirements, as only 2 µg of total nuclear protein were used per replicate. Even with replication, RBR-ID hits, as in any unbiased screen, will contain some false positives and therefore any candidate should be validated before pursuing the functional significance of its interactions with RNA. Identification by RBR-ID requires efficient protein-RNA crosslinks at a site of 4SU incorporation and therefore a substantial false negative rate is also to be expected, as shown by the missed identification of some known RBPs (Figure 2.2 C). This limitation could be mitigated in the future by utilizing other nucleotide analogs (e.g., 6-thio- guanine; [30]), different crosslink strategies, and/or more sensitive MS instruments.

2.4.3 Non-canonical RNA binding in chromatin proteins

Using RBR-ID, we identified 803 RBPs as well as their likely RBRs. Over 50% of these proteins were not present in previous lists from polyA+ RNA purifications or annotation databases. Among these are several chromatin proteins that have been identified by candidate-based approaches, such as EZH2, SUZ12, and CTCF [26, 40, 41, 44], but were missed in previous unbiased screens, either because they bind to polyA− ncRNAs or because the stoichiometry of their interactions with RNA is too low for pull-down purification. These 427 unknown nuclear RBPs were enriched for GO annotations related to chromatin structure, chromosome organization, and transcriptional regulation.

This observation lends further support to the idea that protein-RNA crosstalk plays a central role in epigenetic regulation [1, 2, 5, 25].

At the peptide level, several domains of interest were enriched, including the chromodomain, which was proposed as a potential RBD [73], before its role in recognizing lysine methylation was discovered [74]. Our RBR-ID data suggest that some chromodomains might indeed moonlight as RNA binders. A conspicuous number of putative RBRs map to protein regions that lack domain annotations. Although some of these might reflect incomplete annotation, the slight, but significant, enrichment of predicted disordered regions suggests that some of them might mediate RNA contacts, as in the case of FMRP and LAF-1 [75, 76]. This is particularly relevant in light of the prominent role of RBPs with disordered regions in disease [77].

2.4.4 Use of RBR-ID predictions

The validation of the RBR of TET2 provides an example of the utility of our RBR-ID dataset as a resource. In Drosophila, the TET2 homolog dTET is partially responsible for cytosine hydroxymethylation on RNA [78]. Although no RNA-binding evidence has been obtained for the Drosophila protein, the fact that mouse TETs can use RNA as a substrate *in vitro* [79] suggests that this function might be conserved in mammals. The presence of both TET1 and TET2 in the list of primary RBR-ID candidates strongly supports this hypothesis. Furthermore, the identification and validation of the TET2 RBR provides a useful starting point to study the biological role of this biochemical function for the TET family of epigenetic regulators.

2.4.5 Outlook and conclusion

We applied RBR-ID to ESC nuclei and identified hundreds of RBRs within proteins previously unknown to bind RNA. Because the approach is easily implemented and versatile, we anticipate that variations on this theme will provide even more comprehensive and precise lists of RBRs than the one presented here. Improvements on MS instrumentation and quantification methods, such as “Tandem Mass Tagging” [80], will increase sensitivity, and alternative photoactivatable nucleotides and protease treatments could expand the range of crosslinked peptides, improving resolution.

RBR mapping data are available at http://rbrid.bonasiolab.org. We anticipate that the community will find this a useful resource to design functional experiments aimed at decrypting the complex regulatory language of protein-RNA interactions on chromatin and elsewhere in cells.

2.5 Acknowledgments

The authors thank T. Christopher for technical support, R. Kohli for his generous gifts of TET2 reagents, and S. Berger and D. Reinberg for comments on the manuscript. R.B. was supported by the Searle Scholars Program, the W.W. Smith Foundation (C1404), the March of Dimes Foundation (1-FY-15-344). B.A.G acknowledges support from NIH grant R01GM110174 and DOD grant BC123187P1. J.E.W. is a Rita Allen Foundation Scholar and was supported by NIH grants R00GM104166 and R35GM119735. R.W.-T. was supported in part by NIH training grant T32GM008216.

2.6 Author contributions

C.H. and R.B. conceived the project and designed the experiments with help from S.S. and B.A.G. C.H. carried out all experimental work with help from R.W.-T. Specifically, R.W.-T. performed cloning and PAR-CLIP validations of RBP candidates and developed online RBR-ID web applet. S.S. analyzed samples by MS under the supervision of B.A.G. D.C.T. and J.E.W. provided critical reagents. C.H. and R.B. wrote the manuscript with input from all authors. Text in this chapter was adapted from manuscript by R.W.-T. with minor modifications.

2.7 Disclosures

The authors have no conflicts of interest to disclose.

2.8 Methods

2.8.1 RNA immunoprecipitation

Nuclear extracts were incubated with hemagglutinin (HA) antibody for 3 hr at 4°C and immunocomplexes recovered with protein G Dynabeads. Beads were washed in RIP-W buffer (20 mM Tris [pH 7.94° C], 1 mM MgCl2, 200 mM KCl, and 0.05% IGEPAL CA-630) twice and incubated with TURBO DNase to eliminate potential bridging effects of protein- DNA and DNA-RNA interactions. After two additional washes, RNA was eluted from the beads with TRIzol and purified. We quantified the RNA abundance after immunoprecipitations by measuring the intensity of the bands with ImageJ and normalizing to the IgG background.

2.8.2 PAR-CLIP

HEK293 cells were transiently transfected, pulsed with 100 µM 4-SU for 24 hr, crosslinked with 400 mJ/cm2 UVA (365 nm), and lysed in CLIP buffer (20 mM HEPES [pH 7.4], 5 mM EDTA, 150 mM NaCl, and 2% Empigen) with protease inhibitors, DNase, and RNase inhibitor. HA and StrepTag-fused proteins were first bound to StrepTactin beads in CLIP buffer for 3 hr at 4°C. Beads were washed five times using CLIP buffer and eluted with 2 mM biotin. Next, proteins were incubated with HA antibody overnight at 4°C and recovered with protein G Dynabeads. DNA was removed with DNase, and crosslinked RNA was dephosphorylated with Antarctic phosphatase and labeled with T4 PNK and [γ-32P] ATP. Labeled complexes were resolved on 4%–12% bis-tris gels, transferred to nitrocellulose membrane, and imaged. For NANOG PAR-CLIP, we used E14Tg2A (E14) ESCs pulsed with 500 µM 4-SU for 2 hr and crosslinked with 400 mJ/cm2 UVB (312 nm).

2.8.3 RBR-ID

Cells were pulsed with 500 µM 4SU for 2 hr and crosslinked with 1 J/cm2 UVA, 1 J/cm2 UVB, or 800 mJ/cm2 UVC. We verified that 2 hr was sufficient to incorporate 4SU in virtually all coding and non-coding transcripts by 4SU sequencing (data not shown). Cells were lysed in buffer A (10 mM Tris [pH 7.94° C], 1.5 mM MgCl2, 10 mM KCl, 0.5 mM DTT, and 0.2 mM PMSF) with 0.2% IGEPAL CA-630 for 5 min on ice to isolate nuclei, which were lysed in 9 M urea and 100 mM

Tris (pH 8RT). The lysate was diluted in 50 mM ammonium bicarbonate and reduced with 5 mM dithiothreitol for 45 min at 56°C. Cysteines were alkylated with 20 mM iodoacetamide for 30 min. Trypsinization was performed at an trypsin:sample ratio of 1:100 overnight at 37°C and blocked with 1% trifluoroacetic acid. Peptides were desalted, dried, and resuspended in 0.1% formic acid prior to MS analysis. Crosslinked RNA was removed with Benzonase.

2.8.4 GO and Interpro enrichment

For protein list comparisons, all proteins identifiers were converted to official mouse symbols using the Biomart database (version 84). One-to-one human-mouse orthologs were mapped directly, whereas one-to-many and many-to-many homologs were reduced to one- to-one by considering the protein with highest percentage of homology, according to the Biomart database (version 84). For GO and Interpro annotation, tables were downloaded from the Uniprot and Interpro websites directly. Enrichment values and statistics were obtained using the DAVID web server [81] either using the unique Uniprot accession identifiers or the converted symbols, when needed.

2.8.5 Cells

E14Tg2A.4 mESC lines (E14 mESCs) and HEK 293 cells were cultured as described previously [26]. KH2 ESCs expressing the reverse tetracycline-controlled transactivator (rtTA) [82] were maintained in standard mouse ESC (mESC) culture conditions. For RNA immunoprecipitation experiments, stable KH2 lines were generated by transfection of the relevant pINTAN3 constructs and selecting with 50 µg/ml Zeocin (InvivoGen, CA). Transgene expression was induced with 2 µg/ml doxycycline for 24 h.

2.8.6 Plasmids and sequences

The construction of the backbones for pGEX-6P1 was described previously [83]. The DNA sequence for MS2-CP was synthesized by IDT and subcloned into the pGEX-6P1 expression vector. pINTON3 vector was based on the pINTO system [84] containing three N-terminal epitope tags (FLAG, HA, and Twin-Strep-Tag). For expression inHEK 293 cells, we cloned into pINTON3 L1td1, Mycn, Rarg, Cdkn2aipnl, and Pced1b from mouse cDNA and the catalytic domain of Tet2 from a plasmid kindly provided by Rahul Kohli (University of Pennsylvania) [85]. Pou5f1, Pcgf1, Pcgf2, and Nanog were cloned from mouse cDNA into the pINTAN3 vector, which is based on the Tet-On 3G system (Clontech, CA) and encodes three different N-terminal epitope tags: Flag, HA, and Twin-Strep- Tag [34]. Truncations were obtained by PCR. All oligonucleotide and synthetic DNA sequences used are in Digital Supplemental Table S8.

2.8.7 Antibodies information

The following antibodies were used for Western blots: SNRNP70 (#sc-9571 Santa Cruz Biotechnologies, TX), GST (#sc-33613; Santa Cruz Biotechnologies), FLAG (#F1804; Sigma-Aldrich, MO), HA (#901501 BioLegend, CA). Antibody against HA (#ab9110 Abcam, UK) was used for PAR-CLIP experiment and RNA immunoprecipitation.

2.8.8 Recombinant protein expression and purification

GST fusion MS2-CP were induced with 0.1 mM IPTG and expressed in BL21(DE3) cells for 24 h at 16°C and purified using glutathione-sepharose 4 Fast Flow beads (GE Healthcare Life Sciences, PA). The beads were washed with PBS in a column, and the proteins were eluted in the presence of 10 mM glutathione. The purified proteins were dialyzed against 20 mM Tris pH 7.5, 100 mM KCl and 10% glycerol. Recombinant, FLAG-fused TET2-CD was kindly provided by Rahul Kohli [85].

2.8.9 In vitro RNA pull-down assays

RNA fragments were *in vitro* transcribed using the HiScribe kit (New England Biolabs, MA) and purified by TRIzol (Thermo Fisher). The 5'-terminal HOTAIR RNA fragment used was as previously described [19]; template information for MS2-SL and random 100 nts RNA fragments is in Digital Supplemental Table S8. For the binding assays, recombinant proteins were incubated with total E14 RNAs or *in vitro* transcripts in 1 ml RIP buffer (20 mM Tris pH 8, 0.2 mM EDTA, 100 mM KCl, 3 mM MgCl2, 0.05% IGEPAL CA-630) with the addition of 2 u/µl murine RNAse inhibitor (New England Biolabs) for 30 min at 4°C. Protein-RNA complexes were pulled down using glutathione-sepharose 4 Fast Flow beads (GE) or FlagM2 beads (Sigma). After three washes with RIP buffer, proteins were eluted from the beads in Laemmli sample buffer and nucleic acid with TRIzol. RNAs were resolved on polyacrylamide/urea gels and visualized with SYBR gold (Thermo Fisher).

2.8.10 RNA immunoprecipitation

Nuclear extracts were obtained using an established protocol [86] with minor modifications to minimize RNAse activity. Briefly, cells were washed with PBS and with Buffer A (10 mM Tris pH 7.94oC, 1.5 mM MgCl2, 10 mM KCl, protease inhibitors, phosphatase inhibitors) and lysed in Buffer A plus 0.2% IGEPAL CA-630 for 5 min on ice. Nuclei were isolated by centrifugation at 2,500g for 5 min and lysed in Buffer C (20 mM Tris pH 7.94ºC, 25% glycerol, 400 mM NaCl, 1.5 mM MgCl2, 10 mM EDTA, 0.4 u/µl murine RNAse inhibitor, protease inhibitors, phosphatase inhibitors) for 30 min at 4oC.

Lysates were cleared at 18,000g for 30 min then incubated with HA antibody for 3 h at 4oC. Immunocomplexes were recovered by adding 7 µl of protein G-coupled Dynabeads (Thermo Fisher) per µg of antibody used and incubating for 1 h at 4oC. Beads were washed in RIP-W buffer (20 mM Tris pH 7.94oC, 1 mM MgCl2, 200 mM KCl, 0.05% IGEPAL CA-630) twice and incubated with 2 u of TURBO DNase (Thermo Fisher) in 20 µl RIP-W buffer for 10 min at room temperature, to eliminate potential bridging effects of protein–DNA and DNA–RNA interactions. After two additional washes in RIP-W buffer RNA was eluted from the beads with TRIzol and collected by precipitation with isopropanol. Residual DNA was removed with TURBO DNAse for 20 min at 37oC.

To quantify the RNA abundance after immunoprecipitations, we measured the intensity of the smears using ImageJ and normalized to the background observed in the IgG pull-down.

2.8.11 PAR-CLIP

HEK 293 cells were transiently transfected with plasmids using Lipofectamine 3000 reagent (Thermo Fisher) and pulsed with 100 µM 4-SU (Sigma) for 24 h. Cells were crosslinked with 400 mJ/cm2 UVA (365 nm) using a Spectrolinker (Spectroline, NY) and lysed in CLIP buffer (20 mM HEPES pH 7.4, 5 mM EDTA, 150 mM NaCl, 2% Empigen) with protease inhibitors (Roche), 20 U/ml Turbo DNase (Thermo Fisher), and 200 U/ml murine RNase inhibitor (New England Biolabs).

Tagged proteins were first bound to BSA-blocked Strep-Tactin beads (IBA, Germany) in CLIP buffer for 3 h at 4°C. Beads were washed 5 times using CLIP buffer and eluted in CLIP buffer with 2 mM biotin (Sigma), protease inhibitors (Roche Life Science, IN), and murine RNase inhibitor (New England Biolabs). Eluted proteins were incubated with HA antibody (Abcam) in CLIP buffer overnight at 4°C. Immunocomplexes were recovered with protein G- coupled dynabeads for 45 min at 4°C. DNA was removed with TURBO DNase (2 U in 20 µl). Crosslinked RNA was labeled by incubations with 5 U Antarctic phosphatase and 5 U T4 PNK (both from New England Biolabs) in presence of 10 µCi [γ-32P] ATP (PerkinElmer, MA). Labeled material was resolved on 8% Bis-Tris gels, transferred to nitrocellulose membrane, and exposed to autoradiography films for 1–24 hr.

For NANOG PAR-CLIP, E14 ESCs were transiently transfected with plasmids using Lipofectamine 3000 reagent (Thermo Fisher) and pulsed with 500 µM 4SU (Sigma) for 2 h. Cells were crosslinked with 400 mJ/cm2 UVB (312 nm) using a Spectrolinker.

2.8.12 In vitro RBR-ID

GST-fused MS2-CP was incubated with MS2-SL RNA in binding buffer (1 mM ATP, 10 mM HEPES pH 7.2, 3 mM MgCl2, 5% glycerol, 1 mM DTT, 100 mM KCl) for 30 min at 4°C. RNA- protein complex were crosslinked with 1 J/cm2 UVB (312 nm) using a Spectrolinker. The complexes were treated with RNase A (1 µg/µl) for 30 min at 37°C and the protein digested with trypsin or chymotrypsin. For trypsin digestion, proteins were diluted in 50 mM ammonium bicarbonate (NH_4HCO_3, pH 8) and incubated with 5 mM dithiothreitol (DTT) for 45 min at 56°C for disulfide bond reduction. This was followed by 20 mM iodoacetamide (IAA) incubation for 30 min in the dark for alkylation of the free cysteines. Trypsin was then added at an enzyme:sample ratio of 1:20, overnight at 37°C. For chymotrypsin digestion, proteins were diluted in 100 mM Tris pH 8, 10 mM CaCl2, and incubated with 5 mM dithiothreitol (DTT) for 45 min at 56°C for disulfide bond reduction. This was followed by 20 mM iodoacetamide (IAA) incubation for 30 min in the dark for alkylation of the free cysteines. Samples were then digested using chymotrypsin at an enzyme:sample ratio of 1:20, overnight at 25°C. Reactions were blocked by adding 1% trifluoroacetic acid. Desalting was performed by using in-house packed Stage tips made of C18 material. Eluted peptides were dried and resuspended in 0.1% formic acid prior to nanoLC-MS analysis.

2.8.13 In vivo RBR-ID

Cells were pulsed with 500 µM 4SU (Sigma) for 2 h and then crosslinked with 1 J/cm2 UVA (365 nm), 1 J/cm2 UVB (312 nm), or 800 mJ/ cm2 UVC (254 nm) using a Spectrolinker. We verified that 2 hours was sufficient to incorporate 4SU in virtually all coding and non-coding transcripts by 4SU-sequencing (data not shown). Cells were lysed in Buffer A (10 mM Tris pH 8, 1.5 mM MgCl2, 10 mM KCl, 0.5 mM DTT, 0.2 mM PMSF) with 0.2% IGEPAL CA-630 for 5 min on ice to isolate nuclei. Nuclei were washed with Buffer A and lysed in denaturing lysis buffer (9 M urea, 100 mM Tris pH 8). Lysate was diluted in 50 mM ammonium bicarbonate (NH4HCO3, pH: 8.0) and incubated with 5 mM dithiothreitol (DTT) for 60 min at 25°C for disulfide bond reduction.

This was followed by 20 mM iodoacetamide (IAA) incubation for 30 min in the dark for alkylation of the free cysteines. Samples were then digested using trypsin at an enzyme:sample ratio of 1:100, overnight at 37°C. Reaction was blocked by adding 1% trifluoroacetic acid. Desalting was performed by using in-house packed Stage tips made of C18 material. Eluted peptides were dried and resuspended in 0.1% formic acid prior MS analysis. Crosslinked RNA was removed with Benzonase (250 U in 20 µl).

2.8.14 In vitro denaturing protein–RNA pull-downs

GST fusion MS2-CP were incubated with MS2-SL RNA in binding buffer (1 mM ATP, 10 mM HEPES pH 7.2, 3 mM MgCl2, 5% glycerol, 1 mM DTT, 100 mM KCl) for 30 min at 4°C. RNA-protein complex were crosslinked with 1 J/cm2 UVA (365 nm), 1 J/cm2 UVB (312 nm), or 800 mJ/ cm2 UVC (254 nm) using a Spectrolinker. Biotin-labeled DNA probes were hybridized to crosslinked RNA in hybridization buffer (500 mM NaCl, 1% SDS, 50 mM Bis-Tris pH6.7, 10 mM EDTA, 10% formamide) with protease inhibitor for 4 h at 37°C. DNA probes and protein–RNA complexes were recovered by incubating with streptavidin-conjugated dynabeads (Thermo Fisher) for 30 min at 37°C. After three washes with wash buffer (2X SSC, 0.5% SDS, 0.4 mM PMSF), proteins were eluted by PBS with 0.5 µg/µl RNase A.

2.8.15 In vivo denaturing protein–RNA pull-downs

Cells were pulsed with 4SU (Sigma) for 2 h and then crosslinked with 1 J/cm2 UVA (365 nm), 1 J/cm2 UVB (312 nm), or 800 mJ/ cm2 UVC (254 nm) using a Spectrolinker. Cells were lysed in lysis buffer (50 mM Bis-Tris pH 6.7, 10 mM EDTA, 1% SDS) with protease inhibitor. A further sonication step with a Bioruptor (Diagenode, NJ) was performed to homogenize the cell lysates. Biotin-labeled DNA probes (Digital Supplemental Table S8) were hybridized to crosslinked RNA in hybridization buffer (500 mM NaCl, 1% SDS, 50 mM Bis-Tris pH 6.7, 10 mM EDTA, 10% formamide) with protease inhibitor for 4 h at 37°C. DNA probes and protein–RNA complexes were recovered by incubating with streptavidin-conjugated dynabeads for 30 min at 37°C. After three washes with wash buffer (2X SSC, 0.5% SDS, 0.4 mM PMSF), proteins were eluted by PBS with 0.5 µg/µl RNase A.

2.8.16 Bottom-up nanoLC-MS/MS

Samples were analyzed by using a nanoLC-MS/MS setup. NanoLC was configured with a 75 µm ID x 17 cm Reprosil-Pur C18-AQ (3 µm; Dr. Maisch GmbH, Germany) nano-column using an EASY-nLC nanoHPLC (Thermo Fisher). The HPLC gradient was 0-30% solvent B (A = 0.1% formic acid; B = 95% acetonitrile, 0.1% formic acid) over 120 min for the nuclear proteome experiments and over 45 min for the recombinant protein (MS2-CP) analysis. The gradient proceeded from 30% to 85% solvent B in 5 minutes and 10 min isocratic at 85% B. The flow rate was set to 300 nL/min. NanoLC was coupled with an Orbitrap Fusion mass spectrometer (Thermo Fisher) for the proteome experiments or with an Orbitrap Elite (Thermo Fisher) for the single protein analysis. Spray voltage was set at 2.3 kV and capillary temperature was set at 275 °C. Full scan MS spectrum (m/z 350−1200) was performed in the Orbitrap with a resolution of 120,000 (at 200 m/z) with an AGC target of 5x10^5. For the proteome experiment in the Orbitrap Fusion the Top Speed MS/MS option was set to 2.5 sec, and the most intense ions above a threshold of 50,000 counts were selected for fragmentation. Fragmentation was performed with higher-energy collisional dissociation (HCD) with normalized collision energy of 32, an AGC target of 10^4 and a maximum injection time of 120 msec. For the single protein experiment in the Orbitrap Elite the top 10 most intense ions above a threshold of 10,000 counts were selected for fragmentation. Fragmentation was performed with collisional induced dissociation (CID) with normalized collision energy of 35, an AGC target of 10^4 and a maximum injection time of 150 msec. MS/MS data for both experiment types were collected in centroid mode in the ion trap mass analyzer (normal scan rate). Only charge states 2-5 were included.

2.8.17 MS analysis

All MS/MS spectra were processed through the MaxQuant program (Cox and Mann, 2008). Parameters for MS/MS database searching included the following: precursor mass tolerance 4.5 ppm; product mass tolerance 0.5 Da; enzyme trypsin; missed cleavages allowed 2; static modifications carbamidomethyl (C); variable modifications none; label-free quantification method iBAQ (for protein reports); database used was Mus musculus (Uniprot, September 2015, including not reviewed proteins) for the proteome searches, and a custom database including MS2-CP and RNase A for *in vitro* RBR-ID. PSMs and protein false discovery rate was filtered for < 0.01. Match between runs was enabled using a tolerance of 1 min to extend the peptide identification to MS signals not identified in some of the replicates.

2.8.18 RBR-ID analysis

For each peptide, the maximum intensity of the corresponding extracted chromatogram calculated by MaxQuant was considered and inter-run variability was accounted for by normalizing for the sum of all peptide intensities in each MS run. To calculate the extent of crosslinking-induced depletion we calculated the log2-converted ratio of the mean intensity of each peptide in the +4SU samples divided by the mean intensity of the same peptide in -4SU samples. The list of primary hits contains all peptides showing depletion (i.e. log2(fold-change) < 0 with a P-value < 0.05 (Student's t test). For the extended list we relaxed the P-value requirement to 0.1. For both the primary and extended list we removed peptides that passed the same cutoffs when comparing signals for +4SU and -4SU in absence of UV.

RBR-ID scores were calculated by combining the extent of depletion and the P-value according to the following formula:

$$\text{RBR-ID score} = \log_2\left(\frac{normalized{+}4SUintensity}{normalized{-}4SUintensity}\right) \times (\log_{10}(\text{P-value}))^2.$$

For residue level RBR-ID scores, we summed the RBR-ID score of each peptide overlapping any given amino acid and smoothed the resulting curve using Friedman's 'super smoother' [87].

2.8.19 GO and Interpro enrichment

For protein list comparisons, all proteins identifiers were converted to official mouse symbols using the Biomart database (version 84). One-to-one human-mouse orthologs were mapped directly, whereas one-to-many and many-to-many homologs were reduced to one-to-one by considering the protein with highest percentage of homology, according to the Biomart database (version 84). For GO and Interpro annotation, tables were downloaded from the Uniprot and Interpro websites directly. Enrichment values and statistics were obtained using the DAVID web server [81] either using the unique Uniprot accession identifiers or the converted symbols, when needed.

2.8.20 Data availability

MS raw data are available at the Chorus database (https://chorusproject.org) under project number 1128. Peptide lists and RBR-ID score plots are available at http://rbrid.bonasiolab.org.

2.9 Supplementary data

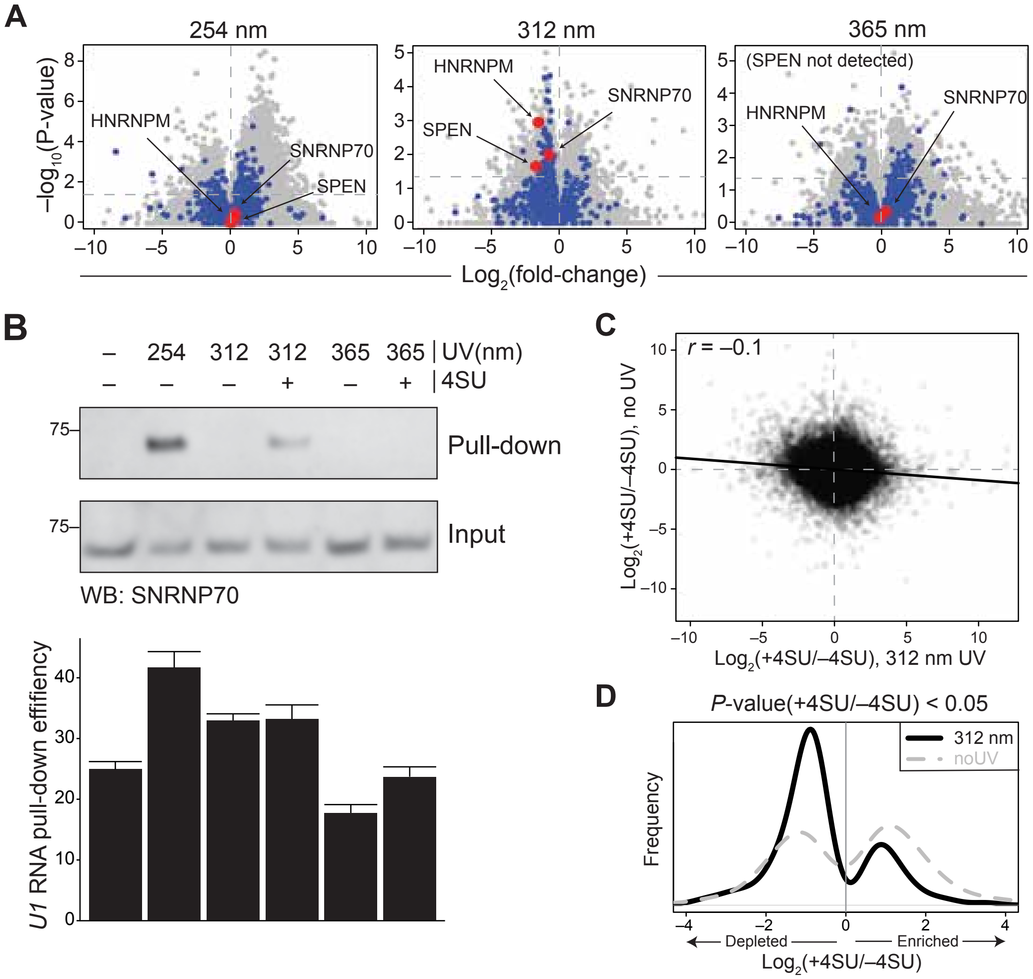


FIGURE 2.7 Comparison of Crosslinking Conditions (related to FIGURE 2.1)
(A) Volcano plots as in Figure 2.1 D, highlighting the position of RRM-spanning peptides from three additional known RBPs. (B) Mouse ESCs were treated or not with 4SU and irradiated with UV at the indicated wavelengths. After lysis U1 RNA-containing complexes were purified with specific U1 probes in denaturing conditions to only recover crosslinked protein, following the ChIRP protocol (Chu et al., 2015). Co-purifying SNRNP70 was detected by western blot (top panels) and comparable efficiencies of RNA pull-downs are shown by RT-qPCR, as a control (bottom panel). Bars represent the mean of three technical replicates + s.e.m. (C) Scatter plot comparing 4SU-dependent depletion of peptides after 312 nm UV crosslinking (x axis) or no UV irradiation (y axis). The regression line is shown in black and the Pearson correlation score is indicated. Data is from six replicates. (D) Density plot showing the distribution of log-fold changes in spectral intensity for peptides with consistent ($P < 0.05$) differences between +4SU and –4SU in the 312 nm irradiation conditions (black line) or no UV conditions (dashed gray line). Experiments and Data generated by C.S.H. and S.S.

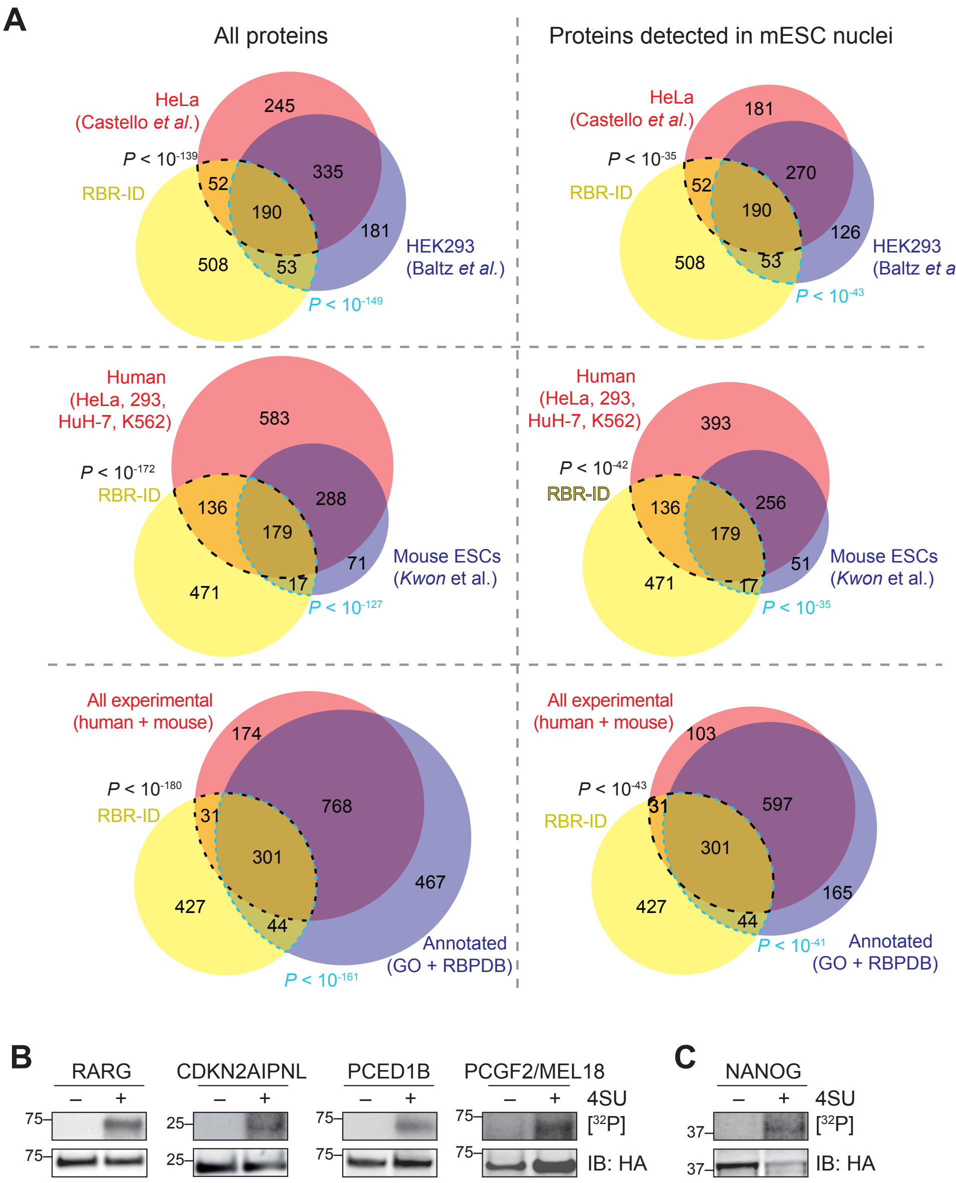


FIGURE 2.8 Analysis and Validation of RBP Candidates Predicted by RBR-ID (related to FIGURE 2.2)

(A) Venn diagrams for the overlaps are shown. On the left ("All proteins"), all protein identifiers were considered; on the right only those proteins that had at least one spectrum in all RBR-ID runs (including different UV wavelengths shown in Figure 2.1) were considered, with the rationale that no overlap could be calculated for proteins that were not expressed/nuclear/detectable in ESCs. Significance of RBR-ID overlaps (shown via dotted lines) against each dataset was determined using one-sided fisher exact test. (B) Validation by PAR-CLIP of four unknown RBPs . Epitope-tagged candidates were expressed in HEK 293 cells and crosslinked to RNA using 365 nm UV. After pull-down, RNA was end-labeled, complexes resolved on SDS-PAGE and revealed by autoradiography (top) or western blot for loading control (bot-tom). (C) Validation as in (B) but using 312 nm UV in mouse ESCs. Experiments and Data generated by C.S.H. and R.W-T.

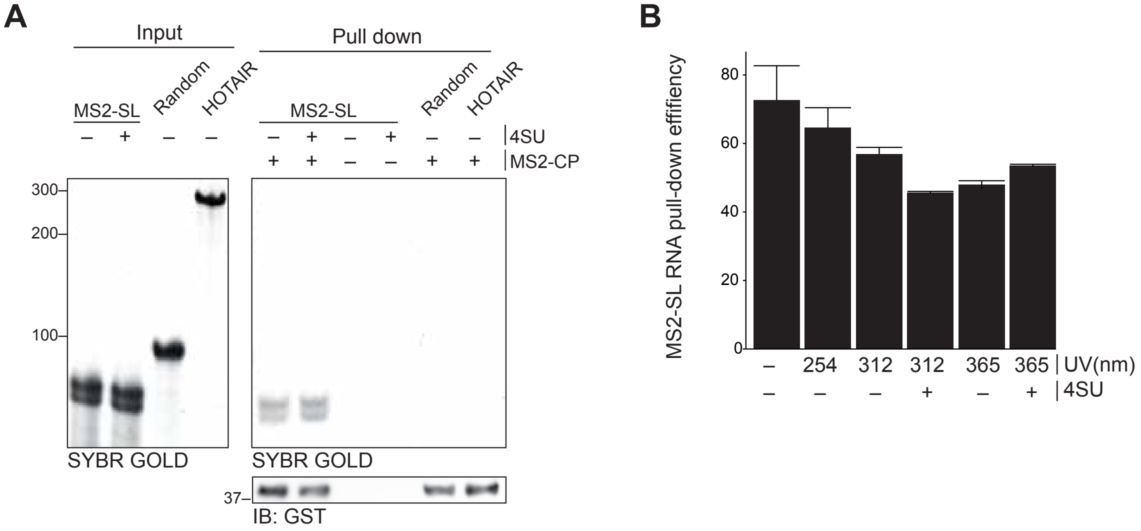


Figure 2.9 Controls for In Vitro MS2 Pull-Downs (related to Figure 2.3)

(A) MS2-CP protein pull-down with glutathione beads in non-denaturing conditions, showing its specific interaction with MS2-SL, regardless of the incorporation of 4SU. A random 100 nts RNA as well as a 5'-terminal fragment of the lncRNA HOTAIR were used as specificity controls. Inputs are shown to the left for RNA and at the bottom for the GST-MS2-CP fusion protein. (B) RT-qPCR for in vitro-transcribed MS2-SL RNA after crosslinking to MS2-CP in the indicated condition and hybridized with biotinylated MS2-SL DNA probe followed by streptavidin pull-down from the experiment shown in Figure 2.3. Data is shown as an RNA pull-down efficiency control. Bars represent the mean of three technical replicates + s.e.m. Experiments and Data generated by C.S.H.

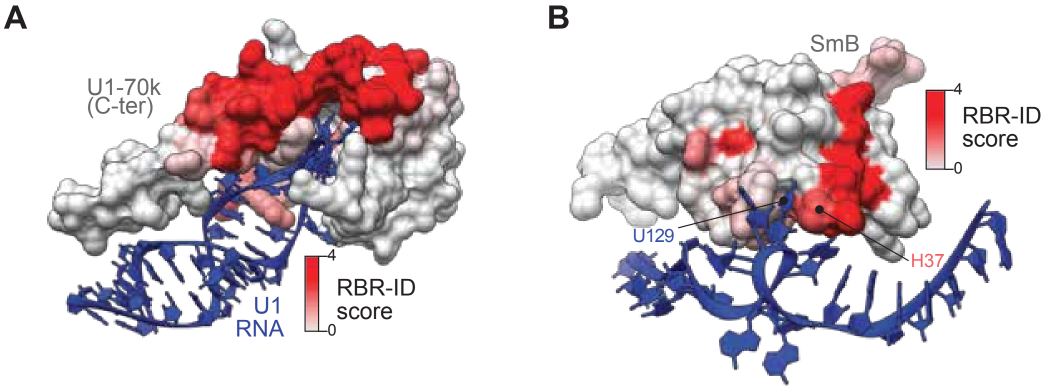


FIGURE 2.10 Additional Examples of RBR Mapping to Known Crystal Structures (related to FIGURE 2.4)

(A) Zoomed-in regions of the crystal structure of the C-terminal RRM of U1-70K bound to stem loop II of U1 snRNA (PDB ID: 4PKD (Kondo et al., 2015)) showing protein surfaces color-coded according to their single-residue RBR-ID score. (B) Same as (A) but for the SmB protein (PDB ID: 4PJO). Experiments and Data generated by C.S.H.

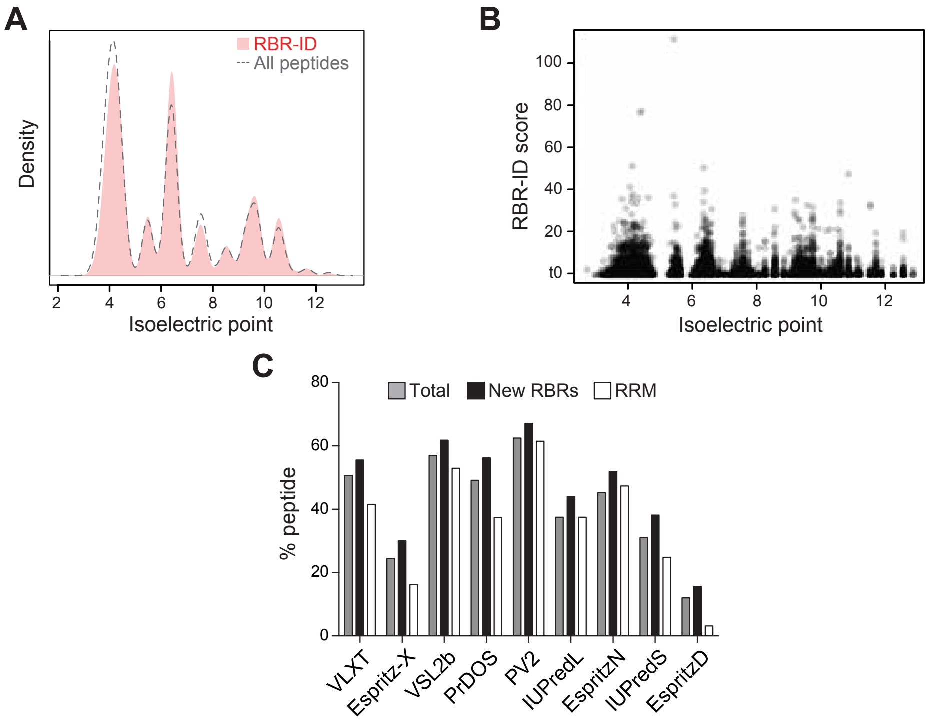


Figure 2.11 Isolectric Point and Disorder Predictions for RNA-Binding Peptides (related to Figure 2.5)

(A) Density plot of the isoelectric point distributions for all peptides (dashed) or newly identified peptides within RBRs (red fill). Values are the same as for Figure 2.5 F. (B) Correlation plot for the isoelectric point and the RBR-ID score for each detected peptide. (C) Percentage of peptides overlapping with disordered regions from various predictors, compiled by Oates et al. in the D2P2 database (Oates et al., 2013). Values are shown for all detected peptides (tot), all top-tier RBR-ID peptides not mapping to a known RNA-binding domain (uRBRs), and all peptides overlapping RRM domains. Experiments and Data generated by C.S.H. and S.S.

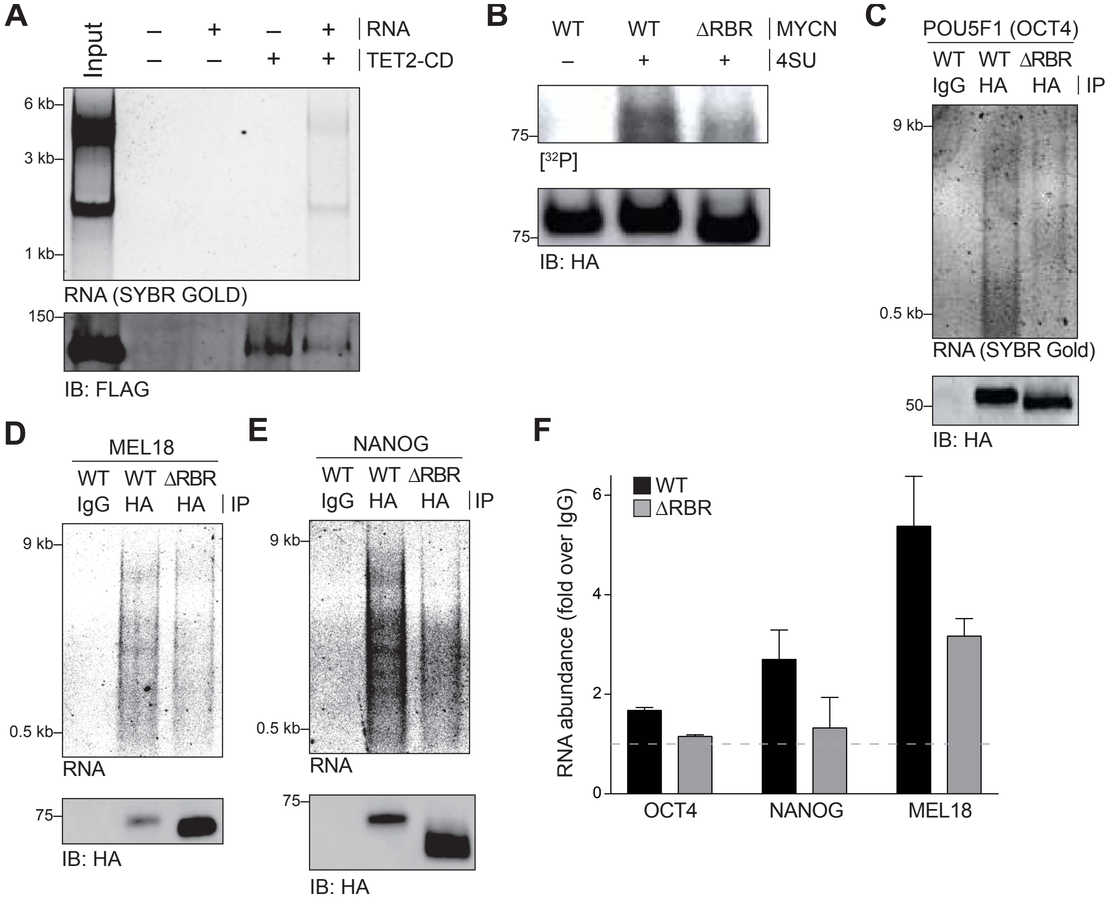


FIGURE 2.12 Additional In Vitro and In Vivo Validations for RBR-ID Candidates (related to FIGURE 2.6)

(A) A recombinant FLAG-tagged fragment encompassing the TET2 catalytic domain and RBR (TET2-CD) was incubated with total RNA from mouse ESCs and immunoprecipitated with anti-FLAG-conjugated agarose beads. The RNA fraction was purified and detected on a denaturing gel with SYBR gold (top). The protein pull-down control is shown at the bottom by western blot. (B) PAR-CLIP for MYCN in presence or absence of 4SU for the WT sequence or a mutant lacking the predicted RBR (ΔRBR). Autoradiography of crosslinked RNA (top) and western blot for loading control (bottom) are shown. (C–E) The indicated proteins either WT or lacking the RBR predicted by RBR-ID were fused to HA and overexpressed in mouse ESCs. After RNA-IP in native conditions, the RNA (top panel) and protein fractions (bottom panels) were separated and resolved on gels. (F) Densitometry on RIP signal from gels shown in (C–E) and additional replicates. The intensity of the SYBR Gold signal in the WT and ΔRBR was normalized to the respective background intensity (IgG lane), shown as a dashed line. Bars show average of two biological replicates + s.e.m. Experiments and Data generated by C.S.H. and R.W-T.

2.10 References

1. Bonasio, R. & Shiekhattar, R. Regulation of transcription by long noncoding RNAs. *Annu Rev Genet* **48,** 433–55 (2014).

2. Holoch, D. & Moazed, D. RNA-mediated epigenetic regulation of gene expression. *Nat Rev Genet* **16,** 71–84 (2015).

3. Goff, L. A. & Rinn, J. L. Linking RNA biology to lncRNAs. *Genome Res* **25,** 1456–65 (2015).

4. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17,** 47–62 (2016).

5. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81,** 145–66 (2012).

6. Wilusz, J. E., Sunwoo, H. & Spector, D. L. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* **23,** 1494–504 (2009).

7. Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161,** 404–16 (2015).

8. McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521,** 232–236 (2015).

9. Minajigi, A. *et al.* Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **349,** 1DUIMMY (2015).

10. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* **46,** 674–90 (2012).

11. Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat Commun* **6,** 10127 (2015).

12. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149,** 1393–1406 (2012).

13. Conrad, T. *et al.* Serial interactome capture of the human cell nucleus. *Nat Commun* **7,** 11212 (2016).

14. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat Struct Mol Biol* **20,** 1122–30 (2013).

15. Brown, J. A., Valenstein, M. L., Yario, T. A., Tycowski, K. T. & Steitz, J. A. Formation of triple-helical structures by the 3'-end sequences of MALAT1 and MENbeta noncoding RNAs. *Proc Natl Acad Sci U S A* **109,** 19202–7 (2012).

16. Wilusz, J. E. *et al.* A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails. *Genes and Development* **26,** 2392–2407 (2012).

17. Lam, M. T., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci* **39,** 170–82 (2014).

18. Wilusz, J. E. Circular RNAs: Unexpected outputs of many protein-coding genes. *RNA Biol* **14,** 1007–1017 (2017).

19. Bonasio, R. *et al.* Interactions with RNA direct the Polycomb group protein SCML2 to chromatin where it represses target genes. *Elife* **3,** e02637 (2014).

20. Kaneko, S. *et al.* Interactions between JARID2 and Noncoding RNAs Regulate PRC2 Recruitment to Chromatin. *Molecular Cell* **53,** 290–300 (2014).

21. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* **8,** 479–90 (2007).

22. Maris, C., Dominguez, C. & Allain, F. H. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* **272,** 2118–31 (2005).

23. Grishin, N. V. KH domain: one motif, two folds. *Nucleic Acids Res* **29,** 638–43 (2001).

24. Chang, K. Y. & Ramos, A. The double-stranded RNA-binding motif, a versatile macromolecular docking platform. *FEBS J* **272,** 2109–17 (2005).

25. Hendrickson, D., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biology* **17** (2016).

26. Kaneko, S. *et al.* Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. *Genes and Development* **24,** 2615–2620 (2010).

27. Saldaña-Meyer, R. *et al.* CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. *Genes and Development* **28,** 723–734 (2014).

28. Greenberg, J. R. Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Res* **6,** 715–32 (1979).

29. Favre, A. *et al.* 4-Thiouridine photosensitized RNA-protein crosslinking in mammalian cells. *Biochem Biophys Res Commun* **141,** 847–54 (1986).

30. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141,** 129–41 (2010).

31. Gorlach, M., Wittekind, M., Beckman, R. A., Mueller, L. & Dreyfuss, G. Interaction of the RNA-binding domain of the hnRNP C proteins with RNA. *EMBO J* **11,** 3289–95 (1992).

32. Hockensmith, J. W., Kubasek, W. L., Vorachek, W. R. & von Hippel, P. H. Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. *J Biol Chem* **261,** 3512–8 (1986).

33. Stiege, W., Kosack, M., Stade, K. & Brimacombe, R. Intra-RNA cross-linking in Escherichia coli 30S ribosomal subunits: selective isolation of cross-linked products by hybridization to specific cDNA fragments. *Nucleic Acids Res* **16,** 4315–29 (1988).

34. Kaneko, S., Son, J., Shen, S. S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20,** 1258–64 (2013).

35. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39,** D301–8 (2011).

36. Bateman, A. *et al.* UniProt: A hub for protein information. *Nucleic Acids Research* **43,** D204–D212 (2015).

37. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106,** 11667–72 (2009).

38. Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D. & Patel, D. J. How chromatin-binding modules interpret histone modifications: Lessons from professional pocket pickers. *Nature Structural and Molecular Biology* **14,** 1025–1040 (2007).

39. Eisen, J. A., Sweder, K. S. & Hanawalt, P. C. Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res* **23,** 2715–23 (1995).

40. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–23 (2007).

41. Tsai, M.-C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329,** 689–693 (2010).

42. Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40,** 939–53 (2010).

43. Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res* **26,** 896–907 (2016).

44. Kanhere, A. *et al.* Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell* **38,** 675–88 (2010).

45. Maison, C. *et al.* SUMOylation promotes de novo targeting of HP1 ± to pericentric heterochromatin. *Nature Genetics* **43,** 220–227 (2011).

46. Castellanos-Rubio, A. *et al.* A long noncoding RNA associated with susceptibility to celiac disease. *Science* **352,** 91–5 (2016).

47. Holz-Schietinger, C. & Reich, N. O. RNA modulation of the human DNA methyltransferase 3A. *Nucleic Acids Res* **40,** 8550–7 (2012).

48. Kung, J. T. *et al.* Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. *Mol Cell* **57,** 361–75 (2015).

49. Sarma, K. *et al.* ATRX directs binding of PRC2 to Xist RNA and Polycomb targets. *Cell* **159,** 869–883 (2014).

50. Sun, S. *et al.* XJpx RNA activates xist by evicting CTCF. *Cell* **153,** 1537 (2013).

51. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat Rev Mol Cell Biol* **14,** 341–56 (2013).

52. Muchardt, C. *et al.* Coordinated methyl and RNA binding is required for heterochromatin localization of mammalian HP1α. *EMBO Reports* **3,** 975–981 (2002).

53. Grahn, E. *et al.* Structural basis of pyrimidine specificity in the MS2 RNA hairpin-coat-protein complex. *RNA* **7,** 1616–27 (2001).

54. Valegard, K., Murray, J. B., Stockley, P. G., Stonehouse, N. J. & Liljas, L. Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature* **371,** 623–6 (1994).

55. Valegård, K. *et al.* The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. *Journal of Molecular Biology* **270,** 724–738 (1997).

56. Kondo, Y., Oubridge, C., van Roon, A. M. & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife* **4** (2015).

57. Wild, T. & Cramer, P. Biogenesis of multisubunit RNA polymerases. *Trends Biochem Sci* **37,** 99–105 (2012).

58. Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* **11,** 394–403 (2004).

59. Ujvari, A. & Luse, D. S. RNA emerging from the active site of RNA polymerase II interacts with the Rpb7 subunit. *Nat Struct Mol Biol* **13,** 49–54 (2006).

60. Aravind, L. & Koonin, E. V. SAP - a putative DNA-binding motif involved in chromosomal organization. *Trends Biochem Sci* **25,** 112–4 (2000).

61. Doerks, T., Copley, R. R., Schultz, J., Ponting, C. P. & Bork, P. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res* **12,** 47–56 (2002).

62. Parker, L. M., Fierro-Monti, I. & Mathews, M. B. Nuclear Factor 90 Is a Substrate and Regulator of the Eukaryotic Initiation Factor 2 Kinase Double-stranded RNA-activated Protein Kinase. *Journal of Biological Chemistry* **276,** 32522–32530 (2001).

63. Wolkowicz, U. M. & Cook, A. G. NF45 dimerizes with NF90, Zfr and SPNR via a conserved domain that has a nucleotidyltransferase fold. *Nucleic Acids Research* **40,** 9356–9368 (2012).

64. Kim, T. *et al.* Aspartate-glutamate-alanine-histidine box motif (DEAH)/RNA helicase a helicases sense microbial DNA in human plasmacytoid dendritic cells. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 15181–15186 (2010).

65. Walbott, H. *et al.* Prp43p contains a processive helicase structural architecture with a specific regulatory domain. *EMBO J* **29,** 2194–204 (2010).

66. Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res* **29,** 943–54 (2001).

67. Närvä, E. *et al.* RNA-binding protein L1TD1 interacts with LIN28 via RNA and is required for human embryonic stem cell self-renewal and cancer cell proliferation. *Stem Cells* **30,** 452–460 (2012).

68. Davidovich, C., Zheng, L., Goodrich, K. J. & Cech, T. R. Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* **20,** 1250–7 (2013).

69. Davidovich, C. *et al.* Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. *Mol Cell* **57,** 552–8 (2015).

70. Kaneko, S., Son, J., Bonasio, R., Shen, S. S. & Reinberg, D. Nascent RNA interaction keeps PRC2 activity poised and in check. *Genes Dev* **28,** 1983–8 (2014).

71. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods* **11,** 1064–70 (2014).

72. Castello, A. *et al.* Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell* **63,** 696–710 (2016).

73. Akhtar, A., Zink, D. & Becker, P. B. Chromodomains are protein-RNA interaction modules. *Nature* **407,** 405–9 (2000).

74. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410,** 116–20 (2001).

75. Elbaum-Garfinkle, S. *et al.* The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 7189–7194 (2015).

76. Phan, A. T. *et al.* Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. *Nat Struct Mol Biol* **18,** 796–804 (2011).

77. Castello, A., Fischer, B., Hentze, M. W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends Genet* **29,** 318–27 (2013).

78. Delatte, B. *et al.* RNA biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* **351,** 282–5 (2016).

79. Fu, L. *et al.* Tet-mediated formation of 5-hydroxymethylcytosine in RNA. *J Am Chem Soc* **136,** 11582–5 (2014).

80. Cheng, L., Pisitkun, T., Knepper, M. A. & Hoffert, J. D. Peptide labeling using isobaric tagging reagents for quantitative phosphoproteomics. *Methods in Molecular Biology* **1355,** 53–70 (2016).

81. Huang da, 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).

82. Hochedlinger, K., Yamada, Y., Beard, C. & Jaenisch, R. Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. *Cell* **121,** 465–77 (2005).

83. Kaelin W. G., J. *et al.* Expression cloning of a cDNA encoding a retinoblastoma-binding protein with E2F-like properties. *Cell* **70,** 351–64 (1992).

84. Gao, Z. *et al.* PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Mol Cell* **45,** 344–56 (2012).

85. Crawford, D. J. *et al.* Tet2 Catalyzes Stepwise 5-Methylcytosine Oxidation by an Iterative and de novo Mechanism. *J Am Chem Soc* **138,** 730–3 (2016).

86. Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res* **11,** 1475–89 (1983).

87. Friedman, J. H. A variable span smoother. *Journal of the American Statistical Association* (1984).

3. RNA EXPLOITS AN EXPOSED REGULATORY SITE TO INHIBIT THE ENZYMATIC ACTIVITY OF PRC2

This chapter is adapted from a published manuscript:

> *Qi Zhang*$^{+}$*, Nicholas J. McKenzie*$^{+}$*, Robert Warneford-Thomson*$^{+}$*, Emma H. Gail, Sarena F. Flanigan, Brady M. Owen, Richard Lauman, Vitalina Levina, Benjamin A. Garcia, Ralf B. Schittenhelm, Roberto Bonasio***, Chen Davidovich**. RNA exploits an exposed regulatory site to inhibit the enzymatic activity of PRC2. *Nature Structural and Molecular Biology*. 2019.

$^{+}$ These authors contributed equally: Qi Zhang, Nicholas McKenzie and Robert Warneford-Thomson.
* co-corresponding authors

3.1 Abstract

Polycomb repressive complex 2 (PRC2) is a histone methyltransferase that maintains cell identity during development in multicellular organisms by marking repressed genes and chromatin domains. In addition to four core subunits, PRC2 comprises multiple accessory subunits that vary in their composition during cellular differentiation and define two major holo-PRC2 complexes: PRC2.1 and PRC2.2. PRC2 binds to RNA, which inhibits its enzymatic activity, but the mechanism of RNA-mediated inhibition of holo-PRC2 is poorly understood. Here we present in vivo and *in vitro* protein–RNA interaction maps and identify an RNA-binding patch within the allosteric regulatory site of human and mouse PRC2, adjacent to the methyltransferase centre. RNA-mediated inhibition of holo-PRC2 is relieved by allosteric activation of PRC2 by H3K27me3 and JARID2-K116me3 peptides. Both holo-PRC2.1 and -PRC2.2 bind RNA, providing a unified model to explain how RNA and allosteric stimuli antagonistically regulate the enzymatic activity of PRC2.

3.2 Introduction

PRC2 is a histone methyltransferase (HMTase) that methylates H3 histones at lysine 27 to form the H3K27me3 mark of facultative heterochromatin (reviewed in [1–4]). The H3K27me3 mark is essential for the epigenetic maintenance of transcriptional repression at developmentally expressed genes. The core PRC2 complex includes a histone methyltransferase subunit—EZH2 or EZH1;

the regulatory subunit EED; one histone-binding subunit—RBBP4 or RBBP7; and SUZ12, which serves as a scaffold [5–7]. The recruitment of PRC2 to chromatin and its HMTase activity are tightly regulated. For instance, after EZH2 introduces the H3K27me3 histone mark, the methylated histone peptide binds to a regulatory site within EED to stimulate the methyltransferase activity of PRC2 [8] through allosteric activation [9–12].

The function of PRC2 is also regulated by its accessory subunits: sub-stoichiometric subunits of PRC2 that are differentially expressed during development [1, 2]. For instance, JARID2 is methylated by PRC2 at lysine 116 (JARID2-K116me3) and then binds to the regulatory center in EED to allosterically activate PRC2 during de novo methylation at target genes [11]. AEBP2 and the three polycomb-like (PCL) proteins—PHF1, PHF19 and MTF2—facilitate DNA binding by PRC2 through direct interactions [13–17]. EPOP (previously termed C17ORF96 or esPRC2p48) is another accessory subunit of PRC2 that facilitates gene repression [18, 19]. Unbiased proteomic studies identified these factors as the most abundant accessory subunits of PRC2 [18–22] and determined that they form two types of holo-PRC2 complexes, including the core subunits and different accessory subunits [21]: PRC2.1 includes one of the PCL accessory subunits (PHF1, PHF19 or MTF2) and EPOP or PALI 1/2, while PRC2.2 includes AEBP2 and JARID2 [21].

Direct interactions with RNA have been proposed to recruit PRC2 to target genes for epigenetic repression, to evict it from active genes and to retain it in a poised state at lowly expressed genes (reviewed in [23–27]). More recently, RNA was shown to inhibit the HMTase activity of PRC2 [28–30]. Experiments using isolated subunits, partial complexes and the core PRC2 complex attributed RNA binding to the core subunits EZH2, EED and SUZ12 [30–35]. Yet, the question of how RNA inhibits different types of holo-PRC2 complexes remains unanswered. RNA competes for nucleosome [36] and DNA binding by PRC2 and the automethylation activity of EZH2 is not affected by RNA [15], suggesting that the competition with DNA might explain the inhibitory effect of RNA on PRC2 [15]. This, however, leaves an unresolved conundrum: how does PRC2 overcome RNA inhibition at target genes while within the RNA-rich environment of the nucleus? Moreover, an earlier study demonstrated that RNA inhibits the HMTase activity of PRC2 also toward biotinylated histone tail peptides [29], which is inconsistent with a model whereby RNA inhibits PRC2 activity exclusively by competing with DNA binding. This suggests the possibility that RNA may inhibit the methyltransferase activity of PRC2 via multiple mechanisms. Testing this possibility is important not only for the understanding of how RNA regulates the HMTase activity of PRC2 at the molecular level, but also because PRC2 methylates non-histone substrates, including transcriptional regulators ([37] and references therein).

Here, we show that RNA binds and inhibits both types of holo-PRC2 complexes—PRC2.1 and PRC2.2. Using in vivo UV crosslinking and mass spectrometry in mouse embryonic stem cells (mESCs), we have mapped RNA-binding regions on the core complex to both regulatory and catalytic centers. In vitro studies on active holo-PRC2 complexes have confirmed binding of RNA to the regulatory site of PRC2, at the interface between EZH2 and EED, near the catalytic center. In agreement with this observation, RNA-mediated inhibition of PRC2 is relieved by peptides that bind to the allosteric regulatory center and RNA inhibited the methyltransferase activity of PRC2 also toward DNA-free substrates. Based on these findings, we provide a mechanistic framework to explain how RNA-mediated inhibition of the two major types of holo-PRC2 complexes takes place and further generalize it for RNA-mediated inhibition of methyltransferase activity toward non-histone substrates.

3.3 Results

3.3.1 RNA binds both PRC2.1 and PRC2.2 in vivo

We previously used protein–RNA crosslinking and mass spectrometry for RNA-binding region identification (RBR-ID) in the nuclear proteome of mESCs [38]. Through reanalysing these data, we detected significant peptide hits within core PRC2 subunits EZH2 and SUZ12, consistent with previous observations [30–35]. We also identified the accessory PRC2.2 subunit AEBP2 (Figure 3.8 A), consistent with our own previous work [34, 39]. However, we did not detect peptides within JARID2, which is known to crosslink to RNA in vivo [28]. We reasoned that the low coverage of this and other PRC2 subunits in the whole nuclear proteome (Figure 3.8 A) might have limited our ability to detect protein-RNA interaction sites in this complex.

To overcome this obstacle, we developed a 'targeted' variant of RBR-ID and utilized an immunoprecipitation step to focus the mass spectrometry analysis on PRC2 (Figure 3.1 A, Figure

3.8 B and Digital Supplemental Tables S9 and S10). The portion of mass spectrometry signal that could be attributed to PRC2 increased 100-fold, from 0.4% in the proteome-wide data to 40% in the targeted RBR-ID experiments (Figure 3.8 C), which allowed us to identify several additional PRC2 peptides that crosslinked to RNA (Figure 3.1 B and Figure 3.8 D). We recovered multiple significant hits in all core subunits, in the PRC2.1 accessory subunits MTF2 and PALI, and in both accessory subunits of PRC2.2—AEBP2 and JARID2 (Figure 3.1 B and Figure 3.8 A). These hits accumulated on the catalytic lobe of PRC2 (Figure 3.1 C, D and Figure 3.8 E) and several mapped to domains previously proposed to bind to RNA, such as the EZH2 RBR (residues 342–368, [32]), the JARID2 RBR (residues 332–358, [28]), the EZH2 CXC and SET domains35, and the RRM-like beta-sheet domain of SUZ12 [5]. In addition to these, we noticed a highly significant crosslinked peptide on EED, very close to the regulatory center, near the stimulatory recognition motif9 (SRM) of EZH2 (Figure 3.1 B, D; EED 336–355 is highlighted).

Thus, our targeted RBR-ID approach not only identified interactions between RNA and the PRC2.2 subunits AEBP2 and JARID2 [15, 28, 30, 34, 38, 39], but also MTF2 (also known as PCL2) and PALI—accessory subunits of the PRC2.1 complex that were previously not known to bind RNA. This strongly suggests that PRC2.1 also binds RNA in vivo.

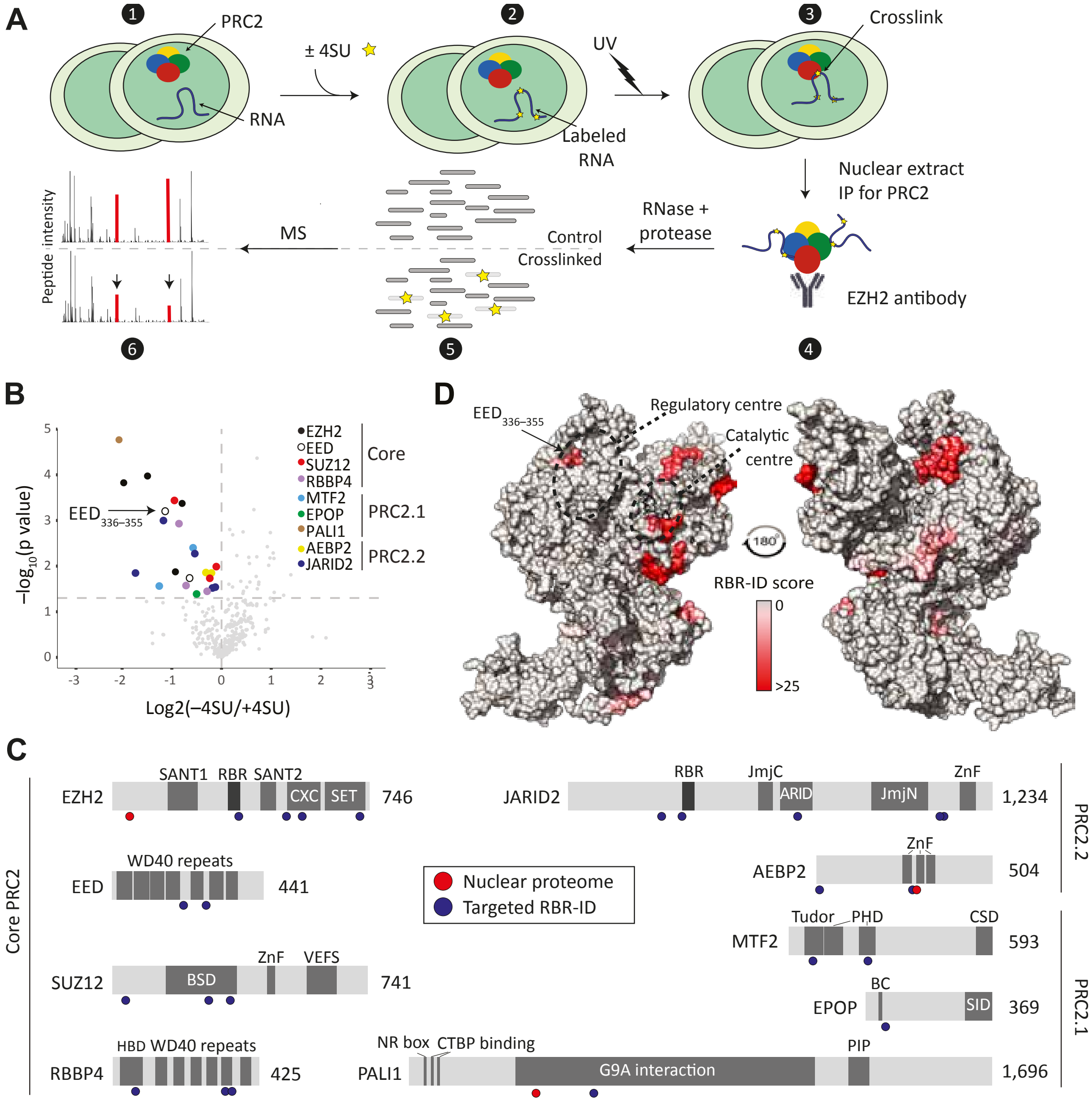


FIGURE 3.1 Targeted RBR-ID of PRC2

(A) Targeted RBR-ID experimental design. Mouse ESCs treated with or without 4SU (1 and 2) were irradiated with UV to generate RNA–protein crosslinks (3). After preparing nuclear extracts, we 'targeted' the RBR-ID technique by performing immunoprecipitations for endogenous PRC2 using an EZH2 antibody (4). Following immunoprecipitation, we treated eluted proteins with RNase and protease to remove crosslinked RNA and generate peptides (5), which were analyzed via high-resolution LC–MS/MS to identify decreases in apparent peptide abundance caused by the crosslink with RNA (6). (B) Volcano plot of peptide intensities comparing material from 4SU-pulsed and control (-4SU) cells. The dashed horizontal line indicates the P value of 0.05. Peptides on core and accessory PRC2 subunits are highlighted. P values were calculated using paired or unpaired two-sided Student's t-tests (see Methods) from three independent experiments and ten total replicates. (C) Mapping to PRC2 subunits of RNA-interacting peptides detected by targeted RBR-ID (blue circles, this study) or proteome-wide RBR-ID [38] (red circles). Known protein domains, including previously identified RNA-binding regions (RBRs) on EZH2 [32] and JARID2 [28] are shown. (D) RBR-ID structural mapping. Residue-level RBR-ID scores were calculated according to the level of 4SU depletion and statistical significance and the resulting heat-map was used to color the surface of a composite PRC2 model using two published PRC2 structures (PDB accession: 5WAI and 6C23, see Methods). The substrate peptide in the catalytic center is shown in black. Experiments and Data generated by R.W-T. and R.L.

3.3.2 RNA binds and inhibits both PRC2.1 and PRC2.2 in vitro.

To further investigate the molecular nature and biochemical function of PRC2-RNA interactions in the context of holo-PRC2 complexes, we continued our studies in vitro, by reconstituting the human core PRC2 complex and a PRC2 complex with either human AEBP2 (PRC2–AEBP2) or the PCL protein PHF19 (PRC2–PHF19) (Figure 3.2 A). Given that PRC2 preferably binds to RNA that contains short repeats of consecutive guanines [34], and since fluorescence anisotropy requires a small labeled ligand, we quantified affinity of PRC2 for an RNA composed of four UUAGGG repeats (G4 24 RNA, Figure 3.2 B and Table 3.1). These four repeats originate from TERRA RNA, fold into a G-quadruplex RNA structure and bind PRC2 in cells [34, 40]. As a negative control, we used a size-matched mutant RNA without G-tracts, composed of four UGAGUG repeats (G4 mt 24 RNA, Figure 3.2 B and Table 3.1). In good agreement with earlier observations [39], the affinity of the core PRC2 complex to RNA (Kd=129±6nM) increased by approximately 2-fold when AEBP2 is in the complex (Kd=54.6±3.8nM). Strikingly, the addition of PHF19 increased the affinity of PRC2 to RNA by nearly 4-fold compared to that of the core PRC2 complex alone. The affinity of PRC2-PHF19 (Kd=34.0±1.8nM) to RNA was even greater than that of the PRC2–AEBP2 complex. The PRC2.1 accessory subunits MTF2 and EPOP also increased the affinity of PRC2 for RNA, with the PRC2– MTF2–EPOP complex having an approximately 2-fold higher affinity for

TABLE 3.1 Affinities of PRC2 complexes to G4 24 and G4 mt 24 RNA

Protein	RNA	K_d (nM)	Hill
PRC2	G4 24	129 ± 6	0.92 ± 0.03
PRC2–AEBP2	G4 24	54.6 ± 3.8	0.88 ± 0.04
PRC2–PHF19	G4 24	34.0 ± 1.8	1.01 ± 0.04
PRC2	G4 mt 24	n.d.	n.d.
PRC2–AEBP2	G4 mt 24	n.d.	n.d.
PRC2–PHF19	G4 mt 24	n.d.	n.d.

RNA (Kd=40.9±3.9nM) compared to PRC2–MTF2 (Kd = 82.4 ± 9.8 nM) and approximately 3-fold higher compared to core PRC2 (Kd=129±6nM; Figure 3.2 B, Table 3.1 and Figure 3.9 G–i). These results indicate that both PRC2.1 and PRC2.2 holo-complexes bind to RNA. None of the PRC2 complexes bound to the mutant RNA (triangles in Figure 3.2 B and Table 3.1), indicating that the RNA-binding specificity of PRC2 toward this G-tract motif34 is preserved in the presence of a PCL subunit.

Since the PCL protein MTF2 interacts with RNA in vivo and *in vitro* and the PRC2–PHF19 complex interacts with RNA in vitro, we wished to determine if PCL proteins allow for RNA-mediated inhibition of PRC2. We performed *in vitro* histone methyltransferase assays using recombinant nucleosome substrates in the presence or absence of a 256-base-long RNA that includes ten UUAGGG repeats flanked by sequences devoid of G-tracts (G4 256 RNA, see Methods), which also bound to PRC2 with nanomolar affinity (Figure 3.10 K, L). In agreement with previous studies, the RNA inhibited the HMTase activity of the core PRC2 complex [29, 30, 41], the PRC2–AEBP2 complex [15] and the PRC2–AEBP2–JARID2 complex [15, 28] toward nucleosome substrates (Figure 3.2 C and Figure 3.9 A). We then performed the same experiments with three reconstituted human PRC2.1 complexes: PRC2–PHF1, PRC2– PHF19 and PRC2–MTF2–EPOP (Figure 3.2 A and Figure 3.9 C, D). Despite variations in the baseline activity, RNA exerted an inhibitory effect regardless of which PCL protein (PHF1, PHF19 or MTF2) was present or whether EPOP was included (Figure 3.2 D and Figure 3.9 B, J, K). This indicates that RNA inhibits the HMTase activity of PRC2 even when assembled with its accessory subunits to form either the PRC2.1 or PRC2.2 complexes.

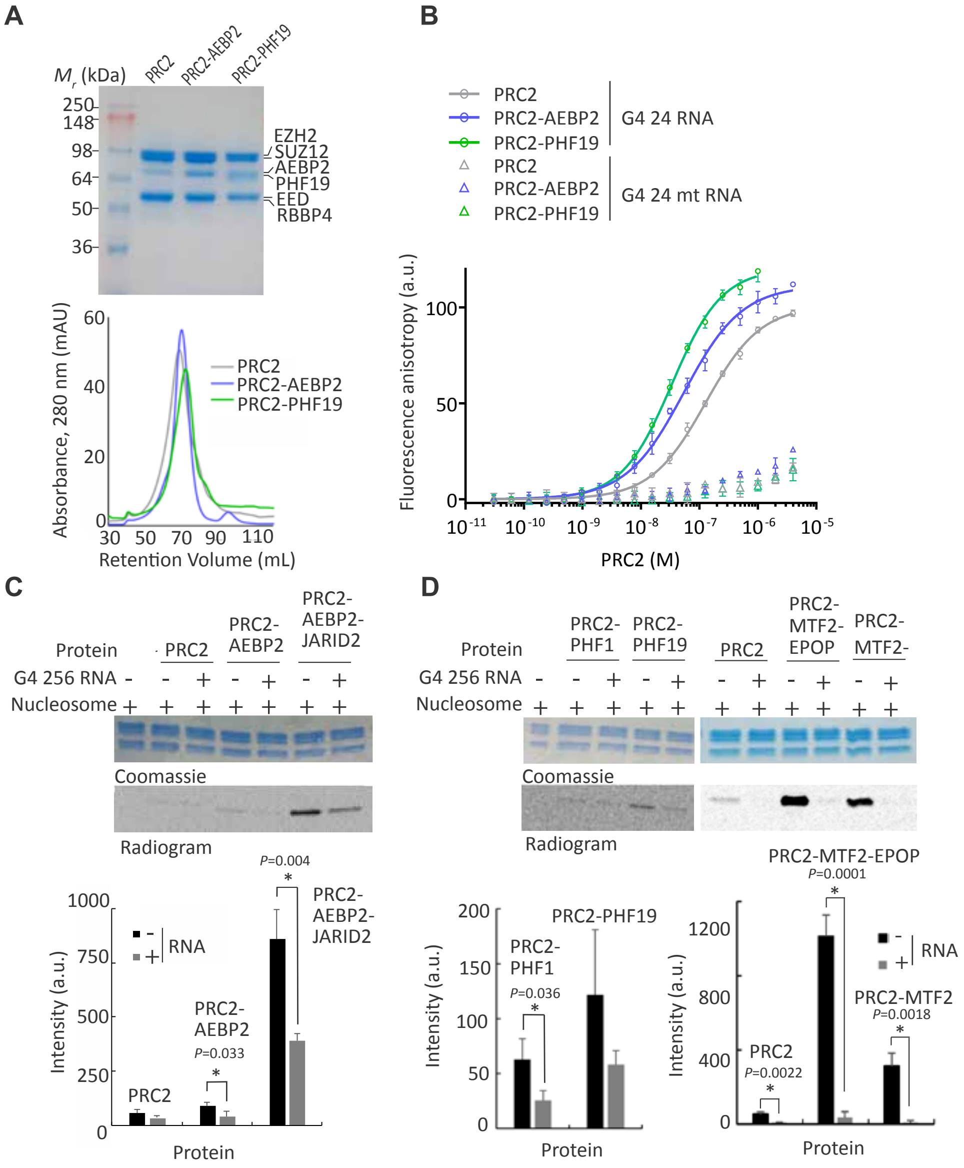


FIGURE 3.2 RNA binds to and inhibits both PRC2.1 and PRC2.2

(A) Coomassie blue-stained SDS–PAGE (top) and gel filtration chromatography (bottom, HiPrep 16/600 Sephacryl S-400 HR) of the PRC2 complexes that were used for binding assays. (B) Fluorescence anisotropy used to quantify the affinity of PRC2 complexes to G4 24 and G4 24 mutant (mt) RNAs. Data represent the mean of three independent experiments that were carried out on different days, error bars represent standard deviation. See Table 1 for dissociation constants (Kd) and Hill coefficients. (C) and (D), HMTase assays of PRC2.2 (C) and PRC2.1 (D) complexes toward nucleosome substrates were carried out in the presence or absence of 8 μM G4 256 RNA. Histone proteins were visualized using Coomassie (upper gel) and methylation levels of H3 were determined by 14C-autoradiography (bottom). Bar plots represent the mean of quantification using densitometry and error bars represent standard deviation based on three independent experiments. P values were determined using unpaired two- tailed Student's t-test; *P $<$ 0.05. See Figure 3.9 for complete gel scans, SDS–PAGE and gel filtration chromatography of holo-PRC2 complexes and evidence for nucleosome reconstitution. mAU, milli-absorbance units; a.u., arbitrary units. Source data are available in Digital Supplemental Table S13. Data and analysis generated by collaborators.

3.3.3 RNA binds to the allosteric regulatory center of PRC2

Under the notion that RNA binds to the PRC2 core complex and to the two types of holo-PRC2 complexes with similar affinity (up to 4-fold Kd, Figure 3.2 B and Table 3.1), specificity and inhibitory activity, we reasoned that RNA inhibits PRC2 through interactions with the core PRC2 subunits. To determine the site of PRC2 that binds to the inhibitory G-tract-containing RNA, we combined an *in vitro* UV crosslinking method [34] with the RBDmap approach [42] for mapping protein-RNA interactions using mass spectrometry (Figure 3.3 A and Methods). We used this approach to detect protein-RNA interactions between the reconstituted PRC2–AEBP2 complex and the G4 256 RNA (Figure 3.3 B and Figure 3.10 A) that, similar to the G4 24 RNA, binds to PRC2 with high affinity (Figure 3.10 K, L). Remarkably, most of the RNA-linked peptides that were identified in independent *in vitro* RBDmap replicates clustered within the same site at the interface between EED and EZH2 (Figure 3.3 B and Digital Supplemental Table S11), overlapping the mouse EED peptide 336–355 identified in vivo by RBR-ID (Figure 3.1 D). These results are also in agreement with earlier UV crosslinking experiments that were carried out with SDS–polyacrylamide gel electrophoresis (SDS–PAGE), without high-resolution mapping, and identified EED and EZH2 as the two subunits that crosslinked to RNA within the context of an assembled PRC2–AEBP2 complex [34]. Intriguingly, this RNA-binding site overlaps with the regulatory site that was previously shown to regulate the HMTase activity of PRC2 through allosteric stimulation [9–12] (Figure 3.3 B).

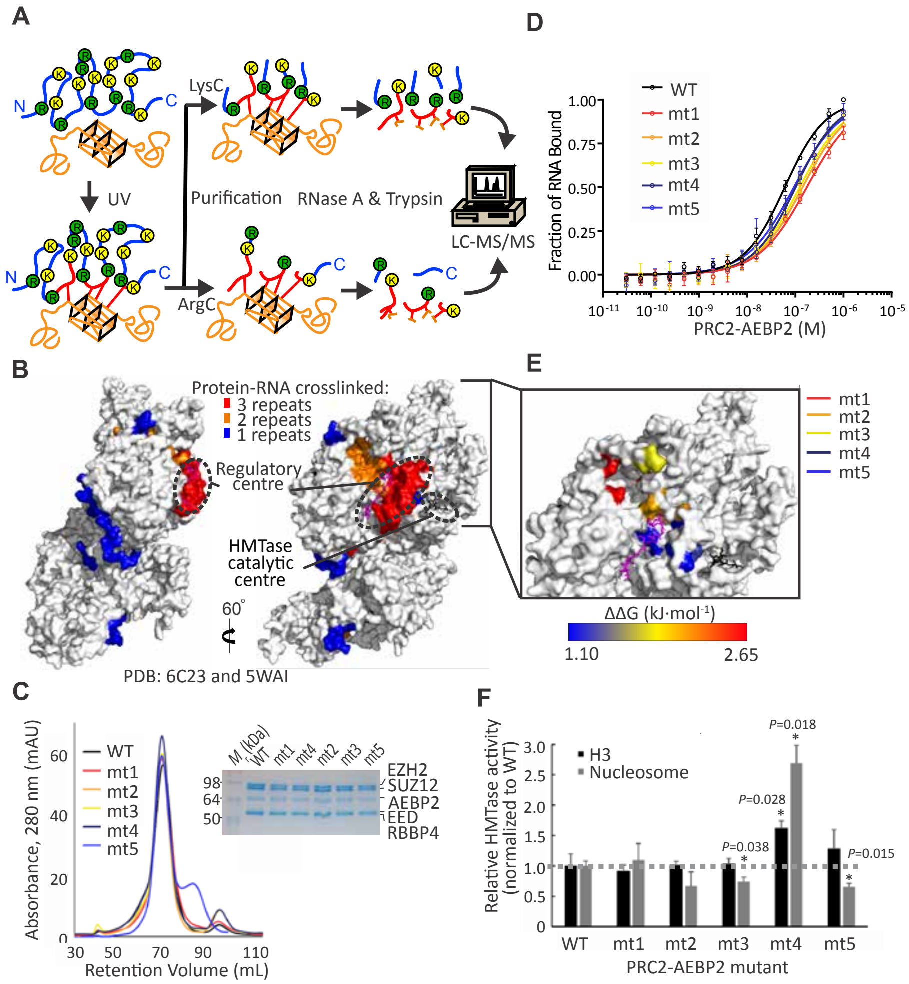


FIGURE 3.3 Mapping of protein-RNA interactions within PRC2–AEBP2 in vitro

(A) Schematic representation of the in vitro RBDmap workflow (see Methods): in vitro reconstituted protein–RNA complexes are crosslinked, followed by tandem proteolytic digestion and LC–MS/MS to reveal peptides adjacent (blue) to the protein–RNA crosslink (red). RNA is shown in orange. (B) RBDmap results: amino acids within the PRC2–AEBP2 structure were colored in blue, orange or red if they resided within peptides that were crosslinked to RNA in 1, 2 or 3 independent RBDmap experiments, respectively. A methylated peptide in the regulatory center is colored magenta and the substrate peptide in the catalytic center is colored black (PDB accession: 6C23 and 5WAI). (C) and (D), Validation using point mutations. The purity and integrity of the mutant complexes were assessed using SDS–PAGE and gel filtration chromatography (HiPrep 16/600 Sephacryl S-400 HR) (C). Fluorescence anisotropy was used to quantify the affinity of the mutants to G4 24 RNA (D). The resulting dissociation constant (Kd), Hill coefficients and the derived $\Delta\Delta G$ are indicated together with details of the mutated amino acids in EZH2 and EED in Table 3.2. Error bars in (D) represent standard deviation based on three independent experiments that were done on different days. (E) The impaired capacity of the mutants to bind RNA is represented in a $\Delta\Delta G$ heat map using the PRC2–AEBP2 structure. Mutated amino acids are mapped to the structure and $\Delta\Delta G$ color code is indicated (bottom). (F) Mean HMTase activity of indicated PRC2 mutants toward H3 histones (black bars) or nucleosomes (gray bars) normalized to the activity of wild-type PRC2–AEBP2 (dashed line). Error bars represent standard deviation based on three independent experiments. P values were determined using paired two-tailed Student's t-test; $^*P < 0.05$. See Figure 3.10 for HMTase radiograms and gel scans, SDS–PAGE analyses and mass spectrometry intensities resulting from the RBDmap process and additional mutants that were assayed. Source data are available in Digital Supplemental Table S14. Data and analysis generated by collaborators.

To validate this finding, we introduced point mutations in EZH2 and EED within the identified RNA-binding site or its vicinity (mt1–8). We reconstituted mutant PRC2–AEBP2 complexes (Figure 3.3 C and Figure 3.10 B) and measured their affinity to G4 24 RNA (Figure 3.3 D, E, Table 3.2 and Figure 3.10 C, D). The mutations reduced the affinity of PRC2 to RNA (Figure 3.3 E, Table 3.2 and Figure 3.10 D-G), with $\Delta\Delta G$ in the range of 1.1–2.7 kJ mol^{-1}.The mutations mt1–3 and mt6–7 caused the largest reduction in affinity ($\Delta\Delta G$ 1.8–2.7 kJ mol^{-1}, Figure 3.3 D, E, Table 3.2 and Figure 3.10 C-G, which supports a direct function of the mutated amino acids in RNA binding. Importantly, these mutations did not lead to adverse effects on complex assembly (Figure 3.3 C and Figure 3.10 B) or on PRC2 activity toward histone substrates, and only mt3, mt5 and mt8 displayed a negative impact on the activity of PRC2 toward nucleosome substrates (Figure 3.3 F and Figure 3.10 H-J). The mutant mt4 displayed a positive effect on HMTase activity, possibly by stabilizing PRC2 in a conformation resembling its allosterically stimulated state, given its location near the regulatory center. The mutants led to only a modest reduction in the affinity of PRC2 to RNA and even in the case of the complex bearing the most effective set of mutations, as determined by its *in vitro* affinity for RNA (mt1, 3-fold reduction of affinity), we could not detect any change in the extent of RNA-mediated inhibition (Figure 3.11 A-D). This is consistent with

TABLE 3.2 **Affinities of PRC2 complexes to G4 24 and G4 mt 24 RNA**

PRC2–AEBP2 mutant	**Mutation sites**	**K$_d$ (nM)**	**Hill**	**$\Delta\Delta$G (kJ mol^{-1})**
WT	n/a	60.5 ± 2.2	1.03 ± 0.04	0
mt1	EZH2 R27A/R31A/F165A EED R355A	173 ± 7	0.87 ± 0.03	2.65
mt2	EZH2 H129A/K156A/H158A EZH2 G159A/R161A EED R306A	141 ± 8	0.91 ± 0.04	2.13
mt3	EZH2 R16A/K17A/R18A/K20A	124 ± 7	0.91 ± 0.04	1.81
mt4	EZH2 K661A	102 ± 4	0.99 ± 0.04	1.32
mt5	EZH2 Y133A/T144A EZH2 F145A/Y153A EED Y308A	93.7 ± 6.3	0.89 ± 0.05	1.10

our photocrosslink mapping (Figs. 1d and 3b), and previous hydrogen deuterium exchange analyses [35], which pointed to multiple relatively large RNA-binding surfaces within PRC2 that might be involved in RNA mediated inhibition of the complex (see below).

In addition to mutations guided by our RBDmap results (Figure 3.3 B), we assayed two RNA-binding-deficient mutants, mt4 (Figure 3.3 C-E) and mt8 (Figure 3.10 C,D,G) that reside externally to the regulatory site and were previously analyzed [35]. In the original study, these mutants were tested in the context of the minimal PRC2 core complex (EZH2, EED and the VEFS domain of SUZ12) and found to reduce the affinity of PRC2 to RNA by up to 12-fold Kd [35]. Within the PRC2–AEBP2 complex, mt4 and mt8 exhibited an affinity change of approximately 2-fold ($\Delta\Delta$G of 1.3 and 1.1 kJ mol^{-1}, respectively). These results are in good agreement with our observations and those of others that AEBP2 and regions of SUZ12—beyond the VEFS domain [5]—interact with RNA (Figure 3.1) and that AEBP2 increases the affinity of PRC2 for RNA (Figure 3.2 B and Table 3.1).

Our results thus indicate that amino acids within the allosteric site of PRC2 and in its immediate vicinity are directly involved in RNA binding. These findings reinforce previous observations of dispersed RNA-binding sites in EZH2 [35], but also show additional significant protein-RNA contacts within the regulatory subunit EED, and point to the regulatory site of PRC2

TABLE 3.3 Affinities of PRC2–AEBP2 complexes to G4 24 RNA in the presence or absence of stimulatory peptides

Competitive peptide	**Sequence**	**K$_d$ (nM)**	**Hill**	**ΔΔG (kJ mol^{-1})**	**n**
No peptide	n/a	19.2 ± 2.7	0.79 ± 0.06	0	11
JARID2-K116me3	KRPRLQAQRK(me3) FAQSQ	46.7 ± 6.9	0.88 ± 0.08	2.24	6
H3K27me3	TKAARK(me3) SAPAT	22.5 ± 3.1	1.05 ± 0.12	0.4	4

as an important determinant for RNA binding.

3.3.4 Stimulatory peptides relieve RNA-mediated inhibition of PRC2

Given the location of a prominent RNA-binding site within the regulatory center of PRC2, we wished to determine the regulatory interplay between peptide ligands that bind to this site and stimulate PRC2 (H3K27me3 and JARID2-K116me3) and inhibitory RNAs (for example, G4 256 RNA and G4 24 RNA). Both stimulatory peptides H3K27me3 and JARID2-K116me3 peptides could significantly overcome RNA-mediated inhibition by G4 24 and G4 256 RNAs (Figure 3.4 A-C, Figure 3.11 A-D). Quantitative binding assays indicated that the JARID2-K116me3 peptide competed with RNA for binding to PRC2, decreasing the ΔG by 2.24 kJ mol–1, but the H3K27me3 peptide did not (Figure 3.4 D and Table 3.3). This is consistent with the observation that the longer JARID2-K116me3 peptide extends into the pocket formed at the EED–EZH2 interface (Figure 3.4 E,F, in purple), which was identified as a primary site of PRC2-RNA interactions by both RBR-ID (Figure 3.1) and RBDmap (Figure 3.2), whereas the short H3K27me3 peptide used in these assays did not (Figure 3.4 F, in green). Similarly, the PRC2 allosteric inhibitor A395 [43] does not bind to the RNA-binding surface (Figure 3.11 I) and it did not compete with RNA binding (Figure 3.11 G, H).

These data indicate that stimulatory peptides of PRC2 relieve the inhibitory activity of RNA, possibly through allosteric modulation and, at least for JARID2-K116me3, in part through competition with RNA at the regulatory center.

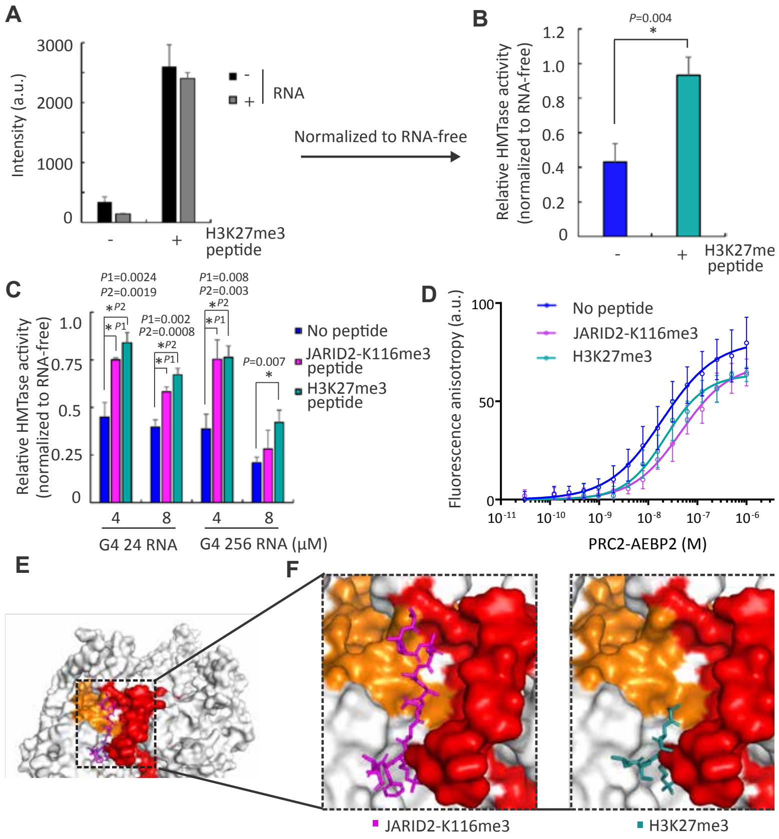


FIGURE 3.4 Stimulatory peptides relieve the RNA-mediated inhibition of PRC2

(A) HMTase assays of PRC2 in the presence (+) or absence (–) of 80 μM H3K27me3 peptide and in the presence (+) or absence (–) of 4.0 μM G4 256 RNA. (B) HMTase activities of PRC2 in its basal and stimulated states, relative to the HMTase activity of an RNA-free PRC2 within the same state: bar plot based on the same data as in (A) after normalizing each RNA-containing sample (gray bars in (A)) to the corresponding RNA-free sample (black bars in (A)) to yield the relative HMTase activity of PRC2 in either its stimulated (H3K27me3 peptide, in green) or basal (no peptide, in blue) state. (C) HMTase activities, relative to an RNA-free sample, of PRC2–AEBP2 in its stimulated (10 μM JARID2-K116me2 peptide, in magenta, or 80 μM H3K27me3 peptide, in green) or its basal (no peptide, in blue) state and in the presence of RNA as indicated (relative activities were calculated as in (B). Error bars in (A)-(C) represent standard deviations based on three independent experiments. All bar plots are represented means. P values were determined using unpaired two-tailed Student's t-test; *P $<$ 0.05. (D) The affinity of the PRC2–AEBP2 complex to G4 24 RNA was quantified using fluorescence anisotropy in the presence or absence of 100 μM H3K27me3 or 10 μM JARID2-K116me3 peptides. The KCl concentration in the binding buffer was reduced to 100 mM (rather than 200 mM KCl that was used in Figure 3.2 B,C) to mimic the conditions used in the HMTase assays presented in this figure. Error bars represent standard deviations in 11, 6 and 4 independent replicates for the binding curves plotted in blue, purple and green, respectively. Values are represented means. See Table 3 for dissociation constants and Hill coefficients. (E) The stimulatory peptides' binding sites in PRC2 (coordinates: PDB 6C23): Orange and red represent RNA-linked polypeptides (color code as in Figure 3.3 B), after superimposing the JARID2-K116me3 peptide (magenta, from PDB: 6C23) and the H3K27me3 peptide (dark green, from PDB: 3IIW). (F) Close-up of the two peptides' binding sites. Source data are available in Digital Supplemental Table S15. Data and analysis generated by collaborators.

3.3.5 RNA inhibits PRC2 in a DNA-independent manner

The RNA binding sites that we identified in PRC2 using RBDmap and targeted RBR-ID are adjacent to and, in some cases, overlap the substrate-binding site in the methyltransferase center (Figure 3.12 A). We therefore wished to determine if RNA can inhibit the methyltransferase activity of PRC2 through a mechanism other than competition for nucleosomes [36] or DNA binding [15]. We repeated the HMTase assays using DNA-free H3 histones as substrates (Figure 3.5 A), rather than nucleosomes. In agreement with the hypothesis of a DNA-independent RNA-mediated inhibition of PRC2, G4 256 RNA inhibited the activity of both PRC2.2 (Figure 3.5 A, left) and PRC2.1 (Figure 3.5 A, right) toward H3 substrate. Unlike the longer G4 256 RNA, the G4 24 RNA did not inhibit PRC2 activity toward H3 histones (Figure 3.12 D, E and Digital Supplemental Table S12), suggesting that its observed inhibitory effect in the context of nucleosomal substrates (Figure 3.4 C) relies on competition with nucleosomal DNA [15]. This indicates that although a 24-base-long G-quadruplex RNA is sufficient to bind PRC2 (Figure 3.2 and [34]) and inhibit HMTase toward nucleosome substrates (Figure 3.4 C), longer RNAs might be required for methyltransferase inhibition toward non-nucleosome substrates. Contributing factors would probably be an increased affinity of the long RNA for PRC2 (G4 256 RNA, Kd=1.09±0.13nM; Figure 3.10 K, L) over the short RNA (G4 24 RNA, Kd = 19.2 ± 2.7 nM; Figure 3.4 D and Table 3.3) and additional steric hindrances that are probably offered by the longer RNA. Although PRC2 activity toward H3 histones outside nucleosomes has limited biological significance, these data indicate that RNA-mediated HMTase inhibition of PRC2 can take place independent of competition for DNA or nucleosome binding. Indeed, the JARID2-K116 unmethylated peptide, as well as the H3K27M oncogenic peptide that binds to the catalytic site but not the regulatory center of PRC2 ([10]; Figure 3.12 A), reduced the affinity of PRC2 for RNA (Figure 3.12 B, C). This suggests that these peptides can directly compete with RNA binding at the catalytic site of the complex in addition to the regulatory center, further supporting our mapping of protein-RNA interactions to this region (Figure 3.1 and Figure 3.13 E).

We next assayed the methyltransferase activity of PRC2 toward two non-histone substrates of PRC2—human TBP (hTBP) and mouse ID2 (mID2)[37] —and confirmed inhibition of PRC2 activity by RNA also for these substrates (Figure 3.5 B). Collectively, these results demonstrate that RNA can inhibit the methyltransferase activity of PRC2 toward a DNA-free substrate, including non-histone substrates.

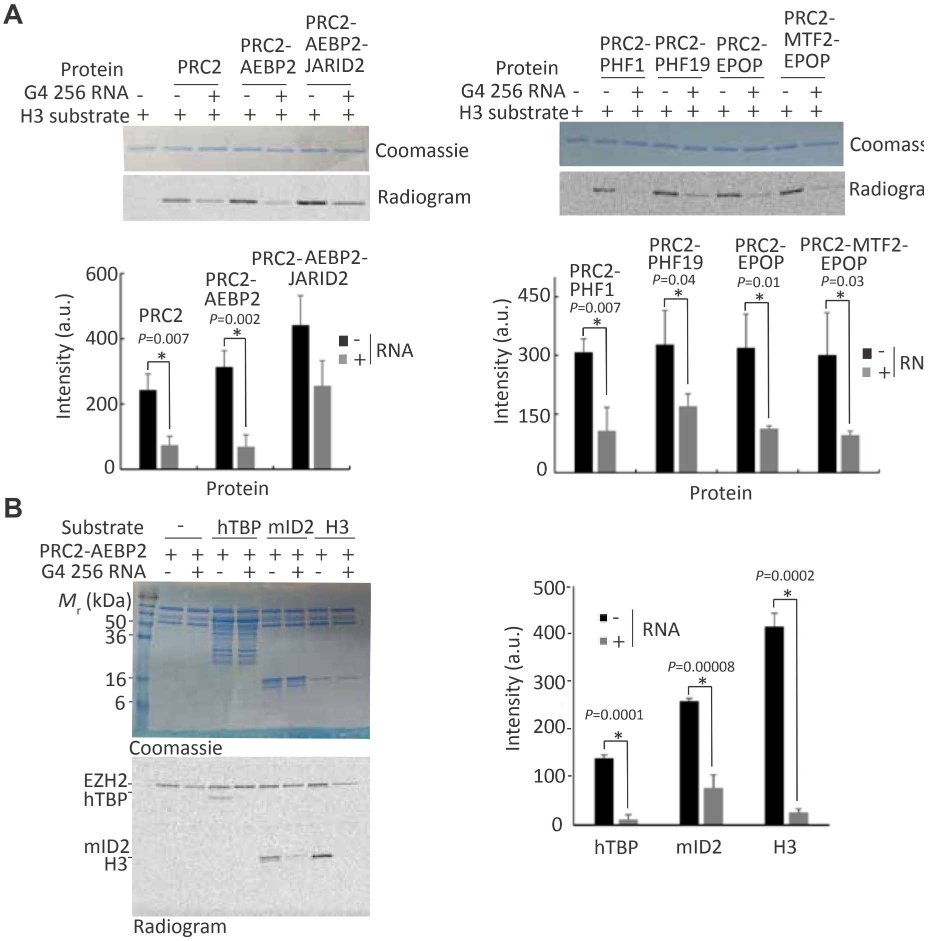


FIGURE 3.5 DNA-independent RNA-mediated inhibition of PRC2

(A) HMTase assays carried out in the presence of 0.5 μM PRC2 complexes as indicated, 4 μM H3 histone substrate and in the presence or absence of 1 μM G4 256 RNA. The bar plot (bottom) represents the activity, as recorded by densitometry after SDS–PAGE (top). (B) HMTase assays were carried out in the presence of 0.5 μM PRC2–AEBP2, 1 μM H3 or non-histone substrates human TBP (hTBP, 20 μM) or mouse ID2 (mID2, 15 μM), and in the presence or absence of 8 μM G4 256 RNA. The bar plot (right) represents the activity, as recorded by densitometry after SDS–PAGE (left). In all plots within the figure, bars represent means and error bars represent standard deviation based on three independent experiments and P values were determined using unpaired two-tailed Student's t-test; *P < 0.05. Complete gel scans are shown in Figure 3.12. Source data are available in Digital Supplemental Table S16. Data and analysis generated by collaborators.

3.3.6 The regulatory center and the RNA-binding site are exposed in PRC2.1

Since RNA binds and inhibits both PRC2.1 and PRC2.2 (Figure 3.2), we hypothesized that the PRC2.1 complex adopts a similar architecture to PRC2–AEBP2, leaving the regulatory site exposed. We mapped protein-protein interactions within PRC2–PHF19 and PRC2–MTF2–EPOP using bis(sulfosuccinimidyl)suberate (BS3) crosslinking with mass spectrometry (BS3 XL–MS, Figure 3.6 B, C). For a direct comparison, we also mapped interactions within the PRC2– AEBP2 complex (Figure 3.6 A), including the 216 amino acids that complete the N-terminal domain of the canonical AEBP2 isoform but were not included in previous structural investigations into PRC2–AEBP2 [5, 7]. Distances between crosslinked lysine pairs within PRC2 core subunits were measured using the high-resolution structure of the PRC2–AEBP2–JARID2 complex5 and resulted in similar distance distributions for the three complexes (Figure 3.13 A, B), supporting a similar structural organization of the core subunits in the PRC2.1 and PRC2.2.

We next mapped crosslinking sites of the accessory subunits PHF19, MTF2 and EPOP from the PRC2–PHF19 and the PRC2– MTF2–EPOP complexes to the core subunits (Figure 3.6 D, see Methods for a full description). We identified interactions between the C-terminal ‘reversed chromodomain’ (RC domain) [44] of the PCL proteins PHF19 and MTF2 to the C2 domain of SUZ12 (Figure 3.6 B, C), in good agreement with binding assays from previous studies [6, 44]. Importantly, no interactions were detected between PCL proteins and domains of core subunits within the catalytic lobe of PRC2. Although the C-terminal of EPOP crosslinked to residues within the catalytic lobe of PRC2, under the SRM of EZH2, amino acids in the regulatory center or the RNA-binding site were not crosslinked (Figure 3.6 C, D and Figure 3.13 C).

While the structure of the N-terminal portion of EED was never determined, our BS3 XL–MS results indicate that it resides within the vicinity of the N-terminal of EZH2 in all the examined PRC2 complexes (Figure 3.6 A–C). Similar crosslinking was previously observed in the PRC2–AEBP2–JARID2 complex [5]. In addition to mutual protein-protein crosslinks (BS3 XL–MS: green lines in Figure 3.6 A–C), the N-terminal portions of EED and EZH2 cluster multiple RNA–protein crosslinked peptides (RBDmap: blue and red spots in Figure 3.6 A). The simplest explanation for these observations is that the N-terminal regions of EZH2 and EED reside in close proximity and form a single RNA-binding site that is probably exposed in both the PRC2.1 and PRC2.2 complexes. This is in good agreement with the ability of an RNA containing short repeats of consecutive guanines to bind the two types of holo-PRC2 complexes with similar affinity and specificity, and to inhibit the methyltransferase activity of both of them (Figure 3.2), possibly through an identical mechanism.

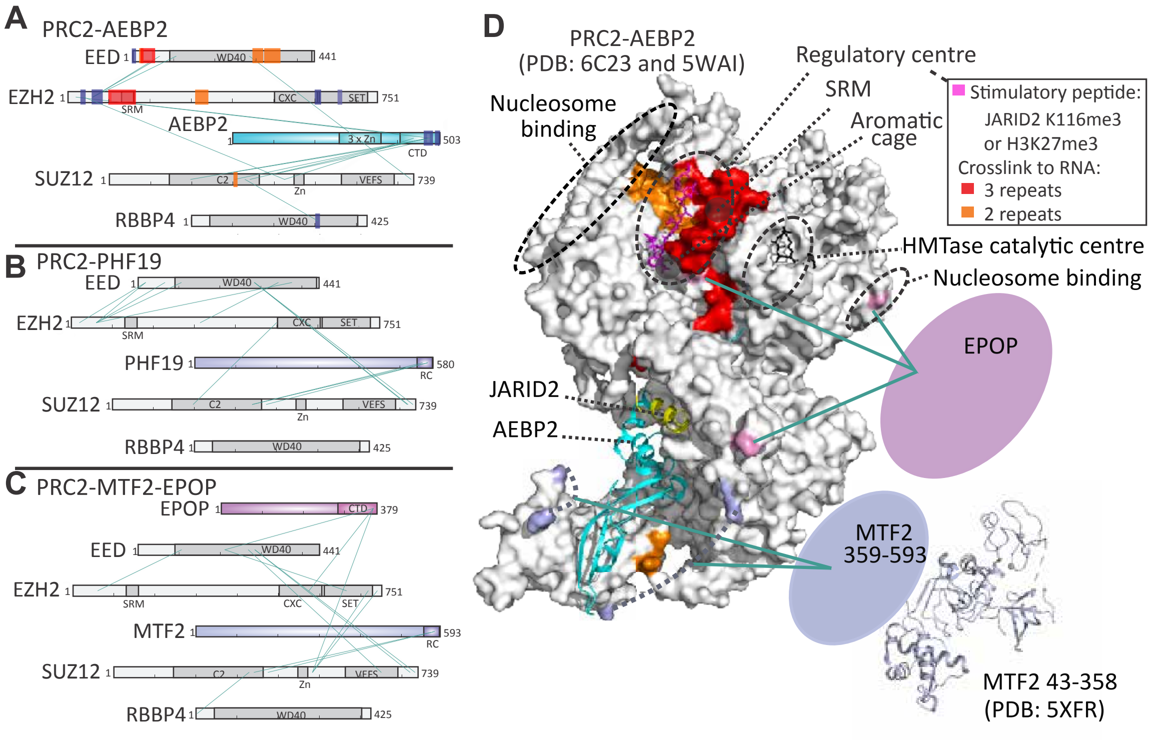


FIGURE 3.6 The RNA-binding site in the regulatory center of PRC2 is exposed within both PRC2.1 and PRC2.2

(A)-(C), BS3 XL–MS results for PRC2–AEBP2 (A), PRC2–PHF19 (B) and PRC–MTF2–EPOP (C). Core subunits are colored gray, accessory subunits are indicated in assorted colors and selected domains are shown in dark colors (see Figure 3.13 D, middle structure, for the same view with the core subunits in assorted colors). Green lines represent inter-molecular protein-protein BS3 crosslinks. Blue, orange and red boxes on the protein representation in a represent RNA–protein crosslinks that were identified in 1, 2 or 3 independent RBDmap experiments, respectively (same data as in the three-dimensional representation in Figure 3.3 B). (D) Accessory proteins and RNA-binding sites within the holo-PRC2 complex: surface view of PRC2 was generated as in Figure 3.3 B. AEBP2 (cyan) and JARID2 (yellow) fragments are shown as a ribbon representation and the N terminus of MTF2 that was determined crystallographically (PDB: 5XFR) is in light blue, to approximate scale. EPOP (pink) and the C-terminal region of MTF2 (light blue) are indicated as blobs, to approximate scale. Green lines indicate crosslinks between PRC2 core subunits to PCL proteins and EPOP. Residues within core PRC2 subunits that were crosslinked to EPOP are indicated in pink. Residues within core PRC2 subunits that reside at the termini of unstructured loops that were crosslinked to PCL proteins are indicated in light blue and linked with dashed arcs. Protein-RNA contacts that were determined in two or three independent RBDmap replicates (see Figure 3.3 for complete data) are indicated in orange and red, respectively. Other key functional centers or structural features are highlighted using dashed black circles. See Figure 3.13 for distance histograms of BS3 XL–MS and different views of the structure presented in (D). Data and analysis generated by collaborators.

3.4 Discussion

The recruitment of PRC2 to chromatin and its regulation at target genes are determined by interactions with multiple factors, including accessory proteins, specific DNA sequences and RNA [2]. Among the protein factors, unbiased proteomic approaches identified JARID2, AEBP2, EPOP, PALI and PCL proteins as key accessory subunits that were reproducibly identified across studies and experimental systems [20–22, 45], albeit in relative abundances that change across developmental stages [20]. Our study indicates that, regardless of subunit composition, RNA binds to and inhibits various subtypes of PRC2 complexes in vivo and *in vitro* (Figs. 1 and 2). Mechanistically, this is achieved through exploiting multiple surfaces on the catalytic lobe of the core complex, including the regulatory center formed at the interface of EED and EZH2 (Figure 3.3), which is exposed both within the PRC2.1 and PRC2.2 (PRC2–AEBP2–JARID2) complexes (Figure 3.6). It implies that PRC2 can bind RNA throughout various stages of development, even when the composition of its accessory subunits varies significantly (Figure 3.7 B).

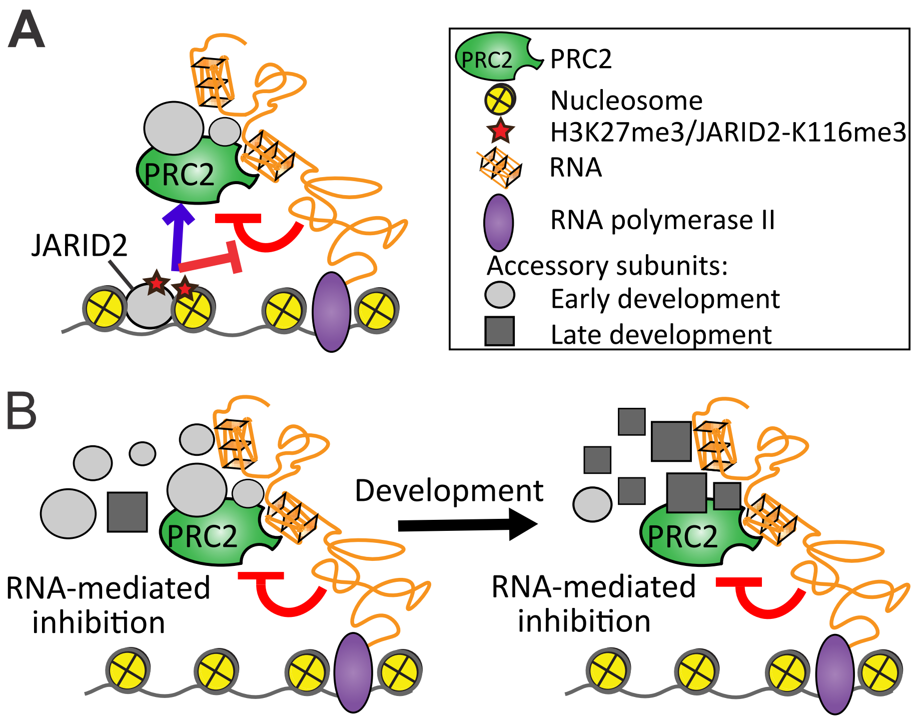


FIGURE 3.7 A model for RNA-mediated inhibition of PRC2

(A) Stimulatory effectors of PRC2—JARID2-K116me3 and H3K27me3—relieve the inhibitory activity of RNA simultaneously with HMTase stimulation. This process provides a molecular mechanism to overcome RNA- mediated inhibition during the nucleation, spreading and propagation of the H3K27me3 mark at polycomb-target genes [8, 11, 46]. (B) A model for RNA-mediated inhibition of PRC2 during development: RNA binds to the allosteric regulatory center of PRC2 and inhibits methyltransferase activity toward histone and non-histone substrates, either when PRC2 is in complex with AEBP2 and JARID2 (PRC2.2) or PHF1, PHF19, MTF2, PALI and EPOP (PRC2.1). RNA-mediated inhibition of PRC2 provides a fail-safe mechanism to prevent substrate methylation by RNA-bound PRC2 at non- target genes, even if the stoichiometry of its common accessory subunits changes during development.

3.4.1 One face of PRC2 clusters binding sites for multiple regulatory factors

Our analysis of PRC2.1 complexes, together with data from previous structural investigations into the architecture of the PRC2–AEBP2–JARID2 complex [5, 6], indicates that binding sites for multiple ligands and factors cluster on one face of PRC2 (Figure 3.6 D and Figure 3.13 C, D). This face also contains binding sites for the PRC2.1 and PRC2.2 accessory subunits that regulate the recruitment of PRC2 to chromatin [1–3]. The same face also includes the catalytic site and the allosteric regulatory center that stimulates HMTase activity upon binding of methylated H3 or JARID2 peptides [8, 11, 12]. All these stimulatory, regulatory and catalytic modules are concentrated on the same face, along with the most prominent RNA-binding region identified by both RBR-ID and RBDmap (Figure 3.13). This structural organization might provide RNA with a simple means to block methyltransferase activity and to simultaneously interact with other RNA-binding regions of EZH2 such as the CXC and SET domains [35], as well as with the accessory subunits JARID2 [28], AEBP2 [38], MTF2 (Figure 3.1) and possibly other PCL proteins (Figure 3.2).

3.4.2 Interplay of RNA and stimulatory peptides at the allosteric regulatory center

Amino acids of EED and EZH2 involved in allosteric activation of PRC2 [8–11], such as for example EZH2 F145 [12], reside within the main RNA-binding site identified by RBR-ID and RBDmap or in its immediate vicinity (Figure 3.6). The RNA-binding site is also in the immediate vicinity of the SRM (Figure 3.6 D), which stabilizes the methyltransferase center within EZH2 [9] during allosteric activation [8]. Here, we report that allosteric stimulation of PRC2 though either H3K27me3 or JARID2-K116me3 peptides relieve RNA-mediated inhibition (Figure 3.4), providing a direct link between the RNA-binding site that we identified in the regulatory center of PRC2 and the process of effector-induced stimulation.

3.4.3 The methylated form of JARID2 relieves RNA-mediated inhibition of PRC2

The JARID2-K116me3 stimulatory peptide was assigned a significant role in nucleating the H3K27me3 mark [11], which in turn stimulates PRC2 selectively at repressed polycomb-target genes [8] to allow for the nucleation and, eventually, spreading and propagation of the H3K27me3 mark [8, 11, 46]. It is plausible that RNA-mediated inhibition of PRC2 is suspended at new PRC2 target sites through stimulatory interactions with the JARID2-K116me3 moiety (Figure 3.7 A).

The PRC2-JARID2 complex, in the absence of AEBP2 and when compared to PRC2, was previously reported to have reduced affinity for RNA and increased HMTase activity in the presence of RNA [30], suggesting that JARID2 weakens PRC2-RNA interactions and relieves catalytic inhibition. Yet these results were seemingly in contrast with those of two other studies: one showing that inclusion of JARID2 in the PRC2-AEBP2 complex does not alter its RNA-binding activity [15] and the other demonstrating that JARID2 binds RNA in vivo and that an RNA-binding region within JARID2 is required for the recruitment of PRC2 to target genes [28]. Our data herein reconciles these findings and is consistent with a model whereby a PRC2-AEBP2-JARID2 complex can bind RNA and be inhibited by it (Figure 3.2), but interactions with the methylated form of JARID2 (JARID2-K116me3) relieve RNA-mediated inhibition (Figure 3.4).

3.4.4 Multi-modal inhibition of PRC2 by RNA

The mechanistic details of RNA-mediated PRC2 inhibition are of great importance to understand the biological meaning of the still poorly understood PRC2-RNA interactions. The fact that RNA does not inhibit the prominent automethylation activity of EZH2 (Figure 3.9 A, B and 3.5 and ref. [15]) previously suggested a model whereby RNA inhibits PRC2 by competing with nucleosomal DNA for binding [15]. Indeed, some of the amino acids that were altered within our mt1 and mt3 mutants, which showed decreased RNA affinity, were previously identified as part of a nucleosome-binding site [47]. Yet, most of the mutated amino acids that affect RNA binding in our experiments (Figure 3.3 D and Table 3.2) do not interact with nucleosomes [47], but rather reside within the regulatory center or its immediate vicinity (Figure 3.3 E). Moreover, recent investigations into the automethylation of EZH2 [48, 49] suggest that it occurs intramolecularly (in cis). Thus, this activity benefits from a high local substrate concentration that could overcome RNA-mediated inhibition. It is thus possible that additional bases, beyond the mere G-quadruplex motif, sterically block the substrate-binding site of PRC2. Indeed, a direct—and possibly inhibitory—interaction between RNA and the SET domain in vivo is supported by our targeted RBR-ID data (Figure 3.1 C, D), *in vitro* by the competition of JARID2 and H3 substrate peptides for RNA binding—presumably at the catalytic site (Figure 3.12 B, C)—and indirectly by hydrogen deuterium exchange results from an independent study [35]. Therefore, RNA can inhibit PRC2 through different mechanisms, including competition for nucleosome binding [36], competition for DNA binding [15] and through blocking methyltransferase activity directly.

3.4.5 Implications for RNA-mediated regulation of PRC2 in vivo

The observation that RNA inhibits the methyltransferase activity of PRC2 independent of competition with nucleosomes or DNA impacts previously proposed models for RNA-mediated regulation of PRC2, especially those envisioning long non-coding RNAs that recruit PRC2 to target genes through direct interactions (reviewed in [23, 25, 26, 50]). These models had already been challenged (see [26] for a critical review) by recent observations that RNA prevents PRC2 from binding nucleosomes [36] or DNA [15], but the possibility remained that lncRNA-PRC2 interactions might tether PRC2 at target genes long enough to nucleate K27 methylation, even while DNA binding was inhibited by the presence of RNA. However, our observations show that the methyltransferase activity of PRC2 is inhibited independently of its ability to bind DNA, suggesting that even after being recruited to a given target gene, the presence of the RNA would continue to inhibit H3K27me3 deposition, unless RNA-mediated inhibition is relieved by a stimulatory effector. However, it is possible that some transcripts regulate PRC2 function via a different mechanism. Specifically, we note that despite a striking overlap of RNA-binding signal on the EED 336-355 peptide, several peptides recovered in vivo by RBR-ID on various subunits were not detected by RBDmap in vitro. It is tempting to speculate that RBR-ID-only peptides comprise regions of PRC2.1 and PRC2.2 that interact with RNAs that we did not test in vitro. Thus, other RNAs might have different regulatory functions in the context of Polycomb silencing. A thorough investigation of separation-of-function mutants suggested by our mapping experiments will be required to test this intriguing hypothesis.

Future studies may reveal if the inhibitory activity of RNA binding to PRC2 could be relieved by specific transcripts, during specific stages of development, in disease states or at specific loci, and possibly through the involvement of cell-type-specific factors beyond the most abundant accessory subunits of PRC2 [20–22]. The location of an exposed RNA-binding site within the allosteric center of the two common forms of holo-PRC2 complexes provides means for RNA-mediated regulation of PRC2 in most cell lineages and during most stages of normal development and suggest a mechanism by which stimulatory peptides relieve RNA-mediated inhibition to allow de novo methylation at PRC2 target loci.

3.5 Acknowledgments

We would like to thank the Monash Biomedical Proteomics Facility for providing instrumentation and technical support. Q.Z. holds an Australian Research Council (ARC) Discovery Early Career Researcher Award (no. DE180100219). N.J.M. is the Isabella and Marcus Foundation Charlee Ferrar Scholar and is also supported through an Australian Government Research Training Program (RTP) Scholarship. R.W.-T. was supported by NIH training grant no. T32GM008216. E.H.G. holds a Biomedicine Discovery Scholarship and is an EMBL-Australia PhD student. B.M.O. is supported through an Australian Government RTP Scholarship and also by the Monash Graduate Excellence Scholarship. R.B. acknowledges support from the NIH (grant no. R01GM127408) and the March of Dimes Foundation (grant no. 1-FY-15–344). C.D. is an EMBL-Australia Group Leader and acknowledges support from the ARC (grant no. DP190103407) and the NHMRC (grant no. APP1162921).

3.6 Author contributions

Q.Z., N.J.M., R.W.-T., S.F.F. and B.M.O. prepared reagents. Q.Z., N.J.M., R.W.-T., S.F.F., B.M.O., R.L., V.L. and R.B.S. carried out experiments. Q.Z., N.J.M., R.W.-T., E.H.G., S.F.F., B.M.O., R.L., V.L., B.A.G. and R.B.S. analyzed data. Q.Z., N.J.M., R.W.-T., R.B. and C.D. wrote the paper. R.B. and C.D. designed and supervised the project. Specifically, R.W-T. performed all experiments and data analysis for RBR-ID experiments (Figure 3.1, 3.8. Text in this chapter was adapted from published manuscript by R.W-T. with minor modifications.

3.7 Disclosures

The authors have no conflicts of interest to disclose.

3.8 Methods

3.8.1 Targeted RBR-ID

For label-free experiments, mESCs were cultured on gelatincoated dishes in KnockOut DMEM (Gibco, no. 10829018) supplemented with 15% FBS (Gibco, no. 10437028), 100 mM non-essential amino acids (Sigma, no. M7145), 0.1 mM 2-mercaptoethanol (Gibco, no. 21985023), 1 mM l-glutamine (Sigma, no. G7513), 100 U ml–1 leukemia inhibitory factor (Millipore, no. ESG1107), 3 μM CHIR99021 (Millipore, no. 361559), 1 μM PD0325901 (Millipore, no. 444966), 50 U ml–1 penicillin and 50 μg ml–1 streptomycin.

For SILAC-assisted quantification, cells were cultured in medium containing 15% dialyzed FBS (Thermo Fisher Scientific, no. 88440), 2 mM proline and 0.47 mM conventional or heavy isotope-labeled amino acids (Arginine-10 + Lysine-8). Following several passages in heavy media, cell extracts were analyzed via mass spectrometry and only used if ¿98% labeling could be confirmed. Cells were crosslinked as previously described [38, 51]. Briefly, cells treated for 2 h with 500 μM 4-thiouridine (4SU) were crosslinked with 1 J cm–2 UVB light. SILAC-labeled cell extracts were prepared as follows: cells were lysed in 50 mM Tris (pH 8 at 25 °C), 150 mM NaCl, 1% IGEPAL CA-630, 0.2 mM EDTA, 2 mM MgCl2, 1 μg ml–1 aprotinin, 1 μg ml–1 leupeptin, 1 μg ml–1 pepstatin and 0.2 mM PMSF, then treated with 2,500 U ml–1 Pierce Universal nuclease (Thermo Fisher Scientific, no. 88700) for 30 min at 25 °C. Extracts were sonicated briefly, then NaCl was added to bring the final concentration to 300 mM and extracts were rotated at 4 °C for 30 min to complete lysis. Extracts were centrifuged at 18,000g for 10 min at 4 °C and supernatants were collected. We quantified cell extracts via Bradford protein assay and prepared 1:1 mixtures of ± 4SU-treated heavy and light labeled extracts. For label-free quantification replicates, nuclear extracts were prepared from cells as previously described [38, 51]. Polyclonal anti-EZH2 antibody was generated by immunizing rabbits with a fragment of mouse EZH2 spanning amino acids 1 to 370 (Uniprot Q61188) and affinity purifying using the same antigen. Immunoprecipitations were performed by adding EZH2 antibody to extracts and incubating overnight at 4 °C, then recovering protein–antibody complexes with protein G Dynabeads (Thermo Fisher Scientific, no. 10003D) pre-blocked with 1 mg ml–1 BSA. Beads were washed three times with immunoprecipitation wash buffer (20 mM Tris (pH 7.9 at 4 °C), 0.2 mM EDTA, 200 mM KCl and 0.05% IGEPAL CA-630). To elute bound proteins, 0.1 M glycine (pH 2.4) was added to beads and incubated at 25 °C with gentle shaking for 10 min, then the eluate was transferred to a fresh tube containing 0.1 M Tris (pH 8 at 25 °C) to neutralize. Fractions of eluted proteins were taken for western blot, and the remainder diluted in trypsin digestion buffer (final: 50 mM NH4HCO3 (pH 8), 1 mM MgCl2 and 1 mM CaCl2) or chymotrypsin digestion buffer (final: 100 mM Tris (pH 8 at 25 °C), 1 mM MgCl2 and 10 mM CaCl2). To remove crosslinked RNA, 12.5 µg ml–1 RNase A was added to samples and incubated for 30 min at 37 °C. To reduce proteins, samples were treated with 5 mM dithiothreitol for 60 min at 25 °C, then cysteines were alkylated with 14 mM iodoacetamide for 30 min in the dark, before additional reduction with 5 mM dithiothreitol for an additional 15 min in the dark. Proteins were digested with trypsin or chymotrypsin at a protease:sample ratio of 1:20 and incubated overnight at 37 °C or 25 °C, respectively. Peptides were quenched with formic acid and desalted on C18 stage tips, then dried and resuspended in 0.1% formic acid before mass spectrometry analysis.

For mass spectrometry, a reverse phase gradient with a nano-LC was performed on a C18 column with a 2–60% binary gradient (mobile phase A: 0.1% formic acid in aqueous; mobile phase B: 80% acetonitrile, 0.1% formic acid) for 90 min. The gradient continued to 95% over 2 min and held for 13 min at 95% mobile phase B. MS was performed using a Thermo Orbitrap Fusion instrument and MS/MS data were collected in centroid mode using the Orbitrap mass analyzer. Fragmentation of peptides for MS/MS was performed using higher energy collisional dissociation and only charge states of 2–5 were included for fragmentation. MS/MS spectra were processed through MaxQuant [52] using a FASTA file comprising PRC2 complex proteins. SILAC and unlabeled samples were processed with the same parameters. To generate the protein-level analyses shown in Digital Supplemental Table S10, MS/MS spectra were processed using a mouse proteome FASTA file.

3.8.2 RBR-ID analysis

After removal of suspected contaminants, MaxQuant peptide abundances were normalized by the mean of all peptide intensities in each MS run, or in the case of SILAC data by the mean heavy or light labeled peptide intensity in each run. For each peptide, a $\log_2$-converted ratio was calculated between samples treated with or without 4SU to assess depletion mediated by RNA-crosslinking.

P values for peptides observed only in SILAC samples were analyzed via a paired, two-sided Student's t-test, while an unpaired two-sided Student's t-test was performed for peptides common between SILAC and label-free quantification samples, to account for missing values in the data matrix. RBR-ID scores, which reflect both the degree and consistency of 4SU-mediated depletion of a peptide38, were calculated as follows. To visualize targeted RBR-ID scores on the high-resolution structure of the human PRC2–AEBP2, RBR-ID scores at each residue were calculated as the sum of the RBR-ID scores of all overlapping identified peptides, and then mouse peptide sequences were aligned to the corresponding human protein sequence using ClustalOmega. Since a large number of amino acids were not resolved in the high-resolution cryo-EM structure of the PRC2–AEBP2– JARID2 complex (PDB: 6C23)5, and in order to allow for maximal coverage of this complex, the crystal structure of SUZ12–RBBP4–JARID2–AEBP2 (PDB: 5WAI)[6] was superimposed on this structure and the non-catalytic lobe of 6C23 was omitted, with the exception of SUZ12 amino acids 497–518, which is absent in 5WAI6.

3.8.3 Protein expression and purification

The full-length sequences encoding human EZH2, SUZ12, RBBP4, EED, AEBP2 and JARID2 (UniProtKB: Q15910-2, Q1502-1, Q0902-1, O7553-1, Q6ZN18-2 and Q92833-1, respectively) were cloned into a pFastBac1 expression vector with PreScission-cleavable N-terminal hexahistidine-MBP tags as previously described [19, 20]. The full-length sequences of PHF1 (O43189-2), MTF2 (Q9Y483-1), PHF19 (Q5T6S3-1) and EPOP (A6NHQ4-1) were synthesised (see Digital Supplemental Table S17) and sub-cloned by Gen9 using XmaI and XhoI sites into the expression vector pFB1.HMBP.A3.PrS.ybbR 20 (derived from the pFastBac1 vector), under the N-terminal hexahistidine-MBP tag and PreScission-cleavable sites (both cohesive ends introduced to the insert using BsaI). For expression and purification of the PRC2-PHF1 and PRC2-AEBP2-JARID2 complexes, genes encoding for PRC2 core subunits EZH2, EED, SUZ12, RBBP4 and the accessory subunit AEBP2 (UniProtKB accession as above) were sub-cloned using the Gibson Assembly® Master Mix (NEB #E2611L) into a pFBOH-MHL baculovirus expression vector (a gift from the lab of Dr. Yufeng Tong, University of Toronto, Addgene #62304) that was digested using BseRI, to obtain a fusion with a TEV-cleavable N-terminal hexahistidine tag (see Digital Supplemental Table S17 for cloning primers). Mutation plasmids of N-terminal hexahistidine-MBP-tagged EZH2 and EED were generated by Pfu polymerase and Takara PrimeSTAR HS DNA Polymerase (Clontech #R045A). Baculovirus stocks were generated as per manufacturer's instructions (ThermoFisher). The titter of each baculovirus stock was quantified using the MTT assay (Promega #G3580). Baculovirus stocks were combined according to the optimal ratio and used to co-infect Trichoplusia ni insect cells at 2 × 106 cells/mL in Insect-XPRESS media (Lonza #12-730Q). Infected cells were incubated for 64 hours at 27 °C and 110 rpm before harvesting by centrifugation (20 min at 4 °C using Beckman JLA-8.1000 rotor at 1500 RCF). The harvested cells were snap-frozen with liquid nitrogen and stored at -80 °C until purification.

For the purification of all PRC2 complexes, with the exception of the PRC2-PHF1 and PRC2-AEBP2-JARID2 complexes (see below), harvested cells were lysed in an ice-cold buffer containing 10 mM Tris-HCl pH 7.5 at 25 °C, 250 mM NaCl, 0.5 % Nonidet-P40 (NP40), 1 mM TCEP and protease inhibitor cocktail (200X ethanol solution containing 30 g/L PMSF, 0.25 g/L Pepstatin A, 0.05 g/L Leupeptin hemisulfate salt and 60 g/L Benzamidine·HCl). The cleared lysates of all the other PRC2 complexes, with the exception of PRC2-PHF1 and PRC2-AEBP2-JARID2, were batch-bound to amylose resin (NEB #E8021), washed using 10 column volumes (c.v.) of ice-cold lysis buffer, 10 c.v. of high salt wash buffer (10 mM Tris-HCl pH 7.5 at 25 °C and 500 mM NaCl) and 10 c.v. of low salt wash buffer (10 mM Tris-HCl pH 7.5 at 25 °C and 150 mM NaCl) before proteins were eluted using ice-cold MBP elution buffer (10 mM Tris-HCl pH 7.5 at 25 °C, 150 mM NaCl, 10 mM maltose and 1 mM TCEP). The eluents were supplemented with NaCl, to a final concentration of 250 mM, before incubation with PreScission protease overnight at 4 °C.

Further purification of tag-free complexes was performed at 4-8 °C, as previously described [21]. In brief, after tag-cleavage complexes were loaded onto a 5 mL HiTrap Heparin HP column (GE #17040701) in buffer A (10 mM Tris-HCl pH 7.5 at 25°C and 150 mM NaCl) and were eluted over a 10 c.v. gradient into 50% buffer B (10 mM Tris-HCl pH 7.5 at 25 °C, 2 M NaCl). Complex-containing fractions were concentrated using 30 kDa Amicon Ultra centrifugal filter and loaded onto a HiPrep 16/60 Sephacryl S-400 HR size exclusion column (GE Healthcare) and fractionated using a buffer containing 20 mM HEPES (pH 7.5 at 25°C), 200 mM NaCl and 1 mM TCEP. The peak fractions were pooled, concentrated using a 30 kDa Amicon Ultra centrifugal filter, snap-frozen in liquid nitrogen and stored at -80 °C as single-use aliquots.

The PRC2-PHF1 and PRC2-AEBP2-JARID2 complexes were purified by tandem affinity purification with a workflow that was designed to allow for near-stoichiometric incorporation of the accessory subunit of interest: first, Ni-NTA agarose was used to purify all complexes; in the second affinity chromatography step, MBP tags on the accessory subunits were used for affinity purification of nearly-stoichiometric complexes, using amylose resin; next, complete removal of hexahistidine-MBP tags from the accessory subunits and hexahistidine tags from the core subunits was done using PreScission and TEV proteases, respectively, and unincorporated subunits and cleaved tags were next removed using subsequent Heparin affinity and gel filtration chromatography. Specifically, harvested cells were lysed in an ice-cold lysis buffer supplemented by 10 mM imidazole before being applied to Ni-NTA resin (Qiagen #30210). The resin was washed using 10 c.v. of lysis buffer as above, 10 c.v. of high salt wash buffer as above, supplemented by imidazole to 25 mM, and low salt wash buffer as above, then proteins were eluted with 5 c.v. of elution buffer (10 mM Tris-HCl pH 7.5 at 25 °C, 150 mM NaCl, 250 mM imidazole and 1 mM TCEP). The eluents were subsequently loaded onto amylose resin by gravity flow, washed with 5 c.v. of low salt wash buffer and eluted with MBP elution buffer, as above. Eluate was supplemented with NaCl to a final concentration of 250 mM before adding TEV and PreScission proteases for tag-cleavage, followed by overnight incubation at 4 °C. Subsequent Heparin affinity and gel filtration chromatography was performed as above.

Plasmids for FLAG-tagged hTBP (N-terminal FLAG tag, pFastBac1 backbone) and mID2 (C-terminal FLAG tag, pFastBac1 backbone) were a gift from the lab of Dr. Robert Kingston, Harvard University. Baculovirus stock preparation and mID2-flag protein expression was carried out as described above. Next, cells were harvested as above and snap-frozen in liquid nitrogen before purification using M2 agarose beads (Sigma-Aldrich #A2220) as previously described [22]. The eluent from the M2 beads was buffer exchanged by centrifugation using a 10 kDa Amicon Ultra centrifugal filter into buffer containing 20 mM Tris-HCl (pH 7.5 at 25°C), 500 mM NaCl and 1 mM TCEP and stored at -80 °C before use. The ratio of absorbance at 260 nm to 280 nm was 0.58, indicating no nucleic acid contamination.

For the purification of hTBP, human TBP gene was sub-cloned from the baculovirus expression plasmid above to the pGEX-MHL E. coli expression vector (a gift from the lab of Dr. Yufeng Tong, University of Toronto) with an N-terminal GST tag (vector was digested using BseRI and sequences of the cloning primers are specified in Digital Supplemental Table S17) and cloning was done using Gibson Assembly® Master Mix. The resulting plasmid, coding for GST-hTBP fusion, was transformed into the BL21(DE3) E.coli strain and grown in LB media at 37 °C, up to an OD600 of 0.6 where induction was carried out using 0.2 mM IPTG. Protein expression was carried out for 14-18 h at 17 °C with shaking at 180 rpm. Harvested cells were lysed by sonication in an ice-cold buffer containing 10 mM Tris-HCl (pH 7.5 at 25 °C), 500 mM NaCl, 1 mM DTT, 10% glycerol and 1 mM PMSF. The cleared lysate was applied to glutathione-agarose beads (Sigma-Aldrich #G4510) for batch purification, washed with 10 c.v. of lysis buffer and elution was carried out using 5 c.v. of lysis buffer supplemented with 10 mM reduced glutathione. The protein was further purified on a Superdex 200 10/300 GL size exclusion column (GE Healthcare) at 4-8 °C, in buffer containing 20 mM HEPES (pH 7.5 at 25 °C), 200 mM NaCl and 1 mM TCEP to separate out soluble aggregates. Protein-containing fractions were pooled, concentrated and stored at -80 °C before use. The ratio of absorbance at 260 nm to 280 nm was 0.59, indicating no nucleic acid contamination.

3.8.4 Fluorescence anisotropy assay

The 3' fluorescein labeled G4 24 RNA $(UUAGGG)_4$ and G4 mt 24 RNA $(UGAGUG)_4$ were synthesized by Integrated DNA Technologies, Inc. RNA was incubated for 2 min at 95 °C in 10 mM Tris-HCl pH 7.5 (at 25 °C) and was then immediately snap-cooled on ice for 2 min. Next, RNA was allowed to fold for 30 min at 37 °C in binding buffer (50 mM Tris-HCl pH 7.5 at 25 °C, 200 mM KCl, 2.5 mM MgCl2, 0.1 mM ZnCl2, 2 mM 2-mercaptoethanol, 0.1 mg ml–1 bovine serum albumin (NEB, no. B9000S), 0.05% Nonidet P40 (Roche, no. 11754599001) and 0.1 mg ml–1 fragmented yeast tRNA (Sigma, no. R5636)).

For assaying the RNA-binding affinity in the presence or absence of stimulatory or substrate peptides, binding buffer was adjusted to mimic conditions used for HMTase assays with these peptides and consisted of 50 mM Tris-HCl pH 7.5 (at 25 °C), 100 mM KCl, 2 mM 2-

mercaptoethanol, 0.05% Nonidet P40 (Roche, no. 11754599001) and 0.1 mg ml–1 bovine serum albumin (NEB, no. B9000S). Serial dilutions of the protein were made separately, in the same buffer, and combined with the RNA solution for a final reaction volume of 40 µl containing 5 nM fluorescently-labeled RNA and the desired final protein concentration. Samples were equilibrated at 30 °C for 30 min before measurement. Fluorescence anisotropy data were collected using a PHERAstar plate reader (BMG Labtech) at 30 °C (excitation wavelength λex = 485 nm, emission wavelength λem = 520 nm). The background was subtracted from protein-free samples. Kd, Hill and standard error values were calculated with GraphPad Prism 7 software using non-linear regression for specific binding with Hill slope function. For comparison between different complexes, non-linear regression was carried out using the background-subtracted anisotropy values. For comparison between wild type and mutants of the same complex, RNA fraction bound was calculated using the background-subtracted anisotropy of the fully-bound wild-type protein sample, which were run on the same plate. Independent replicates of all fluorescent anisotropy experiments were performed on different days.

3.8.5 Electrophoretic mobility-shift assay (EMSA)

EMSA was performed using ^{3}2P end-labeled RNA as previously described [34, 53].

3.8.6 Mapping protein-RNA interactions of an *in vitro* reconstituted complex using RBDmap

The RBDmap workflow was adopted from [42] and was modified to allow for mapping of protein-RNA interactions in an *in vitro* reconstituted complex. DNA templates for *in vitro* transcription of G4 256 RNA were generated through PCR amplification using a synthetic gene (GenScript, see Digital Supplemental Table S17 for the sequence) and forward and reverse primers that were designed to form a 5’ T7 promoter and a 3’ 25-mer polyA region, respectively (see Digital Supplemental Table S17 for primer sequences). Thirty microliters of 5 µM RNA in 20 mM Tris-HCl (pH 7.5 at 25 °C) was incubated for 2 min at 95 °C and was then snap-cooled on ice for 2 min. Then 90 µl of ice-cold Milli-Q ultrapure water and 30 µl of ice-cold 5X RNA-binding buffer (250 mM Tris-HCl (pH 7.5 at 25 °C), 500 mM KCl, 12.5 mM MgCl2, 0.5 mM ZnCl2 and 10 mM 2-

mercaptoethanol) was added to a final volume of 150 µl. The RNA was then allowed to fold at 37 °C for 30 min. Separately, PRC2–AEBP2 was prepared at 4 °C by adding purified protein stock to 1X binding buffer, as above, to a final volume of 150 µl. The entire RNA solution was then combined with the entire protein solution in a single well of a 24well plate to final concentrations of 0.5 µM G4 256 RNA and 1.0 µM PRC2–AEBP2. The mixture was then incubated at room temperature for 30 min to allow for the formation of ribonucleoprotein complexes. The 24-well plate was then placed onto an aluminum block within a container of ice and was irradiated by UV (254 nm, 6 rounds of 0.83 J cm^{-2} each) in a UVP CL-1000 Ultraviolet Crosslinker (Scientifix, no. CL-1000) with the top of the 24-well plate at a distance of 10 cm below five 8-watt tube lamps. Digestion of the crosslinked ribonucleoprotein complexes by ArgC (Promega, no. V1881) and LysC (NEB, no. P8109S) was then performed by adding 5 µl of 100 ng µL^{-1} ArgC or LysC stock solution to the appropriate samples, after which the samples were incubated at room temperature overnight. As input sample, 30 µl was taken and stored at 4 °C until RNase digestion of all samples was performed (see next paragraph below). The remaining sample was applied to 1 ml of oligo d(T)$_{25}$ magnetic beads (NEB, no. S1419S) that were pre-washed withand pre-incubated in 1X RNA-binding buffer for 10 min at 25 °C with gentle agitation. After applying the sample, the beads were incubated with gentle agitation for 10 min at room temperature, and then for 1 h at 4 °C before capture with a magnet. The supernatant was kept for SDS–PAGE analysis ('flow-through'). All steps from this point until the elution were done at 4 °C. After removal of the flow-through, the beads were transferred into a 50 ml conical tube and resuspended with 17 ml of buffer 1 (20 mM Tris-HCl pH 7.5 at 25 °C, 500 mM LiCl, 0.5% lithium dodecyl sulfate (w/v) (Sigma-Aldrich, no. L9781), 1 mM EDTA and 5 mM DTT) with gentle agitation for 5 min before capture by a magnet. The beads were then washed twice as above with bead buffer 2 (20 mM Tris-HCl pH 7.5 at 25 °C, 500 mM LiCl, 0.1% lithium dodecyl sulfate, 1 mM EDTA and 5 mM DTT) and twice with bead buffer 3 (20 mM Tris-HCl pH 7.5 at 25 °C, 500 mM LiCl, 1 mM EDTA and 5 mM DTT). Then the beads were washed twice with bead buffer 4 (20 mM Tris-HCl pH 7.5 at 25 °C, 200 mM LiCl, 1 mM EDTA and 5 mM DTT) as above, and residual bead buffer 4 was used to transfer beads to a 1.7 ml microcentrifuge tube before capture by a magnet, and the remaining bead buffer 4 was removed. RNA-linked polypeptides were then eluted by heating at 55 °C for 2 min in 300 µl of bead elution buffer (20 mM Tris-HCl pH 7.5 at 25 °C and 1 mM EDTA) before capturing the beads using a magnet, and transferring the supernatant (‘eluate’) to another tube. Beads were discarded and never reused in order to avoid cross-contamination.

The ‘input’ and ‘eluate’ samples were then processed for mass spectrometric analysis. The minimal amount of RNase A that was required in order to completely digest the RNA was determined empirically. The appropriate amount of RNase A (Thermo Fisher Scientific, no. EN0531) was then added to the samples, followed by a 1 h incubation at 37 °C. The pH was then adjusted to 8.0 by adding 1 M TrisHCl (pH 8.0 at 25 °C). Disulfide bonds were reduced by adding TCEP (tris(2carboxyethyl)phosphine hydrochloride) to a final concentration of 10 mM and the samples were incubated for 30 min at 65 °C. Free thiol groups were alkylated by adding 2-chloroacetamide (Sigma, no. C0267) to a final concentration of 40 mM and samples were incubated for 20 min at room temperature in the dark. The pH of each sample was adjusted again to 8.0 using 1 M Tris-HCl (pH 8.0 at 25 °C). Trypsin was added to each sample at a trypsin:protein mass ratio of 1:4, where protein mass before crosslinking was considered. The samples were then incubated for 14–18 h at 37 °C, in an orbital shaker at 200 r.p.m. Next, additional trypsin was added, using the same amount as above, and the samples were incubated for additional 2 h at 37 °C in an orbital shaker at 200 r.p.m. The digestion was stopped by the addition of formic acid to pH 3.0. Tryptic peptides were purified using OMIX C18 pipette tips (Agilent Technologies) according to the manufacturer’s instructions. The samples were then dried using a centrifugal vacuum concentrator (Labconco Acid-Resistant CentriVap Concentrator, no. 7810041; CentriVap -105 Cold Trap, no. 7385037; Javac Vector LT-5 High Vacuum Pump, no. VectorLT), and stored at -80 °C until mass spectrometric analysis.

Before mass spectrometric analysis, the peptides were resuspended in 20 µl of 0.1% formic acid, sonicated in a sonicator water bath for 10 min (Grant Instruments XUBA3 Analog Ultrasonic Bath) and centrifuged for 5 min at 21,000g. The samples were then transferred to mass spectrometric vials and analyzed by LC–MS/MS within 48 h.

3.8.7 Tandem mass spectrometry for RBDmap

The peptides were analysed by LC-MS/MS on an Orbitrap Fusion Tribrid Instrument connected to an UltiMate 3000 UHPLC liquid chromatography system (Thermo-Fisher Scientific). Peptides reconstituted in 0.1% formic acid were loaded onto a trap column (Acclaim PepMap100 C18 Nano-Trap, 2 cm × 100 μm i.d., 5 μm particle size and 300 Å pore size; Thermo-Fisher Scientific) at 15 μL/min for 3 min before switching the trap column in-line with the analytical column (Acclaim C18 PepMap RSLC Nanocolumn, 50 cm × 75 μm i.d., 3 μm particle size and 100-Å pore size; Thermo-Fisher Scientific) and the mass spectrometer. The separation of peptides was performed at 250 nL/min using a non-linear gradient over 65 min, starting at 2.5% Buffer B (80% acetonitrile, 0.1% formic acid) and 97.5% Buffer A (0.1% formic acid) and ending at 42% Buffer B. Data were collected in positive mode using data-dependent acquisition. The precursor scan, which covered a range of 375–2000 m/z at a resolution of 120,000, was followed by up to 12 subsequent MS/MS scans measured at a resolution of 60,000. Higher-energy collisional dissociation (HCD) was used to fragment precursor ions with a charge state of 2-7. Other instrument parameters were: (i) injection times of 118 ms for both ms1 and ms2 scans, (ii) AGC (automatic gain control) target of 1×106 and 0.4×106 for ms1 and ms2 scans, respectively, (iii) ion intensity threshold of 1×105, (iv) dynamic exclusion of 30 s and (v) collision energy of 32%.

For data analysis, resulting Thermo Scientific mass spectrometer binary data files (.raw) were analysed using the Andromeda search engine [54] implemented into the MaxQuant software package (version 1.6.0.1) [52]. Only the assayed protein sequences were added to the search database. Acetylation of N-termini of proteins and oxidation of methionines were considered as variable modifications, whilst carbamidomethylation was specified as a fixed modification. False discovery rates of 1% for the peptide spectrum match level were applied by searching a reverse database and peptides that were detected with low confidence (Andromeda score $<$ 20) were eliminated. Tryptic peptides that passed the quality assurance criteria were used to identify the adjacent RNA-linked peptide based on the protease that was used prior to the oligo $d(T)_{25}$ magnetic bead purification and trypsin digestion (either LysC or ArgC, see above) combined with the location of arginine and lysine residues that are flanking the sequence of the identified tryptic peptide, as previously described [42]. Tryptic peptides were considered for unambiguous identification of their adjacent RNA-linked peptide only if their C-terminal residue (either lysine or arginine) was different than the C-terminal residue of their N-terminally adjacent tryptic peptide. R scripts used for data analysis downstream of Andromeda can be obtained at https://github.com/egmg726/crisscrosslinker.

PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.) was used to visualise the RNA-linked peptides in the structure of PRC2-AEBP2 using PDB accessions 6C23 [5] and 5WAI [6], as described for targeted RBR-ID data visualisation above. The JARID2 fragment in the non-catalytic lobe was not shown since JARID2 protein was not used in the RBDmap experiment.

3.8.8 Bis(sulfosuccinimidyl)suberate crosslinking mass spectrometry

0.5 to 1.0 µM PRC2 complex with its accessory subunits was crosslinked with 5–20 µg BS3 (Thermo Fisher Scientific, no. 21585) in 20 mM HEPES pH 7.5 (at 25°C), 150 mM NaCl, 1 mM TCEP, in a total reaction volume of 85 µl. Reactions were allowed to proceed for 30 min at 25 °C before Tris-HCl pH 8.0 (at 25 °C) was added to a final concentration of 30 mM, to quench the reaction, for 15 min at 25°C. Crosslinking efficiency was assessed by subjecting a portion of the product to 10% SDS–PAGE. The crosslinked product solutions were adjusted to pH 8.0 by adding 20 µl of 1 M Tris-HCl pH 8.0 (at 25 °C), reduced in 10 mM TCEP for 30 min at 60 °C and then alkylated in 40 mM 2-chloroacetamide for 20 min at 25 °C in the dark. Mixtures were then digested with 1:100 trypsin:protein mass ratio overnight at 37°C with shaking on a Thermomixer (Eppendorf). Protein digestions were stopped by acidification through adding formic acid to 1% (v/v). Digested peptides were purified using OMIX C18 Mini-Bed pipette tips according to the manufacturer's instructions, dried in a vacuum concentrator and reconstituted into 20 µl of 0.1% formic acid before mass spectrometric analysis.

The peptides were analyzed by LC–MS/MS on an Orbitrap Fusion Tribrid Instrument connected to an UltiMate 3000 UHPLC liquid chromatography system (Thermo Fisher Scientific) as described above for RBDmap.

pLink and pLink2 [55] were used to identify BS3-crosslinked peptides with a false discovery rate of 0.05. In addition, the crosslinked peptides were kept for downstream structural analysis only if they had been identified in at least two independent replicates or if they had been identified with a P value of less than $10^{-}4$. Two-dimensional representations of crosslinked peptides and proteins were generated using xiNET [56]. R scripts used for data retention and analysis downstream of pLink can be downloaded from GitHub: https://github.com/egmg726/crisscrosslinker.

3.8.9 Nucleosome reconstitution

Mononucleosomes were reconstituted as previously described [57]. In brief, recombinant histones were purified from inclusion bodies and reconstituted into histone octamers. Histone octamers and 182-base-pair '601' DNA were combined in a buffer containing 20 mM Tris-HCl pH 7.5 (at 25 °C), 2M KCl, 1mM EDTA, 1mM DTT and were dialyzed for 55h at 4–8°C against a buffer consisting of 20 mM Tris-HCl pH 7.5 (at 25 °C), 25 mM KCl, 1 mM EDTA and 1 mM DTT with a salt concentration that was gradually reduced during the process. The quality of reconstituted nucleosomes was assessed by a 4% TBE (trisborate EDTA) gel and negative stain EM.

3.8.10 Negative stain EM

Negative stain EM experiments were carried out for assessing mononucleosome samples. Two microliters of 0.03 mg ml–1 mononucleosome solution were applied to a glow-discharged continuous carbon grid (EMgrid Australia) and incubated for 30 s before blotting with filter paper and immediately staining by successive incubations, two times for 15 s and one time for 30 s, in drops of 2% (w/v) uranyl acetate. The stain was removed by blotting with filter paper and the grid was air-dried before imaging. The grid was imaged at a magnification of ×52,000 using an FEI Tecnai 12 transmission electron microscope operating at 120 keV.

3.8.11 HMTase activity assays

Unless indicated otherwise, the HMTase activity assays were performed as previously described [15] with minor modifications. Briefly, RNAs were folded as described for binding assays above and were then allowed to bind PRC2 in RNA-binding buffer (50 mM Tris-HCl pH 7.5 at 25 °C, 100 mM KCl, 2.5 mM MgCl2, 0.1 mM ZnCl2 and 2 mM 2-mercaptoethanol) at 25 °C for 30 min. Each 10 µl HMTase reaction contained 500 nM PRC2 complex, RNA as indicated throughout the text, 4 µM H3.1 histone protein (NEB M2503S) or 2 µM mononucleosomes (see ‘Nucleosome reconstitution’ above for *in vitro* nucleosome reconstitution), and 5.0 µM S-[methyl-^{1}4C]-adenosyl-l-methionine (PerkinElmer, no. NEC363050UC). The reactions were incubated for 1 h at 30 °C in HMTase buffer (77.5 mM Tris-HCl pH 8.0 at 30 °C, 155 mM KCl, 3.88 mM MgCl2, 0.2 mM ZnCl2, 3.1 mM 2-mercaptoethanol, 0.1 mg ml–1 BSA (NEB B9000) and 5% v/v glycerol). For HMTase assays in the presence or absence of stimulatory peptides, reaction buffer consisted of 50 mM Tris-HCl (pH 8.0 at 30 °C), 100 mM KCl, 2.5 mM $MgCl_2$, 0.1 mM $ZnCl_2$, 2 mM 2-mercaptoethanol, 0.1 mg ml^{-1} bovine serum albumin and 5% v/v glycerol. The reactions were then stopped by adding 4× LDS sample buffer (Thermo Fisher Scientific, no. NP0007) to a final concentration of 1×. Samples were then incubated at 95 °C for 5 min and the samples were subjected to 16.5% SDS–PAGE. Gels were first stained with InstantBlue Coomassie protein stain (Expedeon, no. ISB1L) and then vacuum-dried for 60 min at 80 °C with the aid of a VE-11 electric aspirator pump (Jeio Tech). Dried gels were exposed to a storage phosphor screen (GE Healthcare) for 1–8 days and the signal was acquired using a Typhoon Trio imager (GE Healthcare). Densitometry was carried out using ImageJ [58]. Relative HMTase activities of mutant PRC2 complexes were obtained through a normalization to the signal obtained from the corresponding wild-type PRC2 complex that was loaded in a different lane on the same gel. All experiments were performed in triplicate.

For assaying RNA inhibition of non-histone substrates, experiments were performed as above, with the exception that the final reaction buffer was 25 mM HEPES pH 8.0 (at 25 °C), 40 mM KCl, 50 mM NaCl, 3 mM $MgCl_2$, 2 mM 2-mercaptoethanol and 10% v/v glycerol.

3.8.12 Structures used for visualization

The high-resolution cryo-EM structure of the PRC2–AEBP2–JARID2 complex (PDB accession 6C23) [5] and the crystal structure of SUZ12–RBBP4–JARID2–AEBP2 (PDB accession 5WAI) [6] were used for PRC2 data representation as described above. The coordinates of the EED-bound H3K27me3, H3K27M and A395 are from PDB accessions 3IIW [8], 5HYN [10] and 5K0M [43], respectively. The coordinates of MTF2 were obtained from PDB accession 5XFR [16].

3.8.13 Cell lines

RBR-ID experiments were carried out in E14Tg2A.4 (E14) mESCs that were obtained from the Reinberg Laboratory, as previously used by Kaneko et al. [41]. When cultured in the described conditions, these E14 cells display the typical morphology of mESCs. In addition, we have accumulated extensive genomic and functional data that confirm their pluripotent state and identity. All cell lines used in the Bonasio Laboratory are routinely tested for mycoplasma. The E14 cells used in these experiments tested negative as recently as October 2017.

3.8.14 Statistics and reproducibility

All P values were calculated using paired or unpaired two-tailed Student's t-test, as explicitly indicated in the respective figure legends. P values were given when $P < 0.05$. All binding curves and bar plots represent means averaged across the number of samples indicated in the respective figure legend.

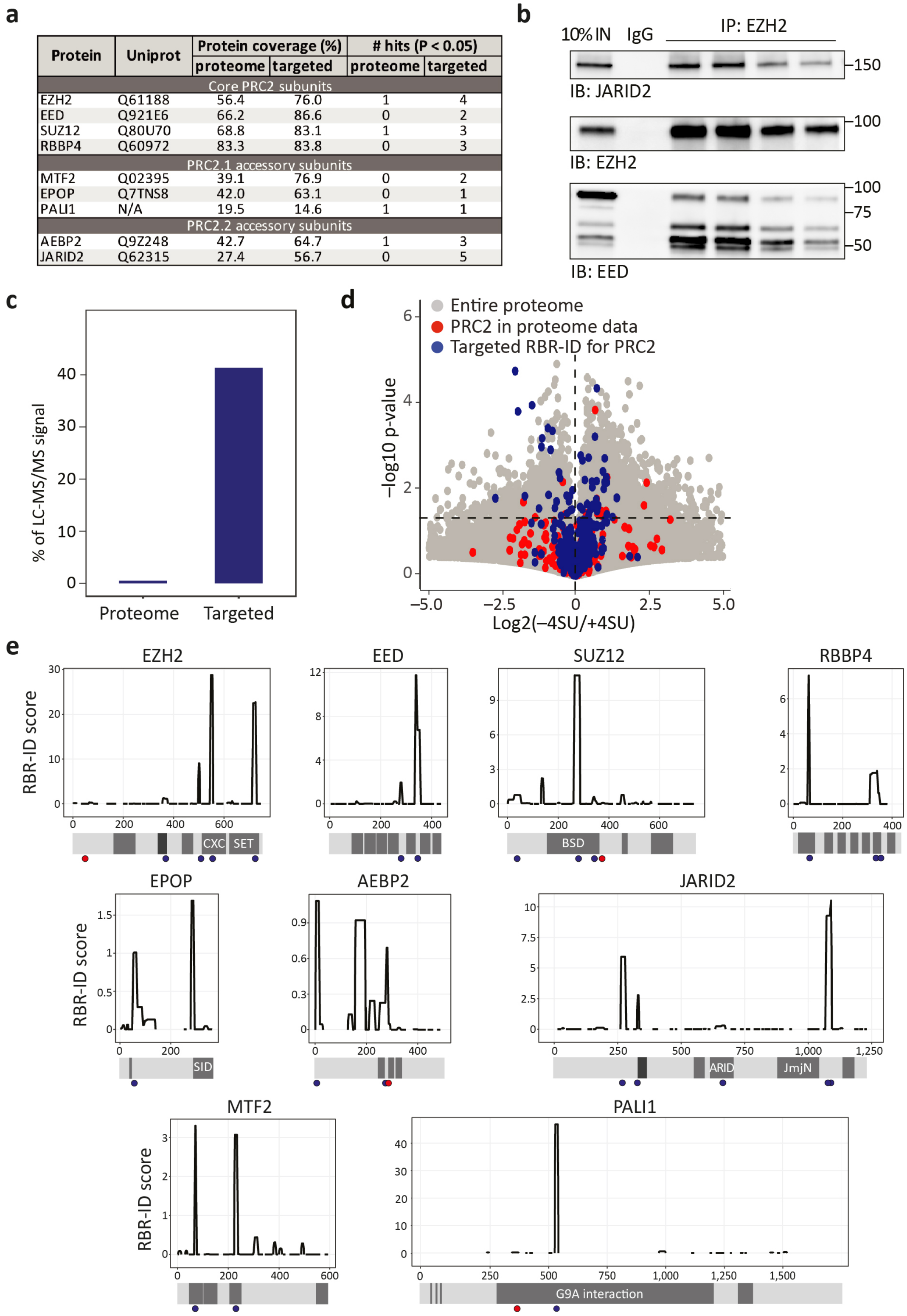
3.9 Code availability

R scripts used for XL–MS and RBDmap data analysis downstream of pLink and Andromeda, respectively, can be downloaded from GitHub: https://github.com/ egmg726/crisscrosslinker.

3.9.1 Data availability

LC–MS raw data for targeted RBR-ID experiments have been deposited at the Chorus project (https://chorusproject.org) with ID 1560. Mass spectrometry data for RBDmap and BS3 XL–MS experiments were deposited at FigShare with DOIs https://doi.org/10.26180/5c3d9751c64ae and https://doi.org/10.26180/5c3d8dd45651b, respectively. Source data for Figures 3.2 - 3.5 and Figures 3.9 - 3.12 are available within Digital Supplemental Table S13-S16, respectively.

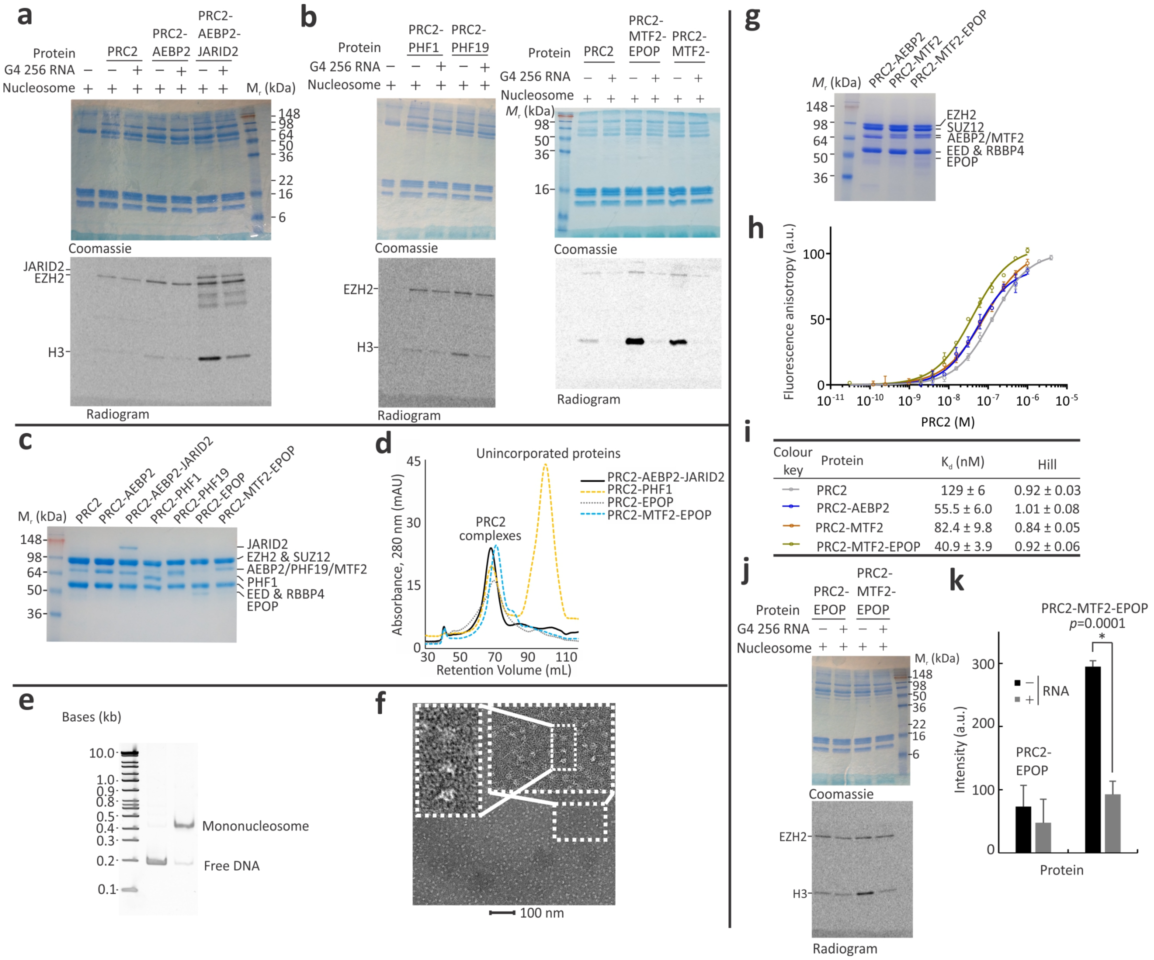
3.10 Supplementary Data



Protein	Uniprot	Protein coverage (%)		# hits (P < 0.05)	
		proteome	targeted	proteome	targeted
Core PRC2 subunits					
EZH2	Q61188	56.4	76.0	1	4
EED	Q921E6	66.2	86.6	0	2
SUZ12	Q80U70	68.8	83.1	1	3
RBBP4	Q60972	83.3	83.8	0	3
PRC2.1 accessory subunits					
MTF2	Q02395	39.1	76.9	0	2
EPOP	Q7TNS8	42.0	63.1	0	1
PALI1	N/A	19.5	14.6	1	1
PRC2.2 accessory subunits					
AEBP2	Q9Z248	42.7	64.7	1	3
JARID2	Q62315	27.4	56.7	0	5

FIGURE 3.8 Additional information about targeted RBR-ID (related to FIGURE 3.1)

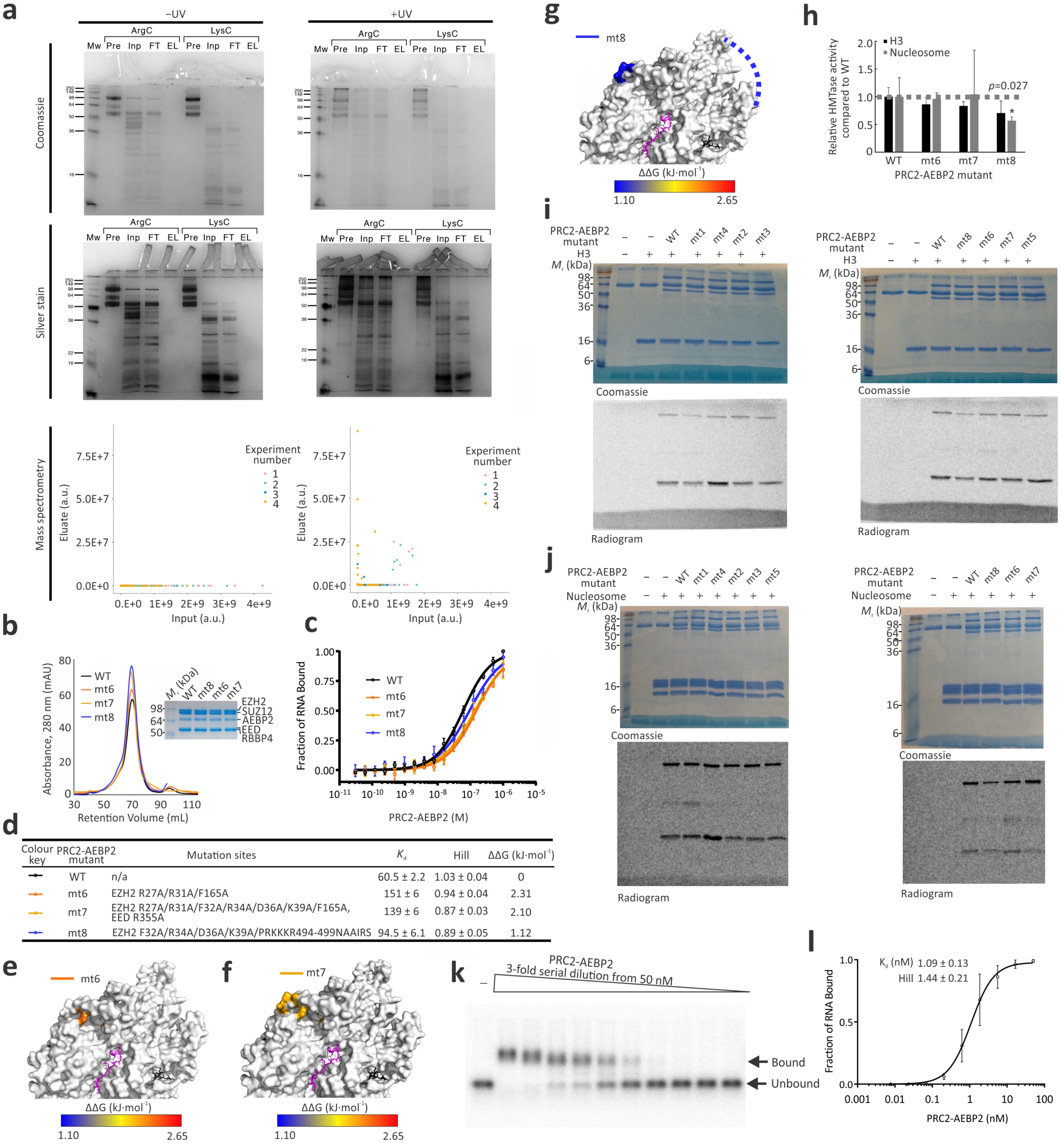
(A) Summary table of PRC2 RBR-ID data. PRC2 proteins, their Uniprot database accession IDs, percent sequence coverage, and the number of significantly ($P < 0.05$) depleted RNA-binding peptides identified by RBR-ID are shown for nuclear proteome data [38] and targeted RBR-ID (this study). Because PALI1 was discovered after our proteome-wide study [45], we reanalysed the dataset separately to include its sequence in the database. (B) Western blot from representative PRC2 immunoprecipitation for core subunits EED and EZH2 and accessory subunit JARID2 (uncropped blots are in Digital Supplemental Table S12). (C) Comparison of the portion of mass spectrometry signal mapping to PRC2 subunits in proteome-wide data1 compared to a representative experiment of SILAC-based targeted RBR-ID. Data are plotted as a percentage of all detected peptides. (D), Volcano plot of peptides in the proteome and targeted RBR-ID data, displaying mean log2-fold changes in ±4SU samples against log-transformed P value, calculated using paired Student's t-test. PRC2 peptides from proteome and targeted RBR-ID are displayed in red and blue respectively. Horizontal dashed line represents P = 0.05. e, RBR-ID score plots (see He, C. et al., Mol Cell. 64, 416-430, 2016)[38] for all PRC2 subunits shown in Figure 3.1 and discussed in this study. Protein domain schematic is shown below each linear plot. Experiments and Data generated by R.W-T. and R.L.



Colour key	Protein	K_d (nM)	Hill
●	PRC2	129 ± 6	0.92 ± 0.03
●	PRC2-AEBP2	55.5 ± 6.0	1.01 ± 0.08
●	PRC2-MTF2	82.4 ± 9.8	0.84 ± 0.05
●	PRC2-MTF2-EPOP	40.9 ± 3.9	0.92 ± 0.06

FIGURE 3.9 PRC2 complex purification and nucleosome reconstitution (related to FIGURE 3.2)

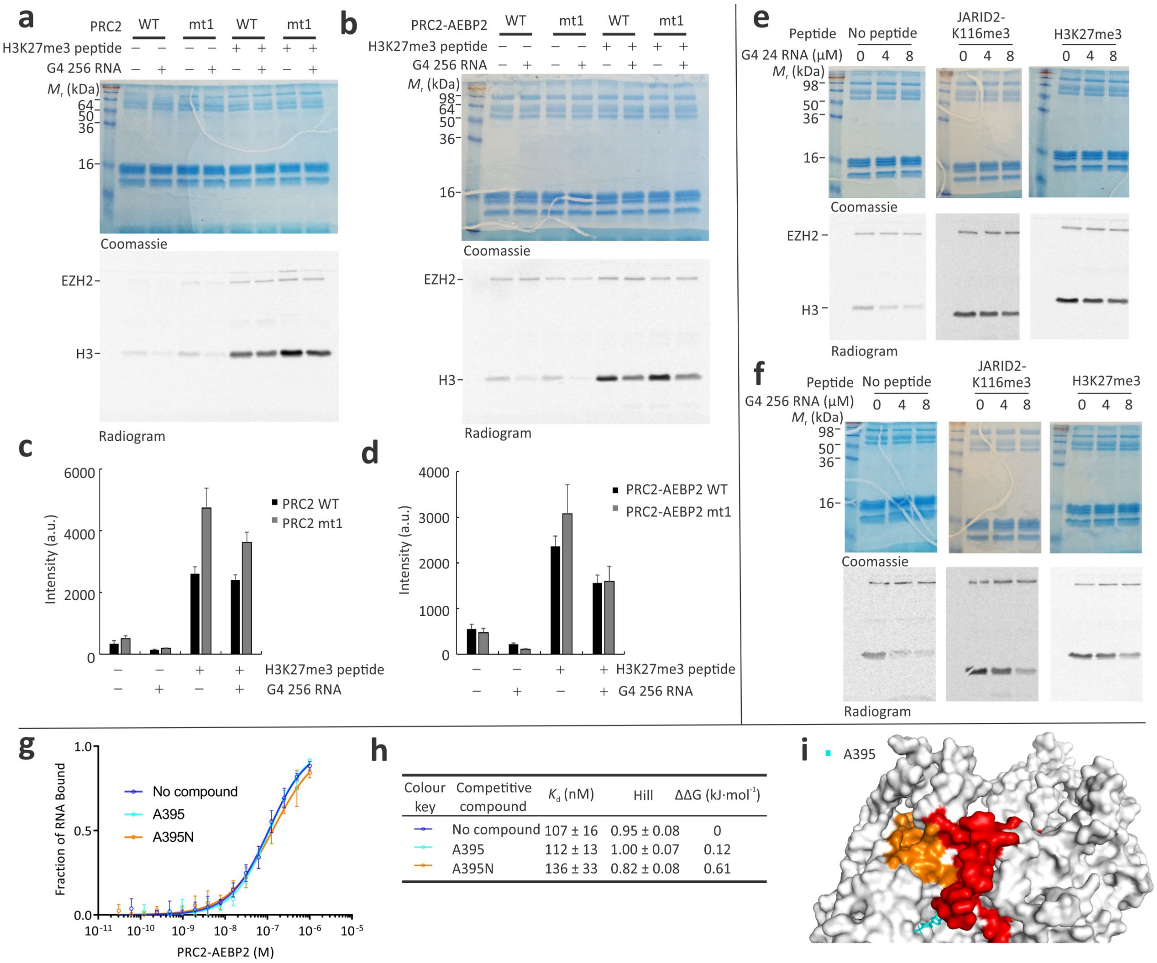
(A)-(B) Full Coomassie blue-stained SDS-PAGE and the corresponding radiogram as shown in Figure 3.2 C,D. (C) Coomassie blue-stained SDS-PAGE gel shows the purity of PRC2 complexes used for HMTase assays in Figure 3.2 C,D. (D) Gel filtration chromatography (Sephacryl S-400 HR resin) of the PRC2 complexes that were used for HMTase assays in Figure 3.2 C,D. Only fractions corresponding to assembled PRC2 complexes were collected and used. (E) Mononucleosomes used for HMTase assays in Figure 3.2 C and 2D were analysed on a 4% polyacrylamide TBE gel and visualised by SYBR Green I post-staining. (F) Mononucleosome homogeneity was assessed using negative stain electron microscopy (representative micrograph at x52,000 magnification). (G) A Coomassie blue-stained SDS-PAGE gel shows the purity of PRC2-MTF2 and PRC2-MTF2-EPOP complexes. (H) Fluorescence anisotropy was carried out to compare the RNA-binding affinities of PRC2-MTF2 and PRC2-MTF2-EPOP. Error bars represent standard deviation based on three independent experiments that were performed on different days. (I) Resulting dissociation constants (Kd) and Hill coefficients are indicated, including the corresponding standard errors. Data for PRC2 was imported from Fig 2B, for a direct comparison. (J)-(K) HMTase assays of the indicated complexes were carried out in the presence or absence of 8.0 μM G4 256 RNA. (J) A representative Coomassie blue-stained SDS-PAGE and the corresponding radiograms. (K) Quantification of HMTase activities from (K), with error bars representing standard deviation calculated from three independent experiments. P values were determined using unpaired two-tailed Student's t-test; *, P<0.05. Data and analysis generated by collaborators.



Colour key	PRC2-AEBP2 mutant	Mutation sites	K_d	Hill	ΔΔG (kJ·mol^{-1})
●	WT	n/a	60.5 ± 2.2	1.03 ± 0.04	0
●	mt6	EZH2 R27A/R31A/F165A	151 ± 6	0.94 ± 0.04	2.31
●	mt7	EZH2 R27A/R31A/F32A/R34A/D36A/K39A/F165A, EED R355A	139 ± 6	0.87 ± 0.03	2.10
●	mt8	EZH2 F32A/R34A/D36A/K39A/PRKKKR494-499NAAIRS	94.5 ± 6.1	0.89 ± 0.05	1.12

FIGURE 3.10 Direct and unbiased detection of protein-RNA interactions within the PRC2-AEBP2 complex (related to FIGURE 3.3)

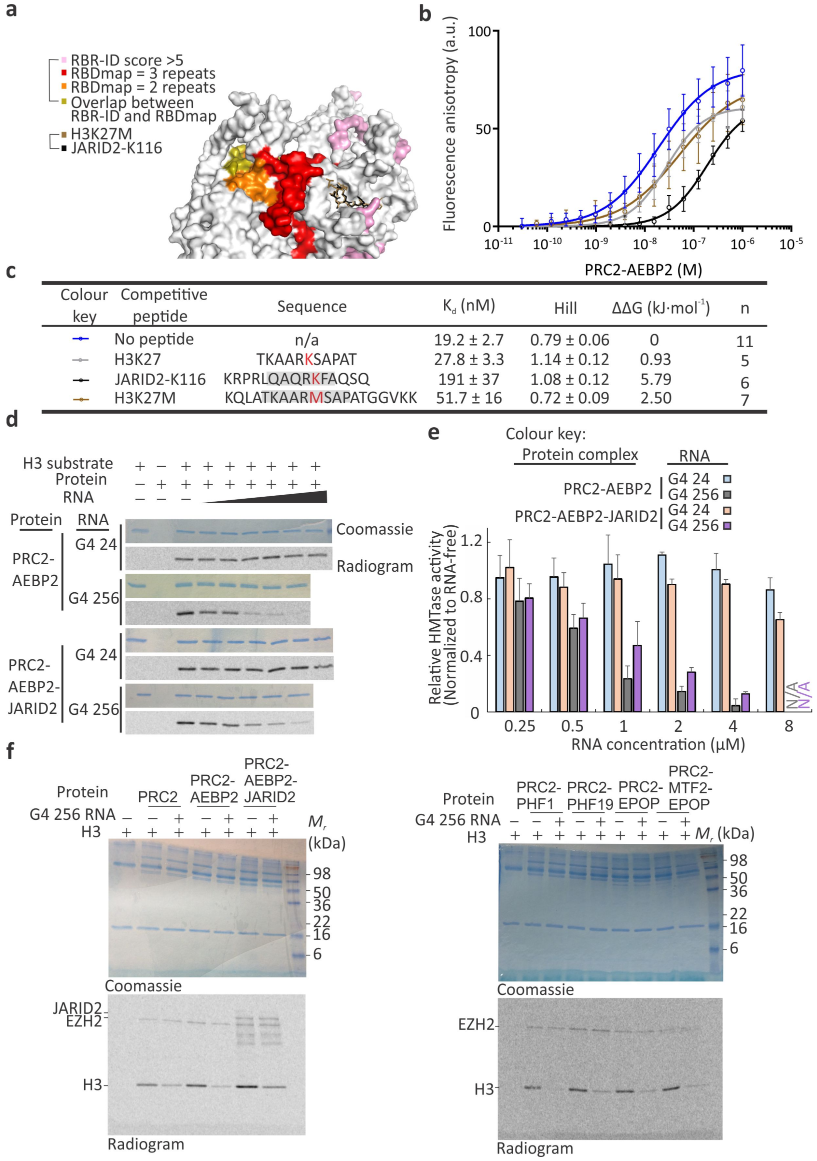
a, Evidence of UV cross-linking, analysed using 18% SDS-PAGE and visualised by Coomassie blue and silver staining. MW: Molecular weight marker; Pre: input before adding LysC or ArgC protease; Inp: input; FT: flow-through; EL: eluate. Scatterplots (bottom) indicate intensities identified by MS/MS for each of the peptides in the input (x-axis) and eluate (y-axis) in four independent RBDmap experiments (in assorted colours). Although the recovered peptides were obtained in quantities below the detection limit of SDS-PAGE (EL lanes in all gels), they were detected by MS/MS only in the +UV sample, indicating the stringency of the purification process. b, PRC2-AEBP2 mutants were evaluated by 10% SDS-PAGE and gel filtration chromatography (Sephacryl S-400 HR resin). c, Fluorescence anisotropy used to quantify the affinity of the mutants to G4 24 RNA. The resulting dissociation constant (Kd), Hill coefficients and the derived $\Delta\Delta G$ are indicated together with details of the mutated amino acids in EZH2 and EED. Error bars in (c) represent standard deviation based on three independent experiments that were performed on different days. Standard errors are indicated in (d) when applicable. e-g, The impaired RNA-binding activity of the mutants and their position on the surface of PRC2 is represented in a G heat map using the PRC2-AEBP2 structure (regulatory and substrate peptides are coloured in magenta and black respectively). h, Bar plot represents the relative HMTase activities of PRC2-AEBP2 mutations toward the H3 substrate compared to the wild-type, which is indicated as a dashed grey line. Error bars indicate standard deviations as measured across three independent experiments. P values were determined using unpaired two-tailed Student's t-test; *, $P<0.05$. I,J, Representative Coomassie blue-stained SDS-PAGE and the corresponding radiograms used for the HMTase assays are in Figure 3.3 G and Figure 3.10 H. K, The affinity of PRC2-AEBP2 to ^{32}P-radiolabeled G4 256 RNA was quantified using EMSA. Data points represent three-fold dilutions of PRC2-AEBP2 starting from 50 nM. l, Quantification was done by fitting the EMSA data to an equilibrium binding curve. Error bars indicate standard deviation based on three independent experiments that were performed on different days. The resulting dissociation constant (Kd), Hill coefficient and standard errors are indicated. Data and analysis generated by collaborators.



Colour key	Competitive compound	K_d (nM)	Hill	ΔΔG (kJ·mol^{-1})
—●—	No compound	107 ± 16	0.95 ± 0.08	0
—●—	A395	112 ± 13	1.00 ± 0.07	0.12
—●—	A395N	136 ± 33	0.82 ± 0.08	0.61

FIGURE 3.11 Stimulatory peptides relieve RNA-mediated inhibition of PRC2 (related to FIGURE 3.4)

a,b, HMTase assays were carried out in the presence of 0.5 µM wild-type (WT) or mutant 1 (mt1) PRC2 or PRC2-AEBP2, 2.0 µM nucleosome substrate, in the presence (+) or absence (-) of 80 µM H3K27me3 peptide and in the presence (+) or absence (-) of 4.0 µM G4 256 RNA. Representative Coomassie blue-stained SDS-PAGE (top) and the corresponding radiograms (middle) are presented, with bar plots (bottom) representing the HMTase activities quantified based on three replicates. The data represented by black bars in panels (c) and (d) were used to generate Figure 3.4 A. e,f, Representative Coomassie blue-stained SDS-PAGE and the corresponding radiograms used for quantifying the HMTase activities presented in Figure 3.4 C. f, HMTase assays were carried out in the presence of 0.5 µM PRC2-AEBP2, 2.0 µM nucleosome substrate and G4 24 RNA (e) or G4 256 RNA (f) at concentrations of either 0, 4 or 8 µM and stimulatory peptides, as indicated. g, Fluorescence anisotropy used to quantify the affinity of PRC2-AEBP2 to G4 24 RNA in the presence or absence of 10 µM of the EED inhibitor A395 or the negative control A395N. Error bars represent standard deviation based on three independent experiments that were performed on different days. h, Resulting dissociation constants (Kd), Hill coefficients and the derived ΔΔG values are indicated. Standard errors are indicated. i, The coordinates of A395, as previously identified by X-ray crystallography (PDB: 5K0M, He, Y. et al., Nat Chem Biol. 13, 389-395, 2017), are presented on the high-resolution cryo-EM structure of PRC2 (PDB: 6C23) by superimposing EED from both structures. Orange and red spots represent RNA-linked polypeptides that were identified in 2 or 3 independent RBDmap experiments respectively. Data and analysis generated by collaborators.



Colour key	Competitive peptide	Sequence	K_d (nM)	Hill	ΔΔG (kJ·mol^{-1})	n
●	No peptide	n/a	19.2 ± 2.7	0.79 ± 0.06	0	11
●	H3K27	TKAARKSAPAT	27.8 ± 3.3	1.14 ± 0.12	0.93	5
●	JARID2-K116	KRPRLQAQRKFAQSQ	191 ± 37	1.08 ± 0.12	5.79	6
●	H3K27M	KQLATKAARMSAPATGGVKK	51.7 ± 16	0.72 ± 0.09	2.50	7

FIGURE 3.12 DNA-independent RNA-mediated inhibition of PRC2 (related to FIGURE 3.5)

a, The location of a substrate peptide (JARID2-K116, in black; PDB 6C23) and an oncogenic inhibitory peptide (H3K27M, in brown; PDB 5HYN) within PRC2 (in grey; PDB 6C23) with respect to RNA-linked polypeptides that were identified using RBR-ID (score of > 5; in pink), RBDmap (identified in 2 or 3 independent experiments; represented in orange and red respectively) or in both assays (in yellow). b, The affinity of PRC2-AEBP2 to G4 24 RNA was quantified using fluorescence anisotropy in the presence or absence of a substrate peptide (10 µM JARID2-K116 or 100 µM H3 histone peptide) or an oncogenic peptide (100 µM H3K27M); see panel c for a colour code. c, Resulting dissociation constants (Kd), Hill coefficients, the derived ΔΔG values and the number of independent replicates (n) are indicated. Peptide sequences are indicated, with the substrate lysines in red; highlighted in grey, are amino acids that were previously traced in the catalytic centre, using high resolution cryo-EM (JARID2-K116; PDB 6C23) or x-ray crystallography (H3K27M; PDB 5HYN). d,e, HMTase assays were carried out in the presence of 0.5 µM PRC2-AEBP2 or PRC2-AEBP2-JARID2, 4.0 µM H3 histone substrate and in the presence or absence of G4 256 RNA or G4 24 RNA at concentrations as indicated under the bar plot. The bar plot represents the relative activity with respect to the no-RNA sample, as recorded by densitometry after SDS-PAGE. See Digital Supplemental Table S16 for the uncropped images of the gels and radiograms. f, Uncropped images of the gels shown in Figure 3.5 A. Data and analysis generated by collaborators.

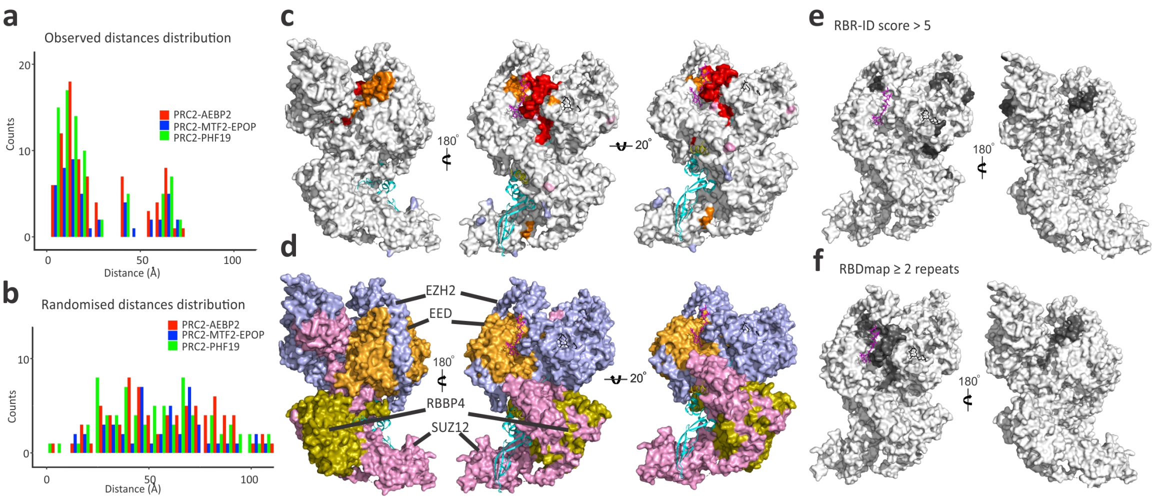


FIGURE 3.13 One face of PRC2 clusters binding sites of multiple regulatory factors (related to FIGURE 3.6)

a, Histogram of distances that were measured between cross-linked lysine pairs within the PRC2 core subunits. BS3 XL-MS data was generated using the three PRC2 complexes as indicated in the colour key (see Figure 3.6 A-c for cross-linking sites) and distances were measured between lysine pairs within the high-resolution cryo-EM structure of the PRC2-AEBP2-JARID2 complex (PDB: 6C23). b, Randomised distances histogram was generated after randomly selecting N lysine pairs and measuring the distances between them over the same structure as in (a), where N is the number of observed cross-linked lysine-pairs in each of the datasets used in (a). c, Front (centre) and, rear (left) views, and 20° rotation with respect to the front view (right), of the PRC2-AEBP2-JARID2 structure presented in Figure 3.6 D, using the same colour code as in Figure 3.6. d, The structure as shown in (c), represented in assorted colours according to the four PRC2 core subunits. AEBP2 and JARID2, as well as the regulatory and substrate peptides, are coloured according to the same colour key as in Figure 3.6 D. e, RNA-linked peptides that were identified using targeted RBR-ID (RBR-ID score > 5) were mapped to the high-resolution structure of PRC2 (PDB: 6C23 and 5WAI). f, RNA-linked peptides that were identified using RBDmap in 2 or 3 replicates are presented on the same structure and views as in (e), for a direct comparison. Data and analysis generated by collaborators.

3.11 References

1. Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* **171,** 34–57 (2017).
2. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469,** 343–9 (2011).
3. Simon, J. A. & Kingston, R. E. Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Mol Cell* **49,** 808–24 (2013).
4. Comet, I., Riising, E. M., Leblanc, B. & Helin, K. Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. *Nat Rev Cancer* **16,** 803–810 (2016).
5. Kasinath, V. *et al.* Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science* **359,** 940–944 (2018).
6. Chen, S., Jiao, L., Shubbar, M., Yang, X. & Liu, X. Unique Structural Platforms of Suz12 Dictate Distinct Classes of PRC2 for Chromatin Binding. *Mol Cell* **69,** 840–852 e5 (2018).
7. Ciferri, C. *et al.* Molecular architecture of human polycomb repressive complex 2. *Elife* **1,** e00005 (2012).
8. Margueron, R. *et al.* Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461,** 762–7 (2009).
9. Jiao, L. & Liu, X. Structural basis of histone H3K27 trimethylation by an active polycomb repressive complex 2. *Science* **350,** aac4383 (2015).
10. Justin, N. *et al.* Structural basis of oncogenic histone H3K27M inhibition of human polycomb repressive complex 2. *Nat Commun* **7,** 11316 (2016).
11. Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. *Mol Cell* **57,** 769–783 (2015).
12. Lee, C. H. *et al.* Allosteric Activation Dictates PRC2 Activity Independent of Its Recruitment to Chromatin. *Mol Cell* **70,** 422–434 e6 (2018).

13. Perino, M. *et al.* MTF2 recruits Polycomb Repressive Complex 2 by helical-shape-selective DNA binding. *Nat Genet* **50,** 1002–1010 (2018).

14. Lee, C. H. *et al.* Distinct Stimulatory Mechanisms Regulate the Catalytic Activity of Polycomb Repressive Complex 2. *Mol Cell* **70,** 435–448 e5 (2018).

15. Wang, X. *et al.* Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. *Nat Struct Mol Biol* **24,** 1028–1038 (2017).

16. Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature* **549,** 287–291 (2017).

17. Choi, J. *et al.* DNA binding by PHF1 prolongs PRC2 residence time on chromatin and thereby promotes H3K27 methylation. *Nat Struct Mol Biol* **24,** 1039–1047 (2017).

18. Beringer, M. *et al.* EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. *Mol Cell* **64,** 645–658 (2016).

19. Zhang, Z. *et al.* PRC2 complexes with JARID2, MTF2, and esPRC2p48 in ES cells to modulate ES cell pluripotency and somatic cell reprograming. *Stem Cells* **29,** 229–240 (2011).

20. Kloet, S. L. *et al.* The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nat Struct Mol Biol* **23,** 682–690 (2016).

21. Hauri, S. *et al.* A High-Density Map for Navigating the Human Polycomb Complexome. *Cell Rep* **17,** 583–595 (2016).

22. Smits, A. H., Jansen, P. W., Poser, I., Hyman, A. A. & Vermeulen, M. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res* **41,** e28 (2013).

23. Brockdorff, N. Noncoding RNA and Polycomb recruitment. *RNA* **19,** 429–42 (2013).

24. Davidovich, C. & Cech, T. R. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA* **21,** 2007–22 (2015).

25. Hekimoglu, B. & Ringrose, L. Non-coding RNAs in polycomb/trithorax regulation. *RNA Biol* **6,** 129–37 (2009).

26. Ringrose, L. Noncoding RNAs in Polycomb and Trithorax Regulation: A Quantitative Perspective. *Annu Rev Genet* **51,** 385–411 (2017).

27. Bonasio, R. & Shiekhattar, R. Regulation of transcription by long noncoding RNAs. *Annu Rev Genet* **48,** 433–55 (2014).

28. Kaneko, S. *et al.* Interactions between JARID2 and Noncoding RNAs Regulate PRC2 Recruitment to Chromatin. *Molecular Cell* **53,** 290–300 (2014).

29. Herzog, V. A. *et al.* A strand-specific switch in noncoding transcription switches the function of a Polycomb/Trithorax response element. *Nat Genet* **46,** 973–981 (2014).

30. Cifuentes-Rojas, C., Hernandez, A. J., Sarma, K. & Lee, J. T. Regulatory Interactions between RNA and Polycomb Repressive Complex 2. *Molecular Cell* **55,** 171–185 (2014).

31. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–6 (2008).

32. Kaneko, S. *et al.* Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. *Genes and Development* **24,** 2615–2620 (2010).

33. Kanhere, A. *et al.* Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell* **38,** 675–88 (2010).

34. Wang, X. *et al.* Targeting of Polycomb Repressive Complex 2 to RNA by Short Repeats of Consecutive Guanines. *Mol Cell* **65,** 1056–1067 e5 (2017).

35. Long, Y. *et al.* Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *eLife* **6** (2017).

36. Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res* **26,** 896–907 (2016).

37. Ardehali, M. B. *et al.* Polycomb Repressive Complex 2 Methylates Elongin A to Regulate Transcription. *Mol Cell* **68,** 872–884 e6 (2017).

38. 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).

39. Davidovich, C., Zheng, L., Goodrich, K. J. & Cech, T. R. Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* **20,** 1250–7 (2013).

40. Montero, J. J. *et al.* TERRA recruitment of polycomb to telomeres is essential for histone trymethylation marks at telomeric heterochromatin. *Nature Communications* **9** (2018).

41. Kaneko, S., Son, J., Shen, S. S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20,** 1258–64 (2013).

42. Castello, A. *et al.* Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell* **63,** 696–710 (2016).

43. He, Y. *et al.* The EED protein-protein interaction inhibitor A-395 inactivates the PRC2 complex. *Nature Chemical Biology* **13,** 389–395 (2017).

44. Ballare, C. *et al.* Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nat Struct Mol Biol* **19,** 1257–65 (2012).

45. Conway, E. *et al.* A Family of Vertebrate-Specific Polycombs Encoded by the LCOR/LCORL Genes Balance PRC2 Subtype Activities. *Mol Cell* **70,** 408–421 e8 (2018).

46. Oksuz, O. *et al.* Capturing the Onset of PRC2-Mediated Repressive Domain Formation. *Mol Cell* **70,** 1149–1162 e5 (2018).

47. Poepsel, S., Kasinath, V. & Nogales, E. Cryo-EM structures of PRC2 simultaneously engaged with two functionally distinct nucleosomes. *Nat Struct Mol Biol* **25,** 154–162 (2018).

48. Lee, C.-H. *et al.* Automethylation of PRC2 fine-tunes its catalytic activity on chromatin. *BioRxivorg* (2018).

49. Wang, X. *et al.* Regulation of histone methylation by automethylation of PRC2. *BioRxivorg* (2018).

50. Davidovich, C. *et al.* Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. *Mol Cell* **57,** 552–8 (2015).

51. Thomson, R. W., He, C., Sidoli, S., Garcia, B. A. & Bonasio, R. Sample preparation for mass spectrometry-based identification of RNA-binding regions. *Journal of Visualized Experiments* **2017** (2017).

52. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26,** 1367–72 (2008).

53. Davidovich, C., Goodrich, K. J., Gooding, A. R. & Cech, T. R. NAR Breakthrough Article: A dimeric state for PRC2. *Nucleic Acids Research* **42,** 9236–9248 (2014).

54. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10,** 1794–805 (2011).

55. Yang, B. *et al.* Identification of cross-linked peptides from complex samples. *Nat Methods* **9,** 904–6 (2012).

56. Combe, C. W., Fischer, L. & Rappsilber, J. xiNET: cross-link network maps with residue resolution. *Mol Cell Proteomics* **14,** 1137–47 (2015).

57. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol* **304,** 3–19 (1999).

58. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9,** 671–5 (2012).

4. COV-ID: A LAMP SEQUENCING APPROACH FOR HIGH-THROUGHPUT CO-DETECTION OF SARS-COV-2 AND INFLUENZA VIRUS IN HUMAN SALIVA

This chapter is adapted from a submitted manuscript:

Robert Warneford-Thomson, Parisha P. Shah, Patrick Lundgren, Jonathan Lerner, Benjamin S. Abella, Kenneth S. Zaret, Jonathan Schug, Rajan Jain, Christoph A. Thaiss, and Roberto Bonasio.* COV-ID: A LAMP sequencing approach for high-throughput co-detection of SARS-CoV-2 and influenza virus in human saliva. medRxiv. 2021 [1].
*corresponding author

4.1 Abstract

The COVID-19 pandemic has created an urgent need for rapid, effective, and low-cost SARS-CoV-2 diagnostic testing. Here, we describe COV-ID, an approach that combines RT-LAMP with deep sequencing to detect as few as 5–10 virions of SARS-CoV-2 in unprocessed human saliva. Based on a multi-dimensional barcoding strategy, COV-ID can be used to test thousands of samples overnight in a single sequencing run with limited labor and laboratory equipment. The sequencing-based readout allows COV-ID to detect multiple amplicons simultaneously, including key controls such as host transcripts and artificial spike-ins, as well as multiple pathogens. Here we demonstrate this flexibility by simultaneous detection of 4 amplicons in contrived saliva samples: SARS-CoV-2, influenza A, human *STATHERIN*, and an artificial SARS spike-in. The approach was validated on clinical saliva samples, where it showed 100% agreement with RT-qPCR. COV-ID can also be performed directly on saliva adsorbed on filter paper, simplifying collection logistics and sample handling.

4.2 Introduction

Within the first year of the COVID-19 pandemic SARS-CoV-2 has swept across the world, leading to more than 70 million infections and over 1.5 million deaths worldwide (as of December 2020). In many countries, non-pharmaceutical interventions, such as school closures and national lockdowns, have proven to be effective, but could not be sustained due to economic and social impact [2, 3]. Regularly performed population-level diagnostic testing is an attractive solution [4], particularly as asymptomatic individuals are implicated in rapid disease transmission, with a strong overdispersion in secondary transmission [5]. Maintenance of population-level testing can be successful in isolating asymptomatic individuals and preventing sustained transmission [6, 7]; however, considerable barriers exist to the adoption of such massive testing strategies. Two such barriers are cost and supply constraints for commercial testing reagents, both of which make it impractical to test large numbers of asymptomatic individuals on a recurrent basis. A third major barrier is the lack of "user-friendly" protocols that can be rapidly adopted by public and private organizations to establish high-throughput surveillance screening. In addition, while COVID-19 testing of symptomatic individuals might be effective during the summer season, when other respiratory infections are rare, new strategies are urgently needed to facilitate rapid differential diagnosis between SARS-CoV-2 and other respiratory viruses in winter.

Recent adaptations of reverse transcription and polymerase chain reaction (RT-PCR) to amplify viral sequence and perform next-generation DNA sequencing have opened promising new avenues for massively parallel SARS-CoV-2 detection. In general, sequencing-based protocols use libraries of amplification primers to tag reads originating from each individual patient sample with a unique index that can be identified and deconvoluted after sequencing, thus allowing pooling of tens of thousands of samples in a single assay. SARSeq, SPAR-Seq, and Swab-seq, directly amplify the viral RNA by RT-PCR and simultaneously introduce barcodes [8–10]. While effective, these methods rely on individual PCR amplification of each patient sample, thus requiring a large number of thermal cyclers for massive scale-up. An alternative approach, ApharSeq, addresses this bottleneck by annealing barcoded RT primers to viral RNA and pooling samples prior to amplification but the need for specialized oligo-dT magnetic beads might constitute a separate adoption barrier for this method [11]. Finally, several recent methods have been designed to take advantage of the extreme sensitivity and isothermal conditions of loop-mediated isothermal amplification (LAMP)[12–14], but these methods either require additional manipulation to introduce barcodes [12, 13] or do not allow for convenient multiplexing [15]. In this study, we present COV-ID, a method for SARS-CoV-2 identification based on LAMP, which enables large-scale diagnostic testing at low cost and with minimal on-site equipment. COV-ID is a robust method that can be used to test tens of thousands of samples for multiple pathogens with modest reagent costs and 2–4 laboratory personnel, generating results within 24 hours. COV-ID uses unpurified saliva or saliva adsorbed on filter paper as input material, thus enabling the massively parallel, inexpensive testing required for population-level surveillance of the COVID-19 pandemic (Figure 4.1 A).

4.3 Results

4.3.1 Two-step amplification and indexing of viral and human sequences via RT-LAMP and PCR

The molecular basis for COV-ID is reverse transcription loop-mediated isothermal amplification (RT-LAMP), an alternative to PCR that has been used extensively for viral DNA or RNA detection in clinical samples [15–18], including SARS-CoV-2 [19, 20]. RT-LAMP requires 4–6 primers that recognize different regions of the target sequence [21, 22] and proceeds through a set of primed and self-primed steps to yield many inverted copies of the target sequence spanning a range of molecular sizes (Figure 4.5). The forward inner primer (FIP) and backward inner primer (BIP), which recognize internal sequences, are incorporated in opposite orientation across the target sequence in the final amplified product (Figure 4.5). Previous studies have shown that the FIP and BIP tolerate insertion of exogenous sequence between their different target homology regions [23]. We exploited this flexibility and introduced 1) patient-specific barcodes as shown previously [12, 14, 23] and 2) artificial sequences that allowed for PCR amplification of a small product compatible with Illumina sequencing library construction (Figure 4.1 , Figure 4.5). These innovations allowed us to pool individually barcoded RT-LAMP reactions and amplify them in batch via PCR, while introducing unique P5 and P7 dual indexes in different pools, thus enabling two-dimensional barcoding and dramatically increasing method throughput (see Table 4.1 for PCR primer sequences). To minimize pool variability, we titrated PCR primers to 100 nM and performed pool PCRs to completion, resulting in each pool being amplified to the same approximate concentration. Uniquely amplified and barcoded pools were mixed into a single "super-pool" and sequenced on an Illumina NextSeq or similar instrument (Figure 4.1 A). Combining individual barcodes embedded in the product at the RT-LAMP step with dual indexes introduced at the pool level during the PCR step allows for deconvolution of thousands or tens of thousands of samples in a single sequencing run.

To determine whether introduction of these exogenous sequences inhibited the isothermal amplification step, we performed RT-LAMP on inactivated SARS-CoV-2 virus using an extensively validated primer set against the N2 region of the nucleocapsid protein [24] including either the conventional BIP and FIP primers or their modified version re-engineered for the COV-ID workflow (Figure 4.1 B). Although the appearance of the amplified viral product was slightly delayed when using COV-ID primers, all reactions reached saturation rapidly and without detectable amplification of negative controls (Figure 4.1 C). Next, we tested whether COV-ID is compatible with RT-LAMP using newly designed primers against a host (human) transcript and whether the second step of COV-ID, direct library construction and indexing via PCR amplification (Figure 4.1 D), yields the desired product. For this, we designed RT-LAMP primers against the human beta-actin (ACTB) transcript including sequences necessary for COV-ID (Figure 4.1 B, S1). After RT-LAMP, reactions were diluted 100-fold before PCR with barcoded Illumina adapters. A PCR product of the expected size was visible in reactions containing total HeLa RNA, while no PCR product was observed in the absence of template (Figure 4.1 E). Sanger sequencing of the PCR product confirmed that RT-LAMP followed by PCR generated the product expected by the COV-ID method design, including the sample barcode introduced during the RT-LAMP step (data not shown).

Thus, our data show that RT-LAMP is tolerant of sequence insertions in the BIP and FIP primers that allow introduction of LAMP-level barcodes as well as sequences homologous to Illumina adapters for direct amplification, indexing, and library construction via PCR.

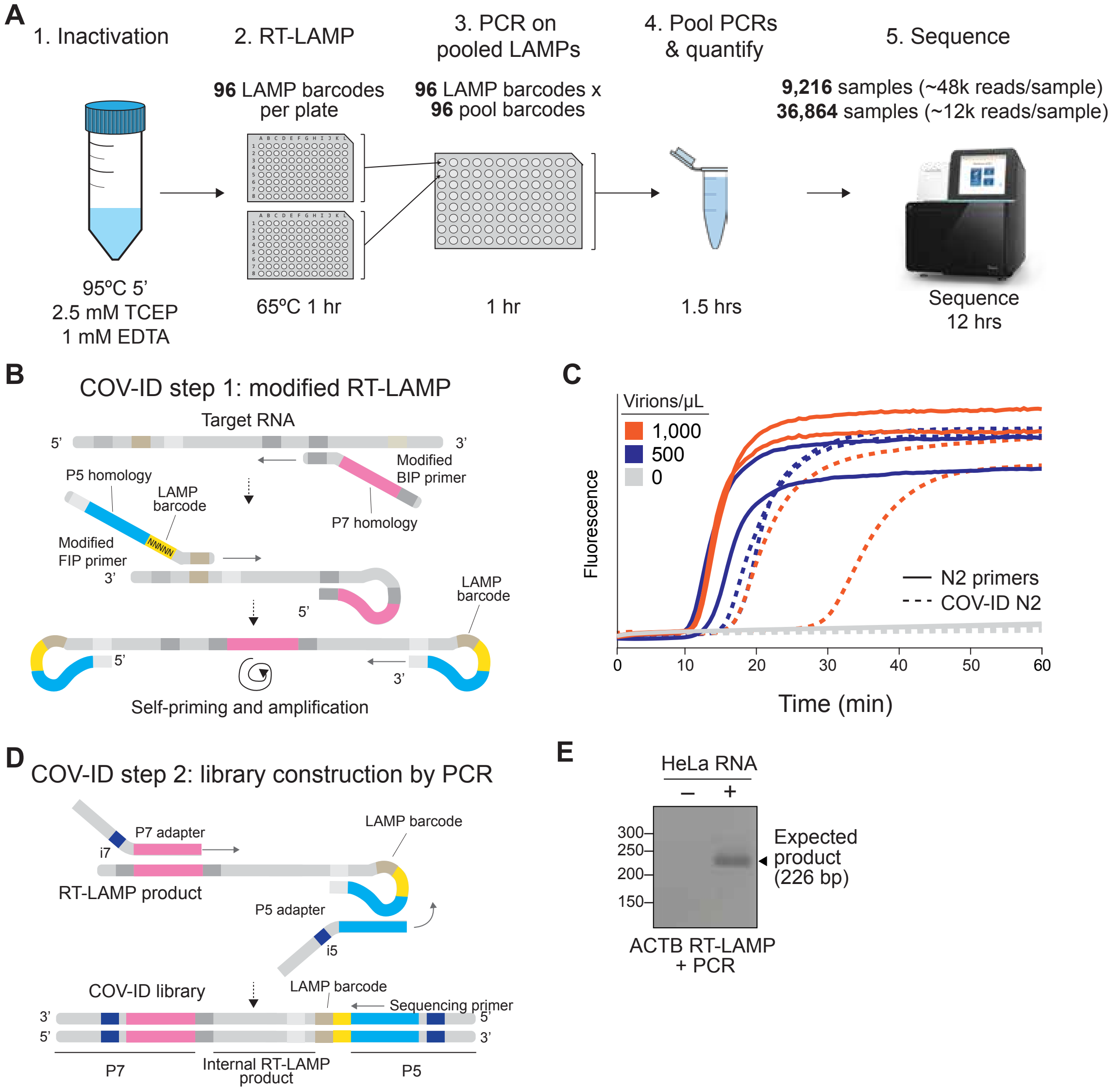


FIGURE 4.1 Barcoding and PCR amplification of RT-LAMP products

(A) Overview of COV-ID. Saliva is collected and inactivated prior to RT-LAMP performed with up to 96 individual sample barcoded primers. LAMP reactions are pooled and further amplified via PCR to introduce Illumina adapter sequences and pool-level dual indexes. A single thermal cycler can amplify 96 or 384 such pools and the resulting "super-pool" can be sequenced overnight to detect multiple amplicons from 9,216 or 36,864 individual patient samples (number of reads in parenthesis assume an output of 450M reads from a NextSeq 500). (B) Schematic of the RT-LAMP (step I) of COV-ID. Selected steps of RT-LAMP reaction are shown to illustrate how the LAMP barcode, shown in yellow, and the P7 and P7 homology sequences (blue and pink, respectively) are introduced in the final LAMP product. A more detailed version of the LAMP phase of COV-ID, including specific sequences, is illustrated in Figure 4.5. (C) Conventional RT-LAMP primers (solid lines) or primers modified for COV-ID (dotted lines) were used for RT-LAMP of SARS-CoV-2. The numbers of inactivated SARS-CoV-2 virions per µL is indicated in the color legend. (D) Schematic of the PCR (step II) of COV-ID. Following RT-LAMP, up to 96 reactions are pooled and purified and Illumina libraries are generated directly by PCR with dual-indexed P5 and P7 adapters in preparation for sequencing. (E) COV-ID primers targeting ACTB mRNA were used for RT-LAMP with HeLa total RNA. LAMP was diluted 1:100, amplified via PCR and resolved on 2% agarose gel. Experiments and Data generated by R.W-T.

4.3.2 Sequencing-based detection of SARS-CoV-2 RNA from saliva using COV-ID

We next evaluated the utility of COV-ID to detect viral RNA in saliva. Unpurified saliva has been shown to be a viable template for nucleic acid amplification via RT-PCR [25], recombinase polymerase amplification (RPA)[26] as well as RT-LAMP [27, 28]. We prepared human saliva for RT-LAMP using a recently described treatment that inactivates SARS-CoV-2 virions, saliva-borne RNases and LAMP inhibitors (Figure 4.2 A) [28]. We performed RT-LAMP followed by PCR (Figure 4.1) on inactivated saliva spiked with water or 1,000 total copies of inactivated SARS-CoV-2 virus. We observed a single band of the expected size in reactions performed on saliva spiked with virus but not in control reactions (Figure 4.2 B). The sequence of the amplified and barcoded viral product was confirmed by Sanger sequencing (Figure 4.6 B). Next, we subjected the libraries to deep sequencing. Reads aligned uniformly to the N gene, the region targeted by the N2 primer set, in COV-ID libraries constructed from viral samples but not in control libraries (Figure 4.2 C).

In several SARS-CoV-2 FDA approved tests, parallel amplification of a host (human) amplicon is implemented as a metric for sample integrity and quality. That is, if no human RNA is amplified from a clinical sample, no conclusion can be drawn from a negative test result [29]. However, in most tests, viral and human amplicons must be detected separately, resulting in a multiplication of the number of reactions to be performed. We reasoned that the deep sequencing nature of COV-ID would allow for simultaneous detection of viral, human, and other control amplicons, without increasing the number of necessary reactions. In fact, given that the PCR handles inserted in the BIP and FIP are the same for all RT-LAMP amplicons (Figure 4.1 D), the same P5 and P7 Illumina primers allow the simultaneous amplification of all RT-LAMP products obtained with COV-ID-modified primer sets. To identify a suitable human control, we compared conventional RT-LAMP primers for the mRNA of ACTB [24] or STATHERIN (STATH), a gene expressed specifically in saliva [30]. To determine which of the two RT-LAMP primer sets was a better proxy to measure RNA integrity in saliva samples, we assayed for amplification of the respective products in presence or absence of RNase. Whereas addition of RNase A abolished the STATH signal, it was ineffectual for ACTB (Figure 4.6 A), suggesting that amplification of genomic DNA made considerable contributions to the RT-LAMP signal observed for the latter. Therefore, we utilized STATH mRNA as a human control in subsequent experiments.

We used COV-ID-adapted primer sets for N2 and STATH (Table 4.1) in multiplex on inactivated saliva spiked with a range of SARS-CoV-2 from 5 to 10,000 virions/µL. Subsequently, each RT-LAMP reaction was separately amplified via PCR using a unique P5 and P7 index combination, pooled, quantified, and deep-sequenced to an average depth of 6,000 reads per sample. After read trimming, alignment, and filtering (see Methods), 76% of reads from saliva COV-ID reactions were informative (Figure 4.6 C). In order to differentiate SARS-CoV-2 positive and negative samples, we calculated the ratio between N2 reads and reads mapping to the human STATH control. Using the highest N2/STATH read ratio in control (SARS-CoV-2 negative saliva) as a threshold, 95% (19/20) of samples with spiked-in virus were correctly classified as positives (Figure 4.2 D). Using COV-ID, we consistently detected SARS-CoV-2 in saliva samples containing as low as 5 virions per µL, a sensitivity comparable and in some cases superior to those of established testing protocols [31].

Scaling COV-ID to handle higher sample numbers requires pooling samples immediately following RT-LAMP prior to the PCR step (Figure 4.1 A). We designed 32 unique 5-nucleotide barcodes for several target LAMP amplicons (Figure 4.6 D and Table 4.2). We first individually validated each barcode and primer combination by real-time fluorescence and PCR efficiency. Certain barcodes inhibited the RT-LAMP reaction, possibly due to internal micro-homology and increased primer self-hybridization [32]. Nonetheless, out of 32 barcodes tested in 3 separate RT-LAMP reactions (N2, ACTB, and STATH), 25 successfully amplified all three target RNAs (Figure 4.6 D). Saliva samples spiked with various concentrations of inactivated SARS-CoV-2 were amplified via barcoded RT-LAMP, then optionally pooled prior to PCR and sequencing (Figure 4.6 E). CoV-2/STATH ratios demonstrated no loss of sensitivity or specificity in the pooled samples compared to the individual PCRs. To test the potential of COV-ID on patient samples, we tested 10 saliva specimens, collected and previously analyzed at the Hospital of the University of Pennsylvania (see Methods). We carried out multiplex barcoded RT-LAMPs on each sample (COV-ID step I, Figure 4.1 B), pooled the reactions and then constructed libraries via PCR (COV-ID step II, Figure 4.1 D). After deep sequencing, analysis of N2/STATH ratios showed 100% (10/10) concordance with viral copy numbers generated by a standard clinical test (RNA purification followed by RT-qPCR) (Figure 4.2 E), demonstrating the effectiveness of the COV-ID approach. Taken together, our data show that COV-ID can be utilized to detect viral and human amplicons in multiplex directly from saliva samples that can be batch amplified and deconvoluted after deep sequencing.

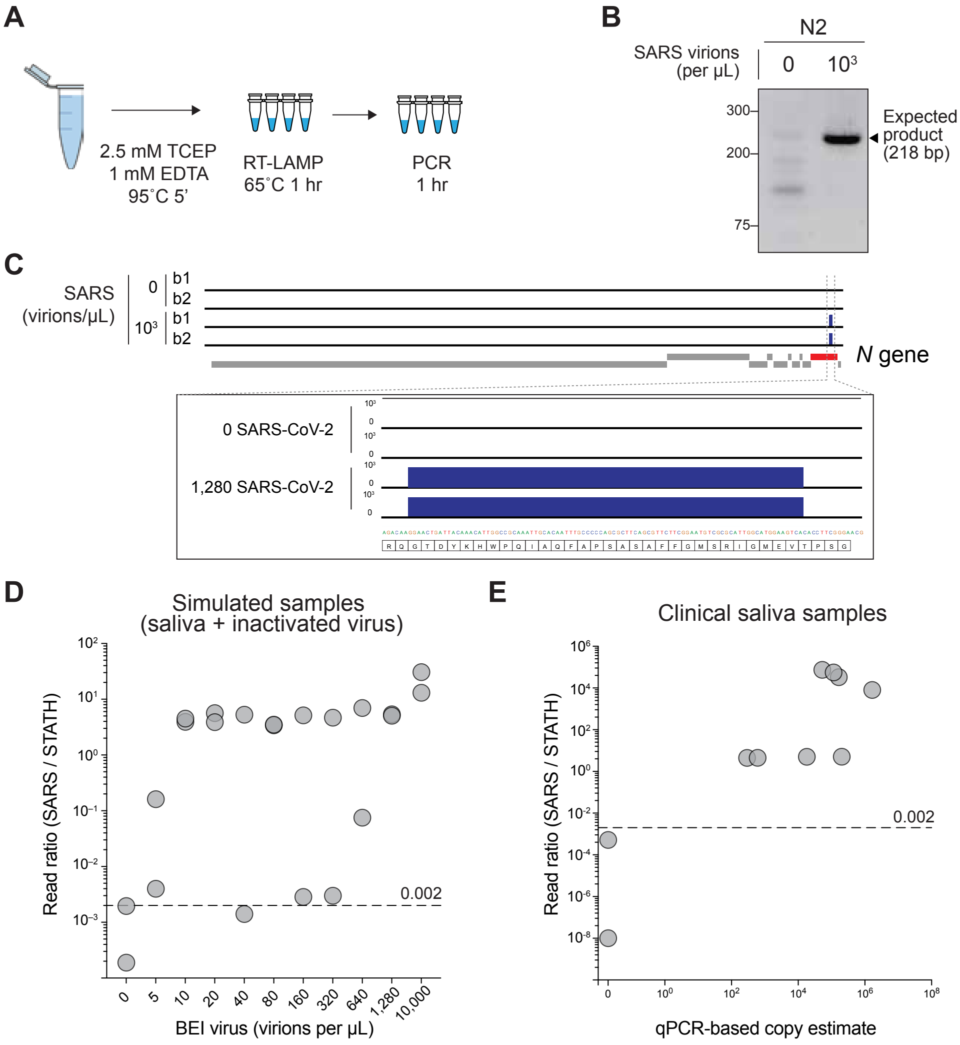


FIGURE 4.2 Sequencing-based detection of SARS-CoV-2 in saliva samples

(A) Saliva preparation. Crude saliva was inactivated via TCEP/EDTA addition and 95ºC incubation prior to RT-LAMP. (B) RT-LAMP followed by COV-ID PCR performed directly on saliva. Saliva with and without addition of 1,000 copies of inactivated SARS-COV-2 templates was inactivated as described in (A), then used as template. (C) Alignment of sequenced reads against SARS-COV-2 genome from COV-ID of inactivated saliva spiked with without 1,280 virions SARS-COV-2 per µL. All SARS-COV-2 reads align exclusively to expected region of the N gene. Open reading frames of viral genome are depicted via gray boxes below alignment. Inset: scale shows reads per 1,000. (D) Scatter plot for the ratio of SARS-CoV-2 / STATH reads obtained by COV-ID (y axis) versus the number of virions per µL spiked in human saliva (x axis). The threshold was set above the highest values scored in a negative control (dashed line). (E) COV-ID performed on clinical saliva samples. The scatter plot shows the SARS-CoV-2 / STATH read ratio (y axis) versus the viral load in the sample estimated by a clinically approved, qPCR-based diagnostic test. The threshold was set based on the negative controls shown in (D). Experiments and Data generated by R.W-T.

4.3.3 Calibration of COV-ID using an artificial spike-in

Existing deep sequencing approaches for massively parallel COVID-19 testing based on RT-PCR incorporate artificial spike-ins, which serve as an internal calibration controls and allow for better estimates of viral loads by end-point PCR [8, 9]. At the same time, adding to the reactions an artificial substrate for amplification helps minimizing spurious signals as it can "scavenge" viral amplification primers in negative samples. Finally, by providing a baseline amplification even in empty samples, a properly designed spike-in strategy can reduce variance in total amounts of final amplified products across samples, which compresses the dynamic-range of sequence coverage for each patient in a complex pool and, therefore, reduces the risk of inconclusive samples due to low sequencing coverage [9]. To our knowledge, a spike-in approach for LAMP-based quantification has not yet been reported, but we reasoned that it would provide similar benefits in the context of COV-ID. To generate a SARS-CoV-2 spike-in, we synthesized a fragment of the N2 RNA that retained all primer-binding regions for RT-LAMP and contained a divergent 7-nt stretch of sequence to distinguish reads originating from the spike-in from those originating from the natural virus (Figure 4.7 A). After confirming that the spike-in template was efficiently amplified via RT-LAMP with the N2 primer set (Figure 4.7 B), we performed pooled COV-ID on virus-containing saliva in the presence of 20 fg of N2 spike-in RNA. As expected [9], addition of a constant amount of viral spike-in across reactions reduced the variability in total read numbers for individual samples in the final pool (Figure 4.7 C). As discussed above, a narrower range in sequencing output across samples in a pool optimizes the utilization of sequencing reads, and ultimately cost per patient. Because the spike-in provides an internal calibration that is independent of the RNA quality found in saliva, we found that in several cases normalization against the spike-in resulted in lower apparent values for negative samples (Figure 4.7 D). This is likely because in cases where very few STATH reads were obtained, possibly due to degradation of host RNA in the saliva sample, the resulting small denominator inflated the N2/STATH ratio even for SARS-CoV-2 signal that was low in absolute terms and likely spurious. Thus, these data show that spike-in strategies are compatible with the COV-ID workflow and provide a means to stabilize total amplification and read allocation per sample while also offering an additional calibration control to better estimate the viral load in samples where the endogenous STATH mRNA might be below detection due to improper collection or handling.

4.3.4 Simultaneous detection of SARS-CoV-2 and influenza A by COV-ID

Given the challenge of distinguishing early symptoms of COVID-19 from other respiratory infections, we evaluated COV-ID for the simultaneous detection of multiple viral pathogens. Multiple distinct products can be simultaneously amplified by RT-LAMP in the same tube by providing the appropriate primer sets in multiplex, as we demonstrated above by co-amplifying N2 and STATH in the same COV-ID reaction (see Figure 4.2). In fact, simultaneous detection of SARS-CoV-2 and influenza virus by RT-LAMP was previously demonstrated, albeit in a fluorescent-based, low-throughput type of assay [33]. We reasoned that the sequencing-based readout of COV-ID would allow extending this approach to the simultaneous detection of multiple pathogens as well as endogenous (host mRNA) and artificial (spike-in) calibration standards, all in a single reaction. To test the ability of COV-ID to simultaneously detect multiple viral templates, we selected and validated a generic "flu" RT-LAMP primer set that recognizes several strains, including influenza A virus (IAV) and influenza B [33, 34], and modified the BIP and FIP sequence to introduce the COV-ID barcodes and PCR handles individual barcodes (Figure 4.6 E and Table 4.1). We added inactivated SARS-CoV-2 virus (BEI resources) and IAV strain H1N1 RNA (Twist Biosciences) to saliva according to a 3 x 3 matrix of 10,000 copies, 1,000 copies, or 0 copies (Figure 4.3 A), as well as the N2 spike-in control. We performed multiplex COV-ID on these samples using primers sets for STATH, N2 (to detect SARS-CoV-2), and IAV (to detect H1N1) and sequenced to an average depth of 11,000 reads per sample. Both H1N1 and SARS-CoV-2 were detected above background and the signal correlated with the amount of the respective template added to saliva (Figure 4.3 B). Overall, multiplex COV-ID correctly identified samples that contained SARS-CoV-2 (6/6), H1N1 (4/6), or both (6/8) (Figure 4.3 C), indicating the potential and versatility of our sequencing-based RT-LAMP assay.

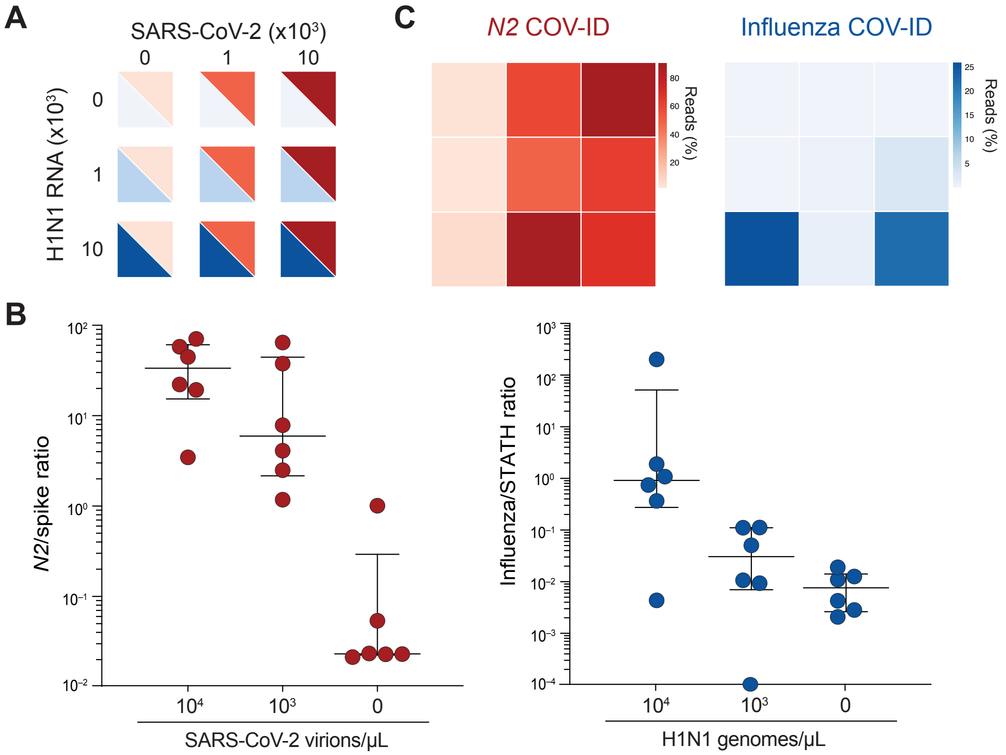
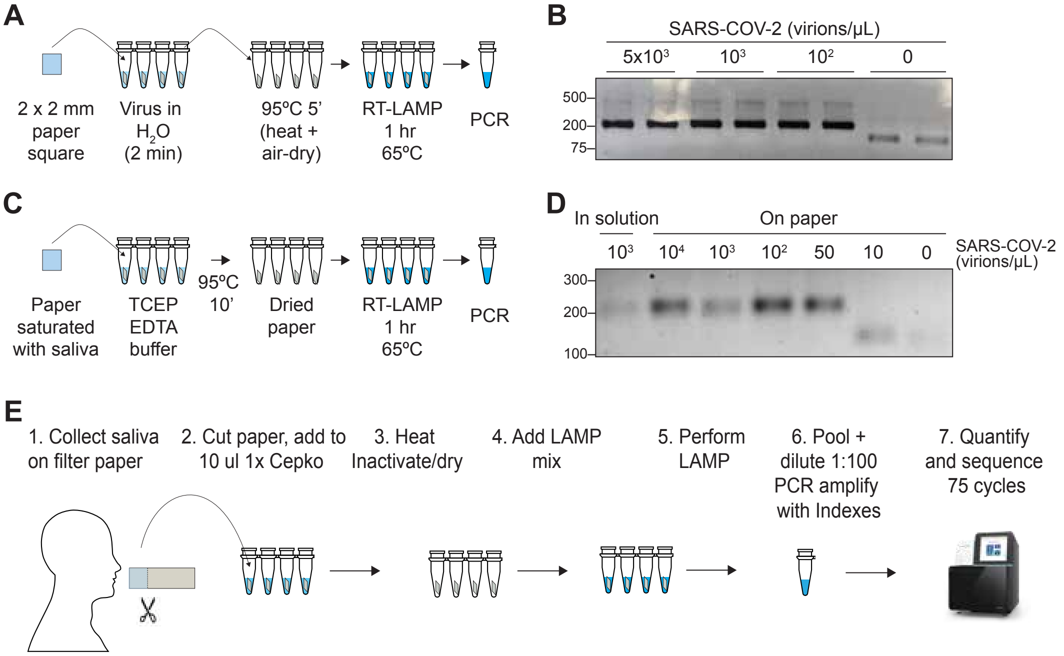


FIGURE 4.3 COV-ID multiplex detection of SARS-COV-2 and Influenza A
(A) TCEP/EDTA treated saliva was spiked with indicated amounts of BEI heat-inactivated SARS-CoV-2 or H1N1 influenza A RNA to the indicated concentration of virions/genomes per µL. 1 µl of saliva was used for COV-ID reactions. (B) COV-ID was performed in technical duplicates on saliva samples from the matrix shown in (A) in the presence of 20 femtograms synthetic N2 spike-in using N2, influenza 32 and STATH COV-ID primers. N2/spike-in and influenza/STATH read ratios are shown. (C) Heatmaps of SARS-CoV-2 (left) or H1N1 (right) COV-ID signal in multiplex reaction. Heatmap is colored according to percentage of viral reads relative to total obtained reads. Experiments and Data generated by R.W-T.

4.3.5 Paper-based saliva sampling for COV-ID

As an additional step toward increasing the throughput of the COV-ID approach, we explored avenues to simplify collection, lower costs, and expedite processing time. Absorbent paper is an attractive alternative to sample vials for collection, given its low cost, wide availability, and smaller environmental footprint. In fact, paper has been used as a means to isolate nucleic acid from biological samples for direct RT-PCR testing [35] as well as RT-LAMP [36, 37]. We sought to determine whether the COV-ID workflow would be compatible with saliva collection on absorbent paper. First, we immersed a small square of Whatman filter paper into water containing various dilutions of inactivated SARS-CoV-2. After 2 min, the paper was removed and transferred to PCR strip tubes followed by heating at 95ºC for 5 minutes to air-dry the sample (Figure 4.4 A). Next, we added the RT-LAMP mix containing the N2 COV-ID primer set directly to the tubes containing the paper squares and let the reaction proceed in the usual conditions. COV-ID PCR products of the correct size were evident in all samples containing viral RNA, with sensitivity of at least 100 virions/µL (Figure 4.4 B) and in none of the controls, demonstrating that the presence of paper does not interfere with the RT-LAMP reaction and subsequent PCR amplification with Illumina adapters. To assay direct COV-ID detection from saliva on paper, we saturated Whatman filter paper squares with saliva containing different amounts of inactivated SARS-CoV-2 virus, which, we reasoned, would be equivalent to a patient collecting their own saliva by chewing on a small piece of absorbent paper. Next, we placed the paper squares into reaction tubes containing TCEP/EDTA inactivation buffer (see Methods) similar to that used for the in-solution samples used in our previous experiments (see Figure 4.1 A). We dried the paper at 95ºC and performed RT-LAMP followed by PCR (Figure 4.4 C). This COV-ID workflow resulted in sequencing-compatible PCR products or the correct size starting from saliva spiked with as few as 50 virions / µL (Figure 4.4 D), suggesting the paper-based approach has similar sensitivity to the conventional in-solution method. Taken together, these data show that the RT-LAMP step of COV-ID is compatible with the presence of paper in the reaction tube and suggest that self-collection of saliva by patients directly on absorbent paper could provide a simple and cost-effective strategy to collect and test thousands of saliva samples for multiple pathogens.



PCR pools	Total samples	Total Reads (Nextseq)	Reads/sample	Sequencing cost/sample
96	9,216	450,000,000	~48,000	$0.16
384	36,864	450,000,000	~12,000	$0.04

FIGURE 4.4 COV-ID on saliva collected on paper

(A) Scheme for COV-ID on viral RNA absorbed on paper. (B) PCR reactions from paper samples immersed in water with indicated viral concentrations then amplified with N2 COV-ID primers. (C) Scheme for COV-ID on saliva spiked with viral and RNA and absorbed on paper. (D) Same as (B) but on saliva absorbed on paper. (E) Paper-based COV-ID workflow and cost calculations. Saliva is collected orally on a precut strip of paper, from which a 2 mm square would be cut out and added to a reaction vessel containing TCEP/EDTA inactivation buffer and processed as shown in (C). Experiments and Data generated by R.W-T. and P.P.S.

4.4 Discussion

Testing strategies are vital to an effective public health response to the COVID-19 pandemic, particularly with the spread of the disease by asymptomatic individuals. An ongoing challenge to COVID-19 testing is the need for massive testing strategies for population-level surveillance that are needed for efficient contact tracing and isolation. As of December 2020, most FDA-approved clinical SARS-CoV-2 diagnostic tests are based on time-consuming and expensive protocols that include RNA purifications and RT-PCR [31] and must be performed by trained personnel in well-equipped laboratories. Point-of-care antigen tests provide a much faster turnaround time and require little manipulation, but there remains limited data on their specificity in real-world applications [38]. Because of reagent limitations and diagnostic testing bottlenecks, prioritization of COVID diagnostic testing continues to be for symptomatic individuals and individuals who are particularly vulnerable for infection after exposure [39]. Private organizations, including colleges and universities, have circumvented some of these challenges by contracting with private laboratories to establish asymptomatic surveillance testing protocols; this is a costly option for population-level surveilling of asymptomatic SARS-CoV-2 infections.

In order to scale testing to the necessary level and frequency, surveillance tests must possess the following qualities: 1) sensitivity, to identify both asymptomatic and symptomatic carriers; 2) simplicity in methodology, to be performed in a number of traditional diagnostic laboratories, without specialized equipment; 3) low cost and easily accessible reagents; 4) ease of collection method; 5) rapid turnaround time to allow for isolation and contract tracing; and 6) ability to co-detect multiple respiratory viruses, given the overlap in patient symptoms. To this end, we have developed COV-ID, an RT-LAMP-based parallel sequencing SARS-CoV-2 detection method that can provide results from tens of thousands of samples per day at relatively low cost to simultaneously detect multiple respiratory viruses.

COV-ID features several key innovations that make it well-suited to high-throughput testing. First, COV-ID uses a two-dimensional barcoding strategy [9], where the same 96 barcodes are used in each RT-LAMP plate, making it possible to pre-aliquot barcodes in 96-well plates ahead of time ("print plates") and store them at -20ºC, simplifying execution of the assay and shortening turnaround times. Second, since RT-LAMP does not require thermal cycling, tens of thousands of samples can be run simultaneously in a standard benchtop-sized incubator or hybridization oven held at 65ºC. Third, individual samples are pooled immediately following RT-LAMP; therefore, a single thermocycler has the potential to process up to 96 or 384 RT-LAMP plates, generating 9,216 or 36,864 individually barcoded samples, respectively (Figure 4.1 A, 4E). Only 96 unique FIP barcodes are required for this scaling; here, we show that 28 out of 32 LAMP barcodes tested were functional for both N2 and STATH. This proof of principle experiment demonstrates the feasibility of generating the library of barcodes required to apply COV-ID to a large population. Notably, COV-ID can generate ready-to-sequence libraries directly from saliva absorbed onto filter paper, which would allow for major streamlining of the often-challenging logistical process of sample collection (Figure 4.4). Thus, COV-ID libraries for thousands and tens of thousands of samples can be generated with relatively minimum effort in biological laboratories with basic equipment and easily accessible reagents. With the pervasive diffusion of deep sequencing technologies in the last few years, most departments and institutes have access to abundant sequencing capacity. With the average throughput of an Illumina NextSeq 500/550, a relatively affordable next-generation sequencer, up to 9,216 (96 RT-LAMPs x 96 pools) can be sequenced at a depth of 48,000 reads per sample, and up to 36,864 (96 RT-LAMPs x 384 pools) can be sequenced at a depth of 12,000 reads, which, we showed, is more than sufficient to obtain information about multiple viral and control amplicons. Considering that reagents for one NextSeq run cost 1,500 U.S. dollars, the theoretical sequencing cost per sample could be as low as $0.04 (Figure 4.4 E). While sequencing instruments are relatively specialized and not ubiquitous, amplified COV-ID DNA libraries could be shipped to remote facilities for sequencing in a cost-effective manner as previously proposed by the inventors of LAMP-seq [14]. Finally, because of the limited sequence space against which reads must be aligned, computational analysis of the resulting data can be performed in a matter of minutes with optimized pipelines, providing results shortly after the sequencing run has completed.

COV-ID has sensitivity of 5–10 virions of SARS-CoV-2 per µL in contrived saliva samples (Figure 4.2 D) and at least 300 virions /µL in saliva collected from patients in a clinical setting (Figure 4.2 E). Given that none of the positive clinical samples that we could test had an estimated viral load $<$ 300 virions /µL, further testing will be required to determine the actual sensitivity of COV-ID. On the other hand, a much larger number of clinical saliva specimens will be needed to determine the extent to which variability in saliva collections and storage might affect sensitivity. Finally, the sensitivity of COV-ID on saliva collected and dried on paper was as low as 50 virions / µL, at least judging by the appearance of a PCR product (Figure 4.4 D). It is possible that deep sequencing of these products would reveal even higher sensitivity. Given that estimates of sensitivity required for effective SARS-CoV-2 surveillance testing are approximately 100 virions / µL [7, 40], COV-ID provides an effective testing platform regardless of the protocol and template used. In conclusion, COV-ID is a flexible platform that can be executed at varying levels of scale with additional flexibility in sample input, making it an attractive platform for surveillance testing. Population-level monitoring of SARS-CoV-2 infections will be critical while vaccines are being distributed to the global population, and continued surveillance will likely remain an effective strategy to protect immune-compromised and unvaccinated members in society and within entities and organizations where regular monitoring is critical to social isolation strategies. To that end, effective, low-cost, multiplexed, and readily-implementable strategies for surveillance testing, such as COV-ID, are important to mitigate the effects of the current and future pandemics.

4.5 Acknowledgments

The authors thank E. Shields for careful proofreading of analysis scripts; B. Morris and R. Collman for the collection and distribution of clinical saliva samples; F. Bushman, S. Sherril-Mix, and Abigail Glascock for sharing RT-qPCR data on the clinical samples; the UPenn rapid assay task force for project feedback; and the gLAMP weekly forum for advice and guidance. R.B. acknowledges financial support from the NIH (R01GM127408 and R01GM138788). R.J. acknowledges financial support from the Burroughs Wellcome Fund and NIH (R01GM137425 and DP2HL147123).

4.6 Author contributions

R.W-T., C.A.T. and R.B. designed the experiments. R.W-T., P.P.S., P.L., and J.L. carried out experiments. R.W-T. performed all computational and statistical analyses. B.S.A. prepared clinical samples. R.W-T., P.P.S, C.A.T. and R.B. wrote the paper. All co-authors read and approved the manuscript.

4.7 Disclosures

Inventors R.W-T., C.A.T. and R.B. have filed a patent related to the COV-ID method presented here.

4.8 Data Availability

Next generation sequencing data generated for this study are available at the NCBI GEO with accession GSE172118.

4.9 Methods

4.9.1 RT-LAMP primer design

Primers against ACTB were designed using PrimerExplorerV5 (https://primerexplorer.jp/e/) using default parameters and including loop primers (Table 4.1).

For COV-ID, priming sequences for PCR were inserted in FIP and BIP primers between the target homology regions (F1c and F2, and B1c and B2, respectively, see Figure 4.5). After testing, we determined that 12 nts and 11 nts were most effective for the P5 and P7 binding regions, respectively, being the shortest insertion that allowed reliable PCR amplification from LAMP products without impacting LAMP efficiency. In addition a 5 nt barcode sequence was inserted at the immediate 3' end of the P5-binding region of the FIP primer.

4.9.2 LAMP barcode design

Starting from the total possible 1,024 unique 5-nt barcodes, we removed those that matched any sequence within the RT-LAMP primers used in this study (Table 4.1) in either sense or anti-sense orientation. From the remaining pool, we selected 32 barcodes with hamming distance of at least 2 between all candidates. We tested FIPs incorporating candidate barcodes for *ACTB*, *STATH*, *N2*, and IAV primer sets on saliva RT-LAMP with 1,000 copies of target amplicon. Primers that failed to show LAMP signal by real time fluorescence monitoring or generate expected PCR product were discarded. Final usable barcodes are provided in (Table 4.2).

4.9.3 Saliva preparation

We prepared 100x TCEP/EDTA buffer (250 mM TCEP, 100 mM EDTA, 1.15 N NaOH) [28]. TCEP/EDTA buffer was added to human saliva at 1:100 volume, then samples were capped, vortexed to mix and heated in a thermocycler (95ºC 5 min, 4ºC hold) until ready to use for RT-LAMP. When indicated, heat-inactivated SARS-CoV-2 (BEI Resources Cat. NR-52286) or H1N1 genomic RNA (Twist Biosciences Cat. 103001) was added to inactivated saliva prior to RT-LAMP.

4.9.4 N2 spike-in synthesis

To prepare the *in vitro* transcription template for SARS-CoV-2 N2 spike-in RNA, we performed RT-PCR using Power SYBR RNA-to-Ct kit (Thermo Cat. 4389986) of heat inactivated SARS-CoV-2 (BEI Resources Cat. NR-52286) using the following primers: N2-B3 and N2-spike-T7 S. PCR product was purified and used as a template for in vitro transcription using HiScribe T7 transcription kit (NEB Cat. E2040S). RNA was purified with Trizol (Thermo Cat. 15596026), quantified via A_{260}, then aliquoted in BTE buffer (10 mM bis-tris pH 6.7, 1 mM EDTA) and stored at -80ºC. Primers used and final spike-in sequence are provided in Table 4.1 .

4.9.5 RT-LAMP

All RT-LAMP reactions were set up in clean laminar flow hoods and all steps before and after LAMP were carried out in separate lab spaces to avoid contamination. RT-LAMP reactions were set up on ice as follow: for each amplicon 5 or 6 LAMP primers were combined into 10x working stock at established concentrations: 16 µM FIP, 16 µM BIP, 4 µM LF, 4 µM LB, 2 µM F3, 2 µM B3. For multiplexed COV-ID reactions 10x working primer mixes for each amplicon were either added proportionally so that the total primer content remained constant, or mixed so that BIP and FIP primers were scaled down depending on amplicon number while remaining primers (LF and/or LB, F3, B3) were kept at same concentration as in single reactions. Each 10 µL RT-LAMP reaction mix consisted of 1x Warmstart LAMP 2x Master Mix (NEB Cat. E1700S), 0.7 µM dUTP (Promega Cat. U1191), 1 µM SYTO-9 (Thermo Cat. S34854), 0.1 µL Thermolabile UDG (Enzymatics Cat. G5020L), 1 µL of saliva template and optionally 20 fg of N2 Spike RNA. Reactions were prepared in qPCR plates or 8-well strip tubes, sealed, vortexed and centrifuged briefly, then incubated in either a QuantStudio Flex 7 or StepOnePlus instrument (Thermo) for 65ºC 1 hr. Real-time fluorescence measurements were recorded every 30 sec to monitor reaction progress but were not used for data analysis. Following LAMP the reactions were heated at 95ºC 5 min to inactivate LAMP enzymes.

4.9.6 Library construction by PCR amplification

All post-LAMP steps were carried out on a clean bench separate from LAMP reagents and workspace. For individual LAMP samples, LAMP amplicons were diluted either 1:100 or 1:1,000 in water. For pooling of individually barcoded LAMP reactions, equal amounts of all LAMP reactions were combined and then either diluted 1:1000 or purified via SPRIselect beads (Beckman Coulter Cat. B23317) using a bead-to-reaction ratio of 0.1x. Purified material was diluted to final 100-fold dilution relative to LAMP. 1 µL of diluted LAMP material was used as a template for PCR using OneTaq DNA polymerase (NEB Cat. M0480L) with 100 nM each of custom dual-indexed Illumina P5 and P7 primers in either 10 or 25 µL reaction (Table 4.1). PCR reactions were incubated as follows: (25 cycles of stage 1 [94ºC x 15 sec, 45ºC x 15 sec, 68ºC x 10 sec], 10 cycles of Stage 2 [94ºC x 15 sec, 68ºC x 10 sec], 68ºC x 1 min, 4ºC x ∞). Note, for initial pilot COV-ID and clinical sample experiments (Figure 4.2 D–E, Figure 4.6 C) PCR incubation was performed as above with modification: [Stage 1 x 10 cycles, Stage 2 x 25 cycles]. PCR products were resolved on 2% agarose gel to confirm library size, then all were pooled and purified via MinElute PCR purification kit (Qiagen Cat. 28004) and quantified using either Qubit dsDNA High Sensitivity kit (Thermo Cat. Q32851) or Kapa Library Quantification Kit for Illumina (Kapa Cat. 07960140001).

4.9.7 Patient samples

Clinical saliva samples were obtained and characterized as part of a separate study at the University of Pennsylvania [41] and collected under Institutional Review Board (IRB)-approved protocols (IRB protocol #842613 and #813913). Briefly, salivary samples were collected from possible SARS-CoV-2 positive patients at one of three locations: (1) Penn Presbyterian Medical Center Emergency Department, (2) Hospital of the University of Pennsylvania Emergency Department, and (3) Penn Medicine COVID-19 ambulatory testing center. Inclusion criteria including any adult (age $>$ 17 years) who underwent SARS-CoV-2 testing via standard nasopharyngeal swab at the same visit. Patients with known COVID-19 disease who previously tested positive previously were excluded. After verbal consent was obtained by a trained research coordinator, patients were instructed to self-collect saliva into a sterile specimen container which was then placed on ice until further processing for analysis.

4.9.8 Paper COV-ID

Squares of Whatman no. 1 filter paper (2 mm x 2 mm) were cut using a scalpel on a clean surface under a laminar flow hood and stored at room temperature until used. Using ethanol-sterilized fine-nosed tweezers a single square was dipped twice into unprocessed, freshly collected saliva with or without added SARS-CoV-2 (BEI Resources Cat. NR-52286) until saliva was saturated on paper by eye. Paper was then transferred to well of 96-well plate containing 10 ul of 1x TCEP/EDTA buffer (2.5 mM TCEP, 1 mM EDTA, 1.15 NaOH). Plate was placed on heat block inside laminar flow hood or inside open thermocycler and incubated at 95ºC x 10 min. 10 ul RT-LAMP mixture was prepared as described above in the absence of the N2 Spike RNA. 10 ul of RT-LAMP reaction mixture was added to each paper strip, then plate was sealed and incubated 65ºC x 1 hr, 95ºC x 5 min in QuantStudio Flex 7 (ThermoFisher). 1 ul of each reaction was diluted 1:100 and PCR amplified as described above.

4.9.9 Sequencing

Libraries were sequenced on one of the following Illumina instruments: MiSeq, NextSeq 500, NextSeq 550, NovaSeq 6000 and sequenced using single end programs with a minimum of 40 cycles on Read 1 and 8 cycles for index 1 (on P7) and index 2 (on P5).

4.9.10 Sequence Analysis

Reads were filtered for optical quality using FASTX-toolkit utility fastq_quality_filter (http://hannonlab.cshl.edu/fastx_toolkit/), then cutadapt [42] was used to remove adapters and demultiplex LAMP barcodes. Reads were aligned to a custom index containing SARS-CoV-2 genome (NC_045512.2), Influenza H1N1 coding sequences (NC_026431.1, NC_026432.1, NC_026433.1, NC_026434.1, NC_026435.1, NC_026436.1, NC_026437.1, NC_026438.1), STATH coding sequence (NM_003154.3), and custom N2 spike sequence (Table 4.1) ,target sequences using bowtie2[43] with options –no-unal and –end-to-end. Alignments with greater than 1 mismatch were removed and the number of reads mapping to each target were extracted and output in a table format for each barcode (sample).

4.9.11 Data Availability

4.9.12 Supplemental Tables

CI_N2	
CI_N2-FIP	TTCCGAAGAACGCTGAAGC CTCTTCCGATCT NNNNN GGAACTGATTACAAACATTGGCC
CI_N2-BIP	CGCATTGGCATGGAAGTCA CATCTCCGAGC CAATTTGATGGCACCTGTGTA
N2-B3	GACTTGATCTTTGAAATTTGGATCT
N2-F3	ACCAGGAACTAATCAGACAAG
N2-LB	CTTCGGGAACGTGGTTGACC
N2-LF	GGGGGCAAATTGTGCAATTTG
CI_Act1	
Act1-FIP	TGCCGCCAGACAGCACTGTG CTCTTCCGATCT NNNNN TGAAGTGTGACGTGGACATC
Act1-BIP	TTGCCGACAGGATGCAGAAGG CATCTCCGAGC GCGCTCAGGAGGAGCAAT
Act1-LB	CCTGGCACCCAGCACAATGAAG
Act1-B3	GCCGATCCACACGGAGTAC
Act1-LF	GGCGTACAGGTCTTTGCG
Act1-F3	GGCATCCACGAAACTACCTT
CI_STATH	
CI_STATH-FIP	AGCTCCAATCATGGAAACCATG CTCTTCCGATCT NNNNN AGCCAACTATGAAGTTCCTTG
CI_STATH-BIP	AAGATTCGGTTATGGGTATGGC CATCTCCGAGC GTTGTGGGTATAGTGGTTGTTC
STATH-F3	GTAGCACATCATCTCTTGAAGCT
STATH-LF	GAGCCAAGATGAAGGCAA
STATH-B3	GCCTCAATAATCATGTCCTGCA
CI_IAV	
CI_IAV-FIP	TTAGTCAGAGGTGACARRATTG CTCTTCCGATCT NNNNN CAGATCTTGAGGCTCTC
CI_IAV-BIP	TTGTKTTCACGCTCACCGTG CATCTCCGAGC TTTGGACAAAGCGTCTACG
IAV-LF	GTCTTGTCTTTAGCCA
IAV-LB	CMAGTGAGCGAGGACTG
IAV-F3	GACTTGAAGATGTCTTTGC
IAV-F3_2	GACTGGAAAGTGTCTTTGC
IAV-B3	TRTTATTTGGGTCTCCATT
IAV-B3_2	TRTTGTTTGGGTCCCCATT
P5 primer (index nt = N)	AATGATACGGCGACCACCGAGATCTACAC NNNNNNNN ACACTCTTTCCCTACACGACGCTCTTCCGATCT
P7 primer (index nt = N)	CAAGCAGAAGACGGCATACGAGAT NNNNNNNN GTCTCGTGGGCTCGGAGATG
N2 Spike T7-S	TAATACGACTCACTATAGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAGTACATG CAATTTGCCCCCAGCGC
N2-B3	GACTTGATCTTTGAAATTTGGATCT
N2_Spike IVT template sequence	TAATACGACTCACTATAGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAGTACATGCAATTTGCCCCCAGCGCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATCAAATTGGATGACAAAGATCCAAATTTCAAAGATCAAGTC

TABLE 4.1 Oligonucleotide sequences

Barcode	**Sequence**	**SARS N2**	**Statherin STATH**	**Actin Act1**	**Influenza A IAV**	**Effective (N2, STATH)**	**Effective (All 4 targets)**
i01	CCTGT	Effective	Effective	Failed	Effective	Yes	No
i02	GTTAC	Effective	Effective	Effective	Effective	Yes	Yes
i03	GCATC	Effective	Effective	Effective	Failed	Yes	No
i04	TGCGA	Effective	Effective	Effective	Effective	Yes	Yes
i05	ATCAT	Effective	Effective	Effective	Failed	Yes	No
i06	GCTTG	Effective	Effective	Effective	Effective	Yes	Yes
i07	CACTG	Effective	Effective	Effective	Failed	Yes	No
i08	AGTCA	Effective	Failed	Effective	Effective	No	No
i09	CTAGA	Effective	Effective	Effective	Effective	Yes	Yes
i10	TAACG	Effective	Effective	Effective	Effective	Yes	Yes
i11	CTAGT	Effective	Effective	Effective	Effective	Yes	Yes
i12	AGCCT	Effective	Effective	Effective	Failed	Yes	No
i13	AAATG	Failed	Effective	Effective	Effective	No	No
i14	AGCCC	Effective	Effective	Effective	Effective	Yes	Yes
i15	ATATC	Effective	Effective	Effective	Effective	Yes	Yes
i16	CCAAG	Effective	Effective	Failed	Effective	Yes	No
i17	CGAGT	Effective	Effective	Effective	Failed	Yes	No
i18	CGATG	Effective	Effective	Failed	Effective	Yes	No
i19	CGCGG	Effective	Effective	Effective	Effective	Yes	Yes
i20	CGGAT	Effective	Effective	Effective	Effective	Yes	Yes
i21	GCGCC	Effective	Effective	Effective	Effective	Yes	Yes
i22	GGCGA	Effective	Effective	Effective	Effective	Yes	Yes
i23	GGTGT	Effective	Effective	Effective	Effective	Yes	Yes
i24	GTCAA	Effective	Effective	Effective	Effective	Yes	Yes
i25	GTCGC	Effective	Effective	Effective	Effective	Yes	Yes
i26	GTGAT	Effective	Effective	Effective	Effective	Yes	Yes
i27	TAAAC	Failed	Effective	Effective	Effective	No	No
i28	TACTA	Effective	Effective	Effective	Effective	Yes	Yes
i29	TAGAG	Failed	Effective	Effective	Effective	No	No
i30	TCTAG	Effective	Effective	Effective	Effective	Yes	Yes
i31	TGAAT	Effective	Effective	Effective	Effective	Yes	Yes
i32	TTATA	Effective	Effective	Effective	Effective	Yes	Yes
					Barcode yield	28/32	20/32

TABLE 4.2 COV-ID barcode validation data

4.10 Supplementary Data

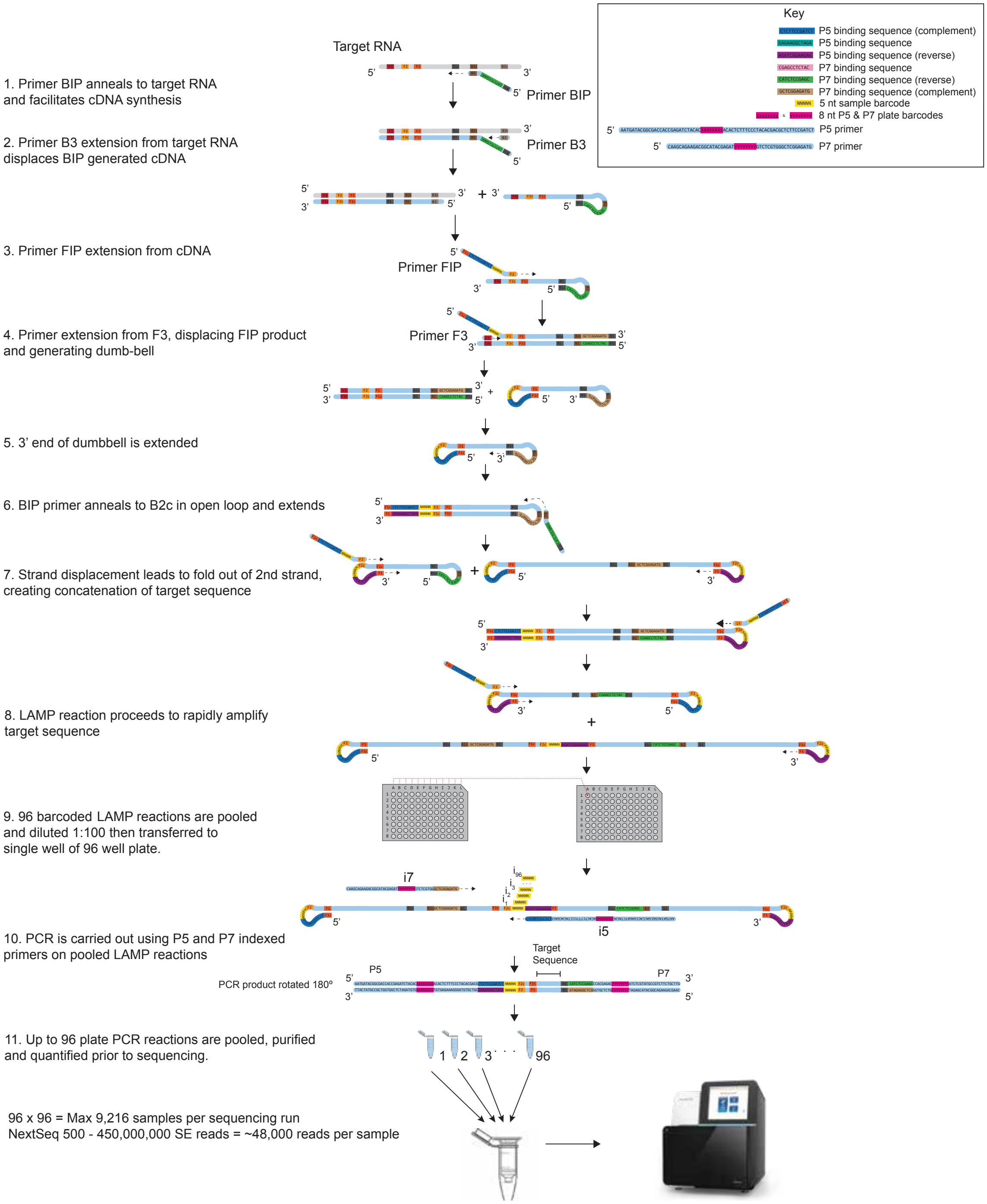
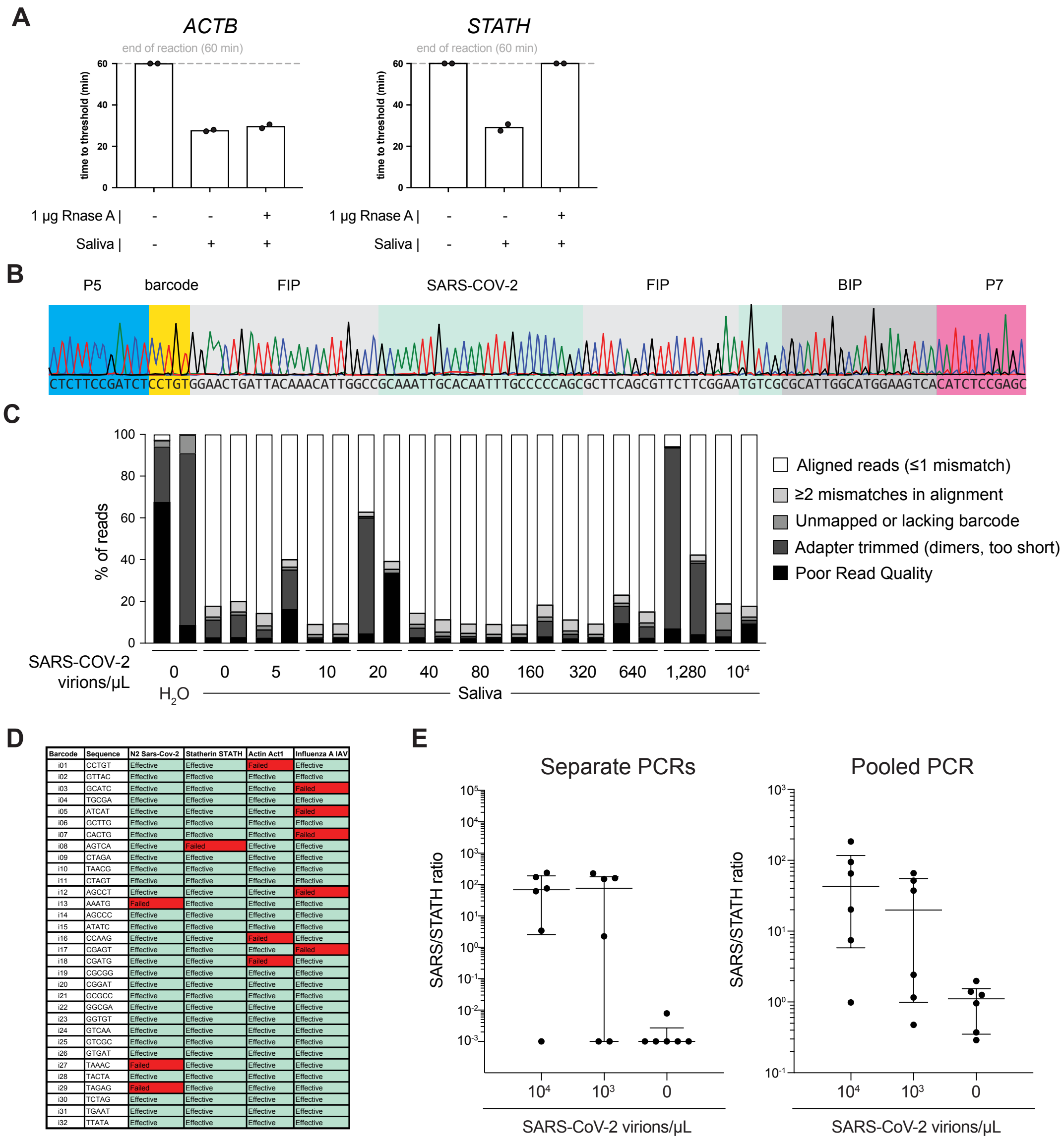

Key
CTCTTCCGATCT P5 binding sequence (complement)
GAGAAGGCTAGA P5 binding sequence
AGATCGGAAGAG P5 binding sequence (reverse)
CGAGCCTCTAC P7 binding sequence
CATCTCCGAGC P7 binding sequence (reverse)
GCTCGGAGATG P7 binding sequence (complement)
NNNNN 5 nt sample barcode
XXXXXXXX & YYYYYYYY 8 nt P5 & P7 plate barcodes
5' AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCT P5 primer
5' CAAGCAGAAGACGGCATACGAGATYYYYYYYYGTCTCGTGGGCTCGGAGATG P7 primer
Target RNA
5'
3'
Primer BIP
1. Primer BIP anneals to target RNA and facilitates cDNA synthesis
2. Primer B3 extension from target RNA displaces BIP generated cDNA
Primer B3
3. Primer FIP extension from cDNA
Primer FIP
4. Primer extension from F3, displacing FIP product and generating dumb-bell
Primer F3
5. 3' end of dumbbell is extended
6. BIP primer anneals to B2c in open loop and extends
7. Strand displacement leads to fold out of 2nd strand, creating concatenation of target sequence
8. LAMP reaction proceeds to rapidly amplify target sequence
9. 96 barcoded LAMP reactions are pooled and diluted 1:100 then transferred to single well of 96 well plate.
A B C D E F G H I J K L
1 2 3 4 5 6 7 8
i7
i96
i3
i2
i1
i5
10. PCR is carried out using P5 and P7 indexed primers on pooled LAMP reactions
Target Sequence
P5
P7
PCR product rotated 180°
11. Up to 96 plate PCR reactions are pooled, purified and quantified prior to sequencing.
1
2
3
. . .
96
96 x 96 = Max 9,216 samples per sequencing run
NextSeq 500 - 450,000,000 SE reads = ~48,000 reads per sample

FIGURE 4.5 Detailed COV-ID Mechanism (related to FIGURE 4.1)

Steps of COV-ID protocol are depicted, showing RT-LAMP mechanism and the ultimate amplicon that is sequenced. For clarity, only selected steps of RT-LAMP reaction are shown and loop primer intermediates are not depicted. For full LAMP mechanism see [21]



Barcode	Sequence	N2 Sars-Cov-2	Statherin STATH	Actin Act1	Influenza A IAV
i01	CCTGT	Effective	Effective	Failed	Effective
i02	GTTAC	Effective	Effective	Effective	Effective
i03	GCATC	Effective	Effective	Effective	Failed
i04	TGCGA	Effective	Effective	Effective	Effective
i05	ATCAT	Effective	Effective	Effective	Failed
i06	GCTTG	Effective	Effective	Effective	Effective
i07	CACTG	Effective	Effective	Effective	Failed
i08	AGTCA	Effective	Failed	Effective	Effective
i09	CTAGA	Effective	Effective	Effective	Effective
i10	TAACG	Effective	Effective	Effective	Effective
i11	CTAGT	Effective	Effective	Effective	Effective
i12	AGCCT	Effective	Effective	Effective	Failed
i13	AAATG	Failed	Effective	Effective	Effective
i14	AGCCC	Effective	Effective	Effective	Effective
i15	ATATC	Effective	Effective	Effective	Effective
i16	CCAAG	Effective	Effective	Failed	Effective
i17	CGAGT	Effective	Effective	Effective	Failed
i18	CGATG	Effective	Effective	Failed	Effective
i19	CGCGG	Effective	Effective	Effective	Effective
i20	CGGAT	Effective	Effective	Effective	Effective
i21	GCGCC	Effective	Effective	Effective	Effective
i22	GGCGA	Effective	Effective	Effective	Effective
i23	GGTGT	Effective	Effective	Effective	Effective
i24	GTCAA	Effective	Effective	Effective	Effective
i25	GTCGC	Effective	Effective	Effective	Effective
i26	GTGAT	Effective	Effective	Effective	Effective
i27	TAAAC	Failed	Effective	Effective	Effective
i28	TACTA	Effective	Effective	Effective	Effective
i29	TAGAG	Failed	Effective	Effective	Effective
i30	TCTAG	Effective	Effective	Effective	Effective
i31	TGAAT	Effective	Effective	Effective	Effective
i32	TTATA	Effective	Effective	Effective	Effective

FIGURE 4.6 Optimization of COV-ID in human saliva (related to FIGURE 4.2)

(A) Validation of control human amplicons for RT-LAMP on saliva. RT-LAMP of TCEP/EDTA inactivated saliva was performed with conventional RT-LAMP primer sets for ACTB and STATH in the presence or absence of RNase A. (B) Saliva COV-ID sequence validation. Single saliva COV-ID reaction using N2 primers was sequenced by the Sanger method. (C) Characterization of COV-ID sequencing libraries. Breakdown of reads for sequence data presented in Figure 4.2D. Samples without added template consist of predominantly adapter dimers. (D) Validation of COV-ID LAMP barcodes. 32 potential barcodes were tested for LAMP primer sets indicated, incompatible barcodes are marked in red. (E) Validation of pooled PCR. COV-ID was performed on saliva samples using unique LAMP barcodes. The RT-LAMP reactions were then amplified either by individual PCR or by first pooling and then performing a single PCR on the pool. Experiments and Data generated by R.W-T.

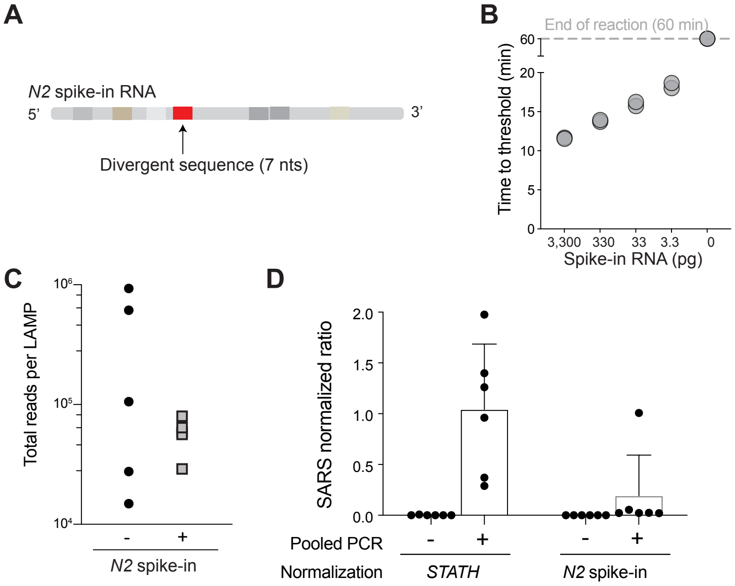


FIGURE 4.7 Spike-in strategy for COV-ID (related to FIGURE 4.3)

(A) Synthetic N2 Spike RNA. SARS-CoV-2 N2 RNA fragment was synthesized including 7 nt divergent sequence inside the forward loop primer-binding site, maintaining all other LAMP primer binding sites and identical GC content. (B) RT-LAMP using COV-ID N2 primers was carried out on indicated amounts of spike-in RNA, showing rapid amplification down to picogram quantities of added template. (C) Total number of reads per barcode in COV-ID pool obtained by including (+) or omitting (-) the N2 spike-in. (D) Spurious COV-ID signal for the N2 amplicon in negative control samples after normalization either to the STATH control in absence of spike-in (left) or to the N2 spike-in control. Experiments and Data generated by R.W-T.

4.11 References

1. Warneford-Thomson, R. *et al.* COV-ID: A LAMP sequencing approach for high-throughput co-detection of SARS-CoV-2 and influenza virus in human saliva. *medRxiv.* https://www.medrxiv.org/content/early/2021/04/23/2021.04.23.21255523 (2021).

2. Haug, N. *et al.* Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat Hum Behav* **4,** 1303–1312 (2020).

3. Tian, H. *et al.* An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368,** 638–642 (2020).

4. Taipale, J., Romer, P. & Linnarsson, S. Population-scale testing can suppress the spread of COVID-19. *medRxiv,* 2020.04.27.20078329 (2020).

5. Endo, A., Centre for the Mathematical Modelling of Infectious Diseases, C.-W. G., Abbott, S., Kucharski, A. J. & Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res* **5,** 67 (2020).

6. Holt, E. Slovakia to test all adults for SARS-CoV-2. *Lancet* **396,** 1386–1387 (2020).

7. Larremore, D. B. *et al.* Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Sci Adv* **7** (2021).

8. Bloom, J. S. *et al.* Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. *medRxiv,* 2020.08.04.20167874 (2020).

9. Yelagandula, R. *et al.* SARSeq, a robust and highly multiplexed NGS assay for parallel detection of SARS-CoV2 and other respiratory infections. *medRxiv,* 2020.10.28.20217778 (2020).

10. Aynaud, M.-M. *et al.* A Multiplexed, Next Generation Sequencing Platform for High-Throughput Detection of SARS-CoV-2. *medRxiv,* 2020.10.15.20212712 (2020).

11. Chappleboim, A. *et al.* ApharSeq: An Extraction-free Early-Pooling Protocol for Massively Multiplexed SARS-CoV-2 Detection. *medRxiv,* 2020.08.08.20170746 (2020).

12. James, P. *et al.* LamPORE: rapid, accurate and highly scalable molecular screening for SARS-CoV-2 infection, based on nanopore sequencing. *medRxiv,* 2020.08.07.20161737 (2020).

13. Dao Thi, V. L. *et al.* A colorimetric RT-LAMP assay and LAMP-sequencing for detecting SARS-CoV-2 RNA in clinical samples. *Sci Transl Med* **12** (2020).

14. Schmid-Burgk, J. L. *et al.* LAMP-Seq: Population-Scale COVID-19 Diagnostics Using a Compressed Barcode Space. *bioRxiv,* 2020.04.06.025635 (2020).

15. Li, S. *et al.* Simultaneous detection and differentiation of dengue virus serotypes 1-4, Japanese encephalitis virus, and West Nile virus by a combined reverse-transcription loop-mediated isothermal amplification assay. *Virol J* **8,** 360 (2011).

16. Shirato, K. *et al.* Detection of Middle East respiratory syndrome coronavirus using reverse transcription loop-mediated isothermal amplification (RT-LAMP). *Virol J* **11,** 139 (2014).

17. Calvert, A. E., Biggerstaff, B. J., Tanner, N. A., Lauterbach, M. & Lanciotti, R. S. Rapid colorimetric detection of Zika virus from serum and urine specimens by reverse transcription loop-mediated isothermal amplification (RT-LAMP). *PLoS One* **12,** e0185340 (2017).

18. Enomoto, Y. *et al.* Rapid diagnosis of herpes simplex virus infection by a loop-mediated isothermal amplification method. *J Clin Microbiol* **43,** 951–5 (2005).

19. Augustine, R. *et al.* Loop-Mediated Isothermal Amplification (LAMP): A Rapid, Sensitive, Specific, and Cost-Effective Point-of-Care Test for Coronaviruses in the Context of COVID-19 Pandemic. *Biology (Basel)* **9** (2020).

20. United States Food and Drug Administration. Color Genomics SARS-CoV-2 RT-LAMP Diagnostic Assay - EUA Summary. https://www.fda.gov/media/138249/download (2020).

21. Nagamine, K., Hase, T. & Notomi, T. Accelerated reaction by loop-mediated isothermal amplification using loop primers. *Mol Cell Probes* **16,** 223–9 (2002).

22. Notomi, T. *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res* **28,** E63 (2000).

23. Yamagishi, J. *et al.* Serotyping dengue virus with isothermal amplification and a portable sequencer. *Sci Rep* **7,** 3510 (2017).

24. Butler, D. J. *et al.* Shotgun Transcriptome and Isothermal Profiling of SARS-CoV-2 Infection Reveals Unique Host Responses, Viral Diversification, and Drug Interactions. *bioRxiv,* 2020.04.20.048066 (2020).

25. Ranoa, D. R. E. *et al.* Saliva-Based Molecular Testing for SARS-CoV-2 that Bypasses RNA Extraction. *bioRxiv,* 2020.06.18.159434 (2020).

26. Myhrvold, C. *et al.* Field-deployable viral diagnostics using CRISPR-Cas13. *Science* **360,** 444–448 (2018).

27. Lalli, M. A. *et al.* Rapid and extraction-free detection of SARS-CoV-2 from saliva with colorimetric LAMP. *medRxiv,* 2020.05.07.20093542 (2020).

28. Rabe, B. A. & Cepko, C. SARS-CoV-2 detection using isothermal amplification and a rapid, inexpensive protocol for sample inactivation and purification. *Proc Natl Acad Sci U S A* **117,** 24450–24458 (2020).

29. Babiker, A., Myers, C. W., Hill, C. E. & Guarner, J. SARS-CoV-2 Testing. *Am J Clin Pathol* **153,** 706–708 (2020).

30. Satoh, T. *et al.* Development of mRNA-based body fluid identification using reverse transcription loop-mediated isothermal amplification. *Anal Bioanal Chem* **410,** 4371–4378 (2018).

31. MacKay, M. J. *et al.* The COVID-19 XPRIZE and the need for scalable, fast, and widespread testing. *Nat Biotechnol* **38,** 1021–1024 (2020).

32. Torres, C. *et al.* LAVA: an open-source approach to designing LAMP (loop-mediated isothermal amplification) DNA signatures. *BMC Bioinformatics* **12,** 240 (2011).

33. Zhang, Y. & Tanner, N. A. Development of Multiplexed RT-LAMP for Detection of SARS-CoV-2 and Influenza Viral RNA. *medRxiv,* 2020.10.26.20219972 (2020).

34. Takayama, I. *et al.* Development of real-time fluorescent reverse transcription loop-mediated isothermal amplification assay with quenching primer for influenza virus and respiratory syncytial virus. *J Virol Methods* **267,** 53–58 (2019).

35. Glushakova, L. G. *et al.* Detection of chikungunya viral RNA in mosquito bodies on cationic (Q) paper based on innovations in synthetic biology. *J Virol Methods* **246,** 104–111 (2017).

36. Kellner, M. J. *et al.* A rapid, highly sensitive and open-access SARS-CoV-2 detection assay for laboratory and home testing. *bioRxiv,* 2020.06.23.166397 (2020).

37. Yaren, O. *et al.* Ultra-rapid detection of SARS-CoV-2 in public workspace environments. *medRxiv,* 2020.09.29.20204131 (2020).

38. Pettengill, M. A. & McAdam, A. J. Can We Test Our Way Out of the COVID-19 Pandemic? *J Clin Microbiol* **58** (2020).

39. Schuetz, A. N. *et al.* When Should Asymptomatic Persons Be Tested for COVID-19? *J Clin Microbiol* **59** (2020).

40. Mina, M. J., Parker, R. & Larremore, D. B. Rethinking Covid-19 Test Sensitivity - A Strategy for Containment. *N Engl J Med* **383,** e120 (2020).

41. Sherrill-Mix, S. *et al.* LAMP-BEAC: Detection of SARS-CoV-2 RNA Using RT-LAMP and Molecular Beacons. *medRxiv,* 2020.08.13.20173757 (2020).

42. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 3 (2011).

43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9,** 357–9 (2012).

5. CONCLUSIONS

5.1 Development of a novel method to identify RNA-protein interactions

In chapter two I described RBR-ID, a method we developed to identify RNA-binding proteins and the protein regions involved in RNA-binding across the nuclear proteome. By using photosensitive nucleotide analogs and ultraviolet radiation, we induce RNA-protein crosslinks that affect quantification via bottom-up (peptide-based) mass spectrometry. With the aid of several replicates to overcome mass spectrometry-associated variability, we can apply RBR-ID to identify RBPs in an unbiased manner. This method dispenses with the need for enrichment of polyA RNA and therefore can profile RBPs binding to any species of RNA. Focusing on the nuclear proteome of mouse embyronic stem cells, we identify 803 RBPs of which 427 were previously unannotated as binding to RNA. The novel putative RBPs were enriched for gene ontology terms involved in chromatin binding and regulation, as well as histone-binding domains such as the chromodomain and bromodomains, and we successfully validated the RNA-binding activity of several novel RBPs *in vivo*. The peptide-level resolution afforded by RBR-ID also permits identification of RBRs in an *in vitro* context using recombinant proteins and RNA. We confirmed the utility of RBR-ID *in vitro* using the well-studied viral RBP MS2CP [1], correctly identifying the known RBR of the protein and observing the highest RBR-ID signal from a peptide directly abutting the protein-RNA interface. Our *in vivo* RBR-ID dataset further revealed RBR regions inside the RBP positive hits, including many well known RNA-binding domains such as RRM, KH and dsRBD. Overlays of RBR-ID data onto published RNA-protein structures showed peptide hits clustered around RNA-binding interfaces and further confirmed the utility of the peptide-level resolution of RBR-ID. We validated 6 RBRs on RBR-ID hits *in vivo*, demonstrating that the RBR identified by RBR-ID was the primary contributor to RNA-binding activity. Development of this method has provided an additional tool to profile protein-RNA interactions in different contexts and revealed more about the characteristics of chromatin-associated RBPs.

5.1.1 Recent advances and future directions in RNA-protein interaction research

Since publication of our manuscript in 2016 it has received over 100 citations, and several groups have validated RBRs that we identified, including BRD4 [2], TET2 [3], and CTCF [4]. In the case of chromatin architectural protein CTCF, RBR-ID identified two RBRs inside the ZF1 and ZF10 zinc finger domains that are not known to bind DNA. Mutants made removing each of the RBRs individually led to a significant reduction in CTCF-RNA binding via PAR-CLIP [4]. Expression of these mutants in mESCs led to loss of CTCF from many sites on chromatin, large scale changes in gene expression and loss of chromatin looping. Fascinatingly, the set of CTCF binding sites lost in each RBR mutant showed virtually no overlap with each other, with ZF1 mutants losing binding primarily inside gene promoters while ZF10 mutant were lost from intronic and intergenic regions [4]. Another study using a similar mutant in the ZF10 domain found loss of CTCF from roughly half of its binding sites [5]. Assuming that these effects are due to changes in RNA-mediated chromatin interactions, it would indicate the modular function of RBRs in binding to specific RNA ligands with perhaps distinct purposes. Investigating this exciting hypothesis could provide a significant step forward in our understanding of how RNA interactions help regulate chromatin structure.

RBR-ID identified an RBR in the catalytic domain of TET2, suggesting that this important regulator involved in active DNA demethylation might also modify RNA, supported by evidence that TET proteins can modify RNA in Drosophila [6]. Follow-up studies confirmed that RNA is modified by TET2 in mammalian cells to generate 5-hydroxymethylcytosine (5-hmC) to destabilize RNAs in different contexts [3, 7]. In addition, work by Chongsheng He and others in our lab have demonstrated that in mESCs, TET2 binds and modifies tRNAs to facilitate generation of tRNA fragments [8]. These findings demonstrate how RBR-ID has helped expand our understanding of several important chromatin regulatory enzymes.

We have also since applied *in vitro* RBR-ID to map other chromatin regulatory proteins. The chromatin remodeler ATRX is required for proper X inactivation and binds RNA–including the lncRNA *Xist*–but the function of these interactions is unclear [9]. We carried out RBR-ID on recombinant ATRX with an *Xist* RNA fragment and identified several peptides within aa 400-750 that we classified as an RBR. This region was necessary and sufficient to bind *Xist in vitro*

[10]. Nuclear fractionation in the presence of RNase lead to loss of WT ATRX from the chromatin fraction, similar to behavior of PRC2 subunits in response to RNase treatment [11] and implicating RNA in regulating ATRX chromatin localization. However, ΔRBR ATRX showed no sensitivity to RNase treatment, providing strong evidence that this region interacts with RNA *in vivo* and is necessary for proper chromatin localization [10]. This example demonstrates the practical utility of our screening approach in identifying RNA-protein interactions and providing tools to address the function of these interactions in chromatin regulation.

’Click’ assisted RNA interactome capture

Interest in RNA-protein interactions has risen steadily in the last few years, concomitant with a rapid succession of new methods to probe RNA-protein interactions. Since our manuscript was published there have been exciting developments in methodology to identify RBPs and their associated RBRs. Updated interactome capture methods have been developed that incorporate 5-ethynyluridine (5EU) into RNA (RICK [12]; CARIC [13]). Following UV crosslinking of proteins and RNA (with or without 4sU), the 5EU can be modified with ’click’ chemistry to add a biotin handle for streptavidin purification. Precipitated RBPs are then analyzed by high resolution mass spectrometry [12, 13]. CARIC identified 597 RBPs, including 167 that were not identified previously by mRNA interactome studies in human cells. Additionally, 186/597 RBPs were also identified in our RBR-ID dataset, suggesting that both methods are able to profile non-polyA RNA-protein interactions [13]. The RICK method identified 720 high-confidence RBPs, of which 295 were unique to the method. To enrich for non polyA RNA binding proteins, the authors depleted polyA RNAs from their extracts before precipitating the crosslinked RNA-protein complexes, identifying 205 RBPs that overlapped with their 295 RICK-unique RBP hits (69.2%). These 205 RBPs therefore are highly likely to bind to non-polyA RNA, and intriguingly when the authors examined what GO terms were enriched in this stringent subset the strongest hit was chromosome organization ($p = 4.18\times10^{-45}$, 74 out of 205), supporting our finding that many chromatin-related proteins exhibit RNA-binding behavior.

Phase-separation assisted identification of RBPs

Novel methods have also emerged to identify RBRs using the physicochemical properties of RNA crosslinked to peptides. Three recent methods take advantage of the the enrichment of RNA-peptide adducts at the interphase of a guanidinium thiocynate /phenol/chloroform (Trizol) mixture [14–16]. In these methods, SILAC-labeled cells are grown prior to UV-crosslinking of one isotope label (heavy or light), then samples are combined and purified via Trizol interface enrichment. These methods allow for purification of crosslinked RNA-protein hybrids and subsequent processing to examine both the RNA and protein components in parallel, such as one group who combined PAR-CLIP with interphase enrichment to greatly simplify the often frustrating method [16]. One of the interphase enrichment methods, protein-crosslinking and RNA extraction (XRNAX), combined interphase purification with silica-column cleanup to further enrich RNA-protein adducts prior to LC-MS/MS, identifying over 1200 RBPs, including 565 that were not identified in previous polyA interactome studies. By comparing the characteristics of the RBPs identified in non polyA versus polyA associated RBPs, the authors noted a strong bias in RRM-containing proteins towards the polyA cohort, while the histone-binding bromodomain was most enriched domain in the non-polyA, consistent with our previous RBR-ID data. Potentially more significant, the authors also found enrichment in the non polyA RBP set for several tripeptide motifs (e.g. GRG, GGG, GSG) found in intrinsically disordered proteins (IDPs) that are linked to liquid-liquid phase separation [15, 17]. These findings highlight the distinct characteristics of non-poly-A RBPs and again reinforce the link between RNA, chromatin regulation and nuclear condensates. Future research should seek to further uncover the properties and function of nuclear RBPs in chromatin regulation. Another similar method, orthogonal organic phase separation (OOPS), sequenced the RNA present in in the Trizol interface of UV-crosslinked extracts, finding consistent loss of read coverage inside 3' UTRs and established RBP binding sites when compared to un-crosslinked samples, suggesting this method provides a way to profile RBP-binding sites across the transcriptome [14]. Additionally, the use of SILAC labeling enables exploration of changes in RNA-protein interactions in response to different biological conditions. The authors take advantage of this to examine changes in RBP behavior upon cell cycle arrest, finding increased abundance of metabolic enzymes in the interphase fraction, showing that RBP activity of these proteins increases in response to stress. These approaches can now be used to examine changes in RBP behavior in response to other variables, such as cell differentiation state or viral infection.

Precise identification of RNA-binding sites

Recent progress has also improved the resolution of mapping RNA-protein interactions. Prior attempts to directly identify crosslinked peptide-RNA adducts using mass spectrometry have performed poorly, due to the considerable challenge of matching nucleotides of many varying masses to a peptide spectra [15, 18]. One promising recent method addresses this issue by substituting inconsistent RNase digestion of UV-crosslinked extracts with hydrofluoric acid treatment to efficiently digest oligonucleotides down to a consistent and minimal moiety to facilitate identification by mass spectrometry [19]. This treatment enabled the use of only a single variable uridine modification in the LC-MS/MS database search, greatly simplifying the analysis and increasing identification efficacy. This approach, called RBS-ID, identified 1970 RNA-binding sites that fell within 642 proteins, a remarkable improvement over previous attempts, which found a maximum of 281 RNA-binding residues [19]. Combining the RBS-ID approach with targeted analyses of chromatin-related complexes could reveal functional residues involved in RNA binding. The current outlook for research studying RNA-protein interactions is positive, as methods to study them are rapidly maturing alongside an emerging understanding of the large scope of RNA function in chromatin biology.

5.1.2 Identification of an RNA regulatory site on PRC2

In chapter 3, I present data detailing our efforts to understand RNA-binding activity of the PRC2 complex. Building on the method we developed in chapter 2, I modified the approach with by incorporating SILAC for improved quantification and used an immunoprecipitation step to enrich for PRC2 to improve coverage relative to our previous proteome-wide dataset. Using this targeted variant of RBR-ID we identified RBRs across all PRC2 subunits identified, demonstrating that RNA-binding is a property shared by all PRC2 variants. Several PRC2 RBRs identified in previously annotated RNA-binding regions [20–22], while others were previously unknown. One

RBR in particular was located on EED directly adjacent to the stimulatory recognition motif (SRM) of PRC2 [23], raising the possibility of cross-talk between peptide-mediated stimulation of PRC2 and RNA inhibition [24]. RBDmap [25] of recombinant PRC2 provided confirmation of the SRM-adjacent RBR on EED. We showed that addition of the H3K27me3 stimulatory peptide could relieve RNA inhibition of PRC2 KMT activity on both histone and non-histone substrates. This showed that RNA inhibition of PRC2 could occur through a mechanism separate from competition of DNA or nucleosome binding [26]. We further showed that addition of a JARID2 stimulatory peptide–that binds the SRM–competed with G-quadruplex (G34) RNA for PRC2 binding, demonstrating that the regulatory center of PRC2 binds RNA in the absence of stimulatory peptides. Based on the antagonistic relationship between RNA and PRC2 binding to chromatin, our findings help understand how PRC2 can overcome inhibition from nascent RNA through stimulation from H3K27me3 or JARID2 K116me3 peptides.

5.1.3 Developments in understanding PRC2-RNA interactions and future directions

Recent evidence has emerged further dissecting the role of RNA in PRC2 regulation. Beltran *et al.* recently focused on the relationship between nascent RNA and PRC2 occupancy, having previously shown loss of RNA leads to PRC2 accumulation on chromatin [11, 27]. iCLIP [28] analysis of SUZ12 showed a strong enrichment for G-rich RNA sequences, consistent with previous findings showing PRC2 binding to G-rich sequences [29–31]. The authors then identified RNA sequences predicted to form G4 structures and found a surprising enrichment of these sequences near the first intron-exon junction of coding transcripts. When they compared these predicted G4 motifs with their iCLIP data they observed a strong overlap with SUZ12 iCLIP peaks, providing in vivo evidence to support the in vitro observations of PRC2 binding to G-quadruplex structures [27]. Recruitment of G4-containing sgRNAs using dCas9 to *Polycomb* target genes led to loss of PRC2 and H3K27me3 levels at the locus, demonstrating that G4 RNA can evict PRC2 from genes. These findings demonstrate how PRC2 responds to transcriptional activity by leaving chromatin and binding to G4 structures on nascent RNA, remaining poised for future silencing. In the presence of methylated JARID2, PRC2 is able to overcome G4 inhibition and place H3K27me3 de novo. While previous work has shown that PRC2 responds to transcriptional status [32], this is one of the first studies to examine the timing of PRC2 regulation. Given that bulk population measurements of various PRC2 subunits show them having largely overlapping occupancy on chromatin [33], time-resolved studies of specific polycomb target loci may help reveal the functions of different PRC2 subunits. Another recent study found that a diminished RNA-binding EZH2 mutant led to aberrant cardiomyocyte differentiation in induced human pluripotent stem cells, suggesting that RNA-PRC2 interactions are involved in cell differentation [34]. Future work should help us understand how PRC2-RNA interactions regulate PRC2 activity and localization in pluripotent and differentiated cell types.

Our targeted RBR-ID analysis of PRC2 identified multiple regions of the complex in binding RNA, including several that were not captured by our *in vitro* RBDmap analysis of PRC2 crosslinked with G4-RNA ligands. This discrepancy points either to technical differences between the two methods, or to the possibility that there are other non-G4 RNA ligands that bind specific sites on PRC2 *in vivo*. Most notably, targeted RBR-ID identified an RBR on EZH2 that was originally described by Kaneko *et al.* (EZH2 aa 342-368) [20]. Removal of this region does not affect PRC2 affinity for G4 RNA *in vitro* [22], yet I have found that EZH2 mutants lacking aa 342-368 show the greatest loss of RNA binding as measured by PAR-CLIP (Figure 5.1 A). This supports the notion that PRC2 binds other RNAs in addition to G4-containing transcripts, and may indicate multiple modes of PRC2-RNA regulation. Systematic analysis of different PRC2 mutants could determine if the various RBRs on PRC2 have distinct RNA ligands, and if so, could determine if the multiple RBRs have distinct functions that might explain the contradictory observations related to RNA-mediated PRC2 recruitment [35].

In addition to EZH2 I was also able to validate an RBR on the PRC2.2 subunit AEBP2. This subunit is known to stimulate PRC2 activity [36] and mediate PRC2 binding to H2AK119Ub [37]. Targeted RBR-ID revealed an RBR inside the 3 tandem zinc finger (ZF) domains of AEBP2 that corresponded closely to one identified in our proteome data (Figure 5.1). RNA pull-down experiments using recombinant protein fragments precipitated a fragment spanning the RBR (Figure 5.1 C), and deletion of the AEBP2 RBR led to a reduction of RNA-binding activity as measured via PAR-CLIP, further validating this region as being involved in binding RNA *in vivo*. To better understand the role of the various AEBP2 ZFs in RNA-binding and PRC2 regulation, I generated

AEBP2 KO mESCs and then attempted rescue experiments with various mutants lacking one or more ZF domains (Figure 5.1 E,F). RNA-seq of the rescue samples showed that while the wild-type AEBP2 was able to partially recapitulate wild-type expression patterns, all the various ZF mutants showed much weaker correlation with wild-type cells. While this indicates the AEBP2 zinc fingers are important functionally for maintaining polycomb function, it does not directly indicate RNA involvement, as recently published work from Eva Nogales has shown the AEBP2 directly binds to H2AK119Ub nucleosomes via its zinc fingers [37]. In order to specifically examine the role of RNA in AEBP2 function more refined separation-of-function mutants would be necessary, potentially using updated approaches as RBS-ID to make pinpointed mutations (see above) [19]. Intriguingly, recent work cataloguing the protein interactors of the *FIRRE* lncRNA identified PRC2 subunits AEBP2 and JARID2, supporting the possibility that PRC2 may recognize specific RNAs through RBRs distinct from those involved in G4-RNA recognition [38]. Systematic high-throughput CLIP-based sequencing of core and various accessory subunits of PRC2 can test this hypothesis, identifying whether any RNAs interact specifically with accessory subunits.

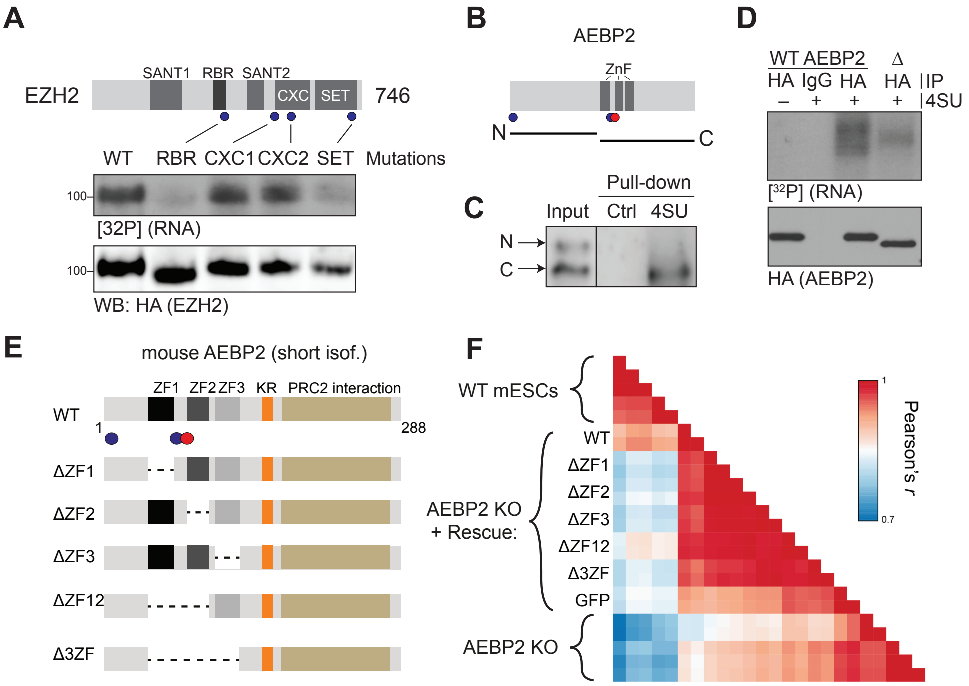


Figure 5.1 Validation of RBRs in PRC2 subunits AEBP2 and EZH2

(A) Localization of proteome-wide (red) and targeted (blue) RBR-ID hits on the domains of EZH2. PAR-CLIP of EZH2 mutants in HEK293 cells with deletions spanning the targeted RBR-ID hits. The "RBR" mutant deletion spans 342–368 [20] (B) Domains of AEBP2 displayed with RBR-ID hits shown as in (A). Recombinant fragments used in (C) are indicated. (C) *In vitro* crosslink and RNA-mediated pull-down of N-terminal and C-terminal fragments. (D) PAR-CLIP in HEK293 overexpressing HA-tagged WT AEBP2 or a mutant lacking the second zinc finger (Δ). RNA signal (top) and protein loading control (bottom) are shown. (E) Overview of AEBP2 mutants generated deleting indicated zinc finger domain(s). (F) AEBP2 Mutants shown in (E) were transfected into AEBP2 KO mESCs alongside GFP control. RNA was collected for RNA-seq analysis and analysed. Correlation heatmap is shown for 822 differentially expressed genes identified in AEBP2 KO cells.

5.1.4 A novel SARS-CoV-2 testing method

In Chapter 4, I present a set of experiments detailing the development of a novel testing method for SARS-CoV-2, based on a combination of barcoded RT-LAMP and deep sequencing. This method, referred to as COV-ID has been developed specifically to support large-scale testing. The method is inexpensive and uses unpurified saliva, a validated diagnostic analyte [39] that does not require invasive collection methods and minimizes sample processing steps and reagent demands. Our preliminary evaluation of the properties of this method revealed a limit of detection of 5–10 virions of SARS-CoV-2 per μL in contrived saliva samples and at least 300 virions /μL in saliva collected from patients in a clinical setting. The latter concentration was the lowest retrieved in the available clinical samples. It is therefore possible that the limit of detection for COV-ID in clinical samples may be lower than such threshold. The sensitivity demonstrated here exceeds the standard recommended for screening and surveillance tests of approximately 100 virions/μL [40, 41].

While most COV-ID data presented was performed using unpurified saliva samples, we also present some preliminary proof-of principle experiments that show that this method can be applied to dried saliva samples collected on filter paper. The possibility of detecting SARS-CoV-2 in dried saliva samples on a paper substrate could dramatically simplify sample collection and remove another barrier to testing. In our preliminary experiments we showed that COV-ID using paper substrates has a sensitivity of at least 50 virions /μL, demonstrating the viability of this testing approach.

5.1.5 Outlook for future SARS-CoV-2 testing initiatives

In the development of COV-ID we have had challenges managing the acute sensitivity of the RT-LAMP method and preventing contamination. This is a constant risk to any diagnostic laboratory but especially for those using RT-LAMP, since contaminating amplicons can self-amplify at room temperature without primers and proceed much more rapidly than PCR. Efforts to mitigate the risk of contamination would strengthen COV-ID and make it a more robust testing solution. For example, preparing all reagents in separate aliquots in an entirely different lab, and then storing them separately from any reaction products. Another outstanding question is how to multiplex RT-LAMP effectively. We were able to multiplex 3 targets in a single reaction, SARS-CoV-2, STATH, and influenza, but we observed inconsistent amplification for influenza. Other attempts at multiplexing RT-LAMP have faced similar issues, suggesting that competing amplification reactions can inhibit each other or expend all available nucleotides or enzyme [42]. To mitigate this issue we tried reducing the concentrations of BIP/FIP primers in multiplexed reactions, but further optimization may yield more consistent data. Also given the rapid emergence of several new mutant strains of SARS-CoV-2 [43], effective tests should be able to detect multiple variants to avoid false negative results. While sequencing can identify variants that occur within the amplicon, variants that impair LAMP primer binding sites may not be amplified efficiently. Therefore if additional amplicons can be multiplexed efficiently, then inclusion of other viral amplicons would be an effective way to simultaneously detect all viral variants in circulation as they emerge..

Regularly performed population-level screening tests are recognized as important elements of a pandemic response plan. Such population surveillance has a remarkable impact not only for clinical care, as it will allow the early identification of suspected cases and improve individual care, but also for public health. This strategy in fact can facilitate the identification of high-risk population pockets that may require specific public health measures. Furthermore, because vaccination campaigns have recently started [44], ongoing testing will help monitor the vaccination efficacy and guide further vaccine development.

The implementation of large-scale testing presents technical and economic challenges. As a result, although the number of tests performed in The United states has markedly increased since the first outbreak, the relative ratio of positive to total tests performed has only slightly increased. Although such numbers may be affected by the reporting method used (multiple testing of same individuals, retrospective reports, lack of follow-up data) and may not reflect a real-time assessment of the cases, they overall indicate that a large fraction of the supposedly asymptomatic population may still not have access to testing.

5.1.6 Final Remarks

In conclusion, in the above chapters I show how methods to profile RNA-binding proteins can enable more effective interrogation of chromatin regulation, and use these methods to reveal a new mode of regulation of the epigenetic regulator PRC2. Following that work I switched focus to development of a SARS-CoV-2 diagnostic method due to the unprecedented challenge that the COVID-19 pandemic has presented in terms of meeting sufficient testing capacity, without which public health measures are much less effective.

5.2 References

1. Valegard, K., Murray, J. B., Stockley, P. G., Stonehouse, N. J. & Liljas, L. Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature* **371,** 623–6 (1994).
2. Rahnamoun, H. *et al.* RNAs interact with BRD4 to promote enhanced chromatin engagement and transcription activation. *Nat Struct Mol Biol* **25,** 687–697 (2018).
3. Shen, Q. *et al.* Tet2 promotes pathogen infection-induced myelopoiesis through mRNA oxidation. *Nature* **554,** 123–127 (2018).
4. RNA Interactions Are Essential for CTCF-Mediated Genome Organization. *Mol Cell* **76,** 412–422 e5 (2019).
5. Hansen, A. S. *et al.* Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF. *Mol Cell* **76,** 395–411 e13 (2019).
6. Delatte, B. *et al.* RNA biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* **351,** 282–5 (2016).
7. Guallar, D. *et al.* RNA-dependent chromatin targeting of TET2 for endogenous retrovirus control in pluripotent stem cells. *Nat Genet* **50,** 443–451 (2018).
8. He, C. *et al.* TET2 chemically modifies tRNAs and regulates tRNA fragment levels. *Nat Struct Mol Biol* **28,** 62–70 (2021).
9. Sarma, K. *et al.* ATRX directs binding of PRC2 to Xist RNA and Polycomb targets. *Cell* **159,** 869–883 (2014).
10. Ren, W. *et al.* Disruption of ATRX-RNA interactions uncovers roles in ATRX localization and PRC2 function. *Nat Commun* **11,** 2219 (2020).
11. Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res* **26,** 896–907 (2016).
12. Bao, X. *et al.* Capturing the interactome of newly transcribed RNA. *Nat Methods* **15,** 213–220 (2018).

13. Huang, R., Han, M., Meng, L. & Chen, X. Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. *Proc Natl Acad Sci U S A* **115,** E3879–E3887 (2018).

14. Queiroz, R. M. L. *et al.* Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat Biotechnol* **37,** 169–178 (2019).

15. Trendel, J. *et al.* The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell* **176,** 391–403 e19 (2019).

16. Urdaneta, E. C. *et al.* Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nat Commun* **10,** 990 (2019).

17. Brangwynne, C. P., Tompa, P. & Pappu, R. V. Polymer physics of intracellular phase transitions. *Nature Physics* **11,** 899–904 (2015).

18. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods* **11,** 1064–70 (2014).

19. Bae, J. W., Kwon, S. C., Na, Y., Kim, V. N. & Kim, J. S. Chemical RNA digestion enables robust RNA-binding site mapping at single amino acid resolution. *Nat Struct Mol Biol* **27,** 678–682 (2020).

20. Kaneko, S. *et al.* Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. *Genes and Development* **24,** 2615–2620 (2010).

21. Kasinath, V. *et al.* Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science* **359,** 940–944 (2018).

22. Long, Y. *et al.* Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *eLife* **6** (2017).

23. Jiao, L. & Liu, X. Structural basis of histone H3K27 trimethylation by an active polycomb repressive complex 2. *Science* **350,** aac4383 (2015).

24. Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* **171,** 34–57 (2017).

25. Castello, A. *et al.* Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell* **63,** 696–710 (2016).

26. Wang, X. *et al.* Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. *Nat Struct Mol Biol* **24,** 1028–1038 (2017).

27. Beltran, M. *et al.* G-tract RNA removes Polycomb repressive complex 2 from genes. *Nat Struct Mol Biol* **26,** 899–909 (2019).

28. Konig, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17,** 909–15 (2010).

29. Hendrickson, D., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biology* **17** (2016).

30. Kaneko, S., Son, J., Shen, S. S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20,** 1258–64 (2013).

31. Wang, X. *et al.* Targeting of Polycomb Repressive Complex 2 to RNA by Short Repeats of Consecutive Guanines. *Mol Cell* **65,** 1056–1067 e5 (2017).

32. Riising, E. M. *et al.* Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. *Mol Cell* **55,** 347–60 (2014).

33. Hojfeldt, J. W. *et al.* Non-core Subunits of the PRC2 Complex Are Collectively Required for Its Target-Site Specificity. *Mol Cell* **76,** 423–436 e3 (2019).

34. Long, Y. *et al.* RNA is essential for PRC2 chromatin occupancy and function in human pluripotent stem cells. *Nat Genet* **52,** 931–938 (2020).

35. Davidovich, C. & Cech, T. R. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA* **21,** 2007–22 (2015).

36. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298,** 1039–43 (2002).

37. Kasinath, V. *et al.* JARID2 and AEBP2 regulate PRC2 in the presence of H2AK119ub1 and other histone modifications. *Science* **371** (2021).

38. Graindorge, A. *et al.* In-cell identification and measurement of RNA-protein interactions. *Nat Commun* **10,** 5317 (2019).

39. Vogels, C. B. F. *et al.* SalivaDirect: A Simplified and Flexible Platform to Enhance SARS-CoV-2 Testing Capacity. *Med (N Y)* (2020).

40. Larremore, D. B. *et al.* Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Sci Adv* **7** (2021).

41. Mina, M. J., Parker, R. & Larremore, D. B. Rethinking Covid-19 Test Sensitivity - A Strategy for Containment. *N Engl J Med* **383,** e120 (2020).

42. Zhang, Y. & Tanner, N. A. Development of Multiplexed RT-LAMP for Detection of SARS-CoV-2 and Influenza Viral RNA. *medRxiv,* 2020.10.26.20219972 (2020).

43. Baric, R. S. Emergence of a Highly Fit SARS-CoV-2 Variant. *N Engl J Med* **383,** 2684–2686 (2020).

44. Williams, T. C. & Burgers, W. A. SARS-CoV-2 evolution and vaccines: cause for concern? *Lancet Respir Med* (2021).