HyPR-MS for Multiplexed and Splice Variant-Specific Discovery of RNA-Protein Interactomes

By

Rachel A. Knoener

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2017

Date of final oral examination: 12/04/2017

The dissertation is approved by the following members of the Final Oral Committee:

Lloyd M. Smith, Professor, Chemistry
Nathan Sherer, Associate Professor, Molecular Virology and Oncology
Helen Blackwell, Professor, Chemistry
Ying Ge, Associate Professor, Chemistry, Cell and Regenerative Biology

ACKNOWLEDGEMENTS

I would like to thank my advisor, Prof. Lloyd Smith, for providing mentorship and for modeling the attributes of a genuine scientist. I greatly appreciate his mentoring philosophies that allowed me to grow scientifically and personally.

The staff scientists in the Smith Lab, Dr. Michael Shortreed, Dr. Brian Frey, and Dr. Mark Scalf, have been tremendous assets to my graduate career. Each provided insights personally and professionally that could not be replicated in another environment. I thank them for their time, patience, and encouragement through the entire process. In particular, I thank Brian Frey for his thoughtful input on the questions I asked, the multitude of information I didn't even know I needed, and his personal guidance and understanding. I also thank the graduate students, past and present, who have celebrated and commiserated with me throughout the years. I especially thank Gloria Sheynkman for her enthusiasm for science, pep-talks during difficult times, and her mentorship.

I have been very fortunate to have a collaborative relationship with Prof. Nathan Sherer and Dr. Jordan Becker in the Molecular Virology and Oncology department. I thank them for sharing with me the amazing world of viruses and allowing me to apply my science to a meaningful study. I also want to thank my committee for their input and guidance through this process and for taking the time to participate in my dissertation and defense.

Finally, I thank my family: my parents for their encouragement and understanding, my sisters for their enthusiasm and support, and particularly my partner M'bark Baddouh. M'bark provides me daily encouragement and understanding. We traveled this journey together, sharing the triumphs and hardships. Words are not suffice to express my gratitude for your example, your strength, and your faith in me. Thank you.

AUTHOR CONTRIBUTIONS

---

1. **Knoener, R. A.**; Becker, J. T.; Scalf, M.; Sherer, N. M.; Smith, L. M., Elucidating the in vivo interactome of HIV-1 RNA by hybridization capture and mass spectrometry. *Scientific reports* **2017,** *7* (1), 16965.

Author Contribution: HyPR-MS design and implementation, RT-qPCR analysis, protein sample preparation, mass spectrometry data collection, siRNA knockdown with viral infection and fluorescence detection, data analysis and writing of manuscript were conducted by R.A.K.

2. **Knoener, R.A**.; Spiniello, M.; Steinbrink, M.I.; Yang, B.; Cesnik, A.J.; Buxton, K.E.; Scalf, M.; Jarrard, D.F.; Smith, L.M., HyPR-MS for Multiplexed Discovery of the MALAT1, NEAT1, and NORAD lncRNA Protein Interactomes. Submitted to *Molecular Cell*.

Author Contribution: Design and optimization of technology parameters, protein data analysis, writing of manuscript were conducted by R.A.K.

3. **Knoener, R.A**.; Becker, J.T.; Scalf, M.; Sherer, N.M.; Smith, L.M., Discovery of Differential Protein Interactomes of HIV-1 Splice Variants. Validation of findings in progress.

Author Contribution: R.A.K. conceived of splice variant specific isolation of HIV-1 RNAs and designed and optimized the strategy for achieving purification. R.A.K performed all capture experiments, RT-qPCR experiments, mass spectrometric experiments and all data analysis. R.A.K. has started and will continue the siRNA knockdown and fluorescence microscopy analysis of proteins identified as HIV binders in this study.

4. Shortreed, M. R.; Frey, B. L.; Scalf, M.; **Knoener, R. A**.; Cesnik, A. J.; Smith, L. M., Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements. *J Proteome Res* **2016,** *15* (4), 1213-1221

Author Contribution: R.A.K. established protocol for preparation of protein samples for intact mass analysis by mass spectrometry and prepared the intact protein samples for analysis in this study.

5. Cesnik, A.J.; Shortreed, M.R.; Schaffer, L.V.; **Knoener, R.A.**; Frey, B.L.; Scalf, M.; Solntsev, S.K.; Dai, Y.; Gasch, A.P.; Smith, L.M., Proteoform Suite: software for constructing, quantifying, and visualizing proteoform families. *J Proteome Res* 2017, In Press.

Author Contribution: R.A.K. contributed to the design of the normal versus salt-stressed yeast biological experiment with NeuCode lysine for proteoform quantification. R.A.K. established cell culture protocols and optimized protein isolation and separation of the yeast samples for intact mass analysis. R.A.K. prepared all samples for mass spectrometric analysis. R.A.K. contributed data interpretation.

TABLE OF CONTENTS

ABSTRACT

Gene expression is largely regulated by the interactions of RNA and protein molecules. These interactions are temporal, dynamic, and dependent on physiological conditions. Therefore, to gain a comprehensive understanding of how the expression of any particular gene is regulated requires powerful tools for discovery of specific RNA-protein interactions in space and time. Presented here is a technology, Hybridization Purification of RNA-protein complexes and Mass Spectrometry (HyPR-MS), for identification of the *in vivo* protein interactome of specific RNAs and the use of that technology for multiplexed and splice variant-specific isolation of RNAs. Chapter 2 introduces the technology for identification of the HIV-1 genomic RNA interactome and establishes efficacy of the strategy. Chapter 3 expands that technology for multiplexed isolation of three regulatory lncRNAs, MALAT1, NEAT1, and NORAD, and discovery of their protein interactomes in PC3 cells. Finally, Chapter 4 harnesses the attributes of HyPR-MS to isolate the principle HIV-1 splice variants from infected cells and discover their differing protein interactomes for deeper comprehension of the role of RNA-protein interactions in HIV replication.

1. INTRODUCTION

1.1 Regulation of Gene Expression. Gene expression is the conversion of information stored in a gene into functional molecules to exhibit a particular phenotype. Broadly and traditionally speaking, DNA is transcribed into messenger RNA (mRNA), and mRNA is translated into the functional product, proteins. However, a deeper investigation into this simplified central dogma of molecular biology reveals complex mechanisms that regulate many intermediary steps to promote or repress a gene's expression. Furthermore, research has revealed a class of RNAs called non-coding RNAs (ncRNAs); ncRNAs do not code for functional proteins but in themselves exhibit regulatory functions[1]. Additionally, protein products from the same gene can display variable post-translational modifications (PTMs) that affect the protein's function but are not explicitly dictated by the gene from which the protein is coded[2]. These findings demand that the complex molecular interactions that convert the information stored in DNA into a phenotype be deciphered in order to understand how the expression of any given phenotype is regulated. Not least among these molecular interactions are RNA-protein interactions.

1.2 RNA-Protein Interactions. RNA-protein interactions influence many steps of gene expression. Proteins can bind to ncRNA to regulate transcription of an adjacent gene[3]. Proteins perform and regulate the splicing of nascent transcripts to express different isoforms of the same gene[4]. Sequestration of RNA into nuclear speckles via protein interactions prevents its splicing or export into the cytoplasm[5]. Sequestration into RNA-protein complexes called P-bodies or stress granules in the cytoplasm can induce the RNA's

degradation or postpone its translation[6]. Similarly, ncRNAs can bind to proteins to prevent specific proteins from acting on their known RNA targets[7]. Proteins interact with RNAs to translocate the RNA to the cellular location in which its protein product functions[8]. These interactions, and many others not mentioned, are transient, sometimes concurrent, and variable depending on the requirements of the cell under different conditions or stresses[9,10]. Often, the dysregulation of these interactions can result in disease. Many neurodegenerative diseases have been linked to anomalies in RNA-binding proteins[11] and recently RNA-modifications such as N6-methyladenosine have been shown to influence RNA-protein interactions and show evidence of affecting tumor development[12]. Furthermore, many viruses, including human immunodeficiency virus 1 (HIV-1), have RNA genomes and depend on specific RNA-protein interactions to achieve replication[13]. The expansive role that RNA-protein interactions play in gene regulation and disease begs for the elucidation of the actors that maintain or disrupt this regulation. However, since each gene has different expression requirements depending on cellular conditions, the respective RNA must be investigated individually to determine the pertinent interactions for its regulation or dysregulation.

1.3 HIV / AIDS. Human immunodeficiency virus 1 (HIV-1) is a retrovirus that causes Acquired Immune Deficiency Syndrome (AIDS)[14]. In 2015, over 36 million people globally were living with HIV, 54% of which were without treatment. Approximately 1 million people died of AIDS-related complications in 2015[15]. Though disconcerting, these numbers show improvement in the number of AIDS related deaths (down 43% since 2003) due, in large part, to an expansion of HIV treatment using antiretroviral therapy (ART)[15]. Though ART is able to suppress viral production in patients with HIV, there still remain latent HIV

virus in some CD4+ T-cells that are resistant to the treatment and can become active if treatment is stopped[16]. Thus, though effective at limiting a patient's viral count, ART does not cure a person of HIV or of immunological damage caused by HIV infection. A permanent cure for HIV has been elusive, thus continued research into HIV and its process of replication is crucial to achieve global eradication of this devastating disease[17].

1.4 HIV Replication. The HIV virus has an approximately nine-kilobase RNA genome that is spliced into 3 major classes of splice variants and codes for nine proteins or polyproteins (Figure 1.1)[18].

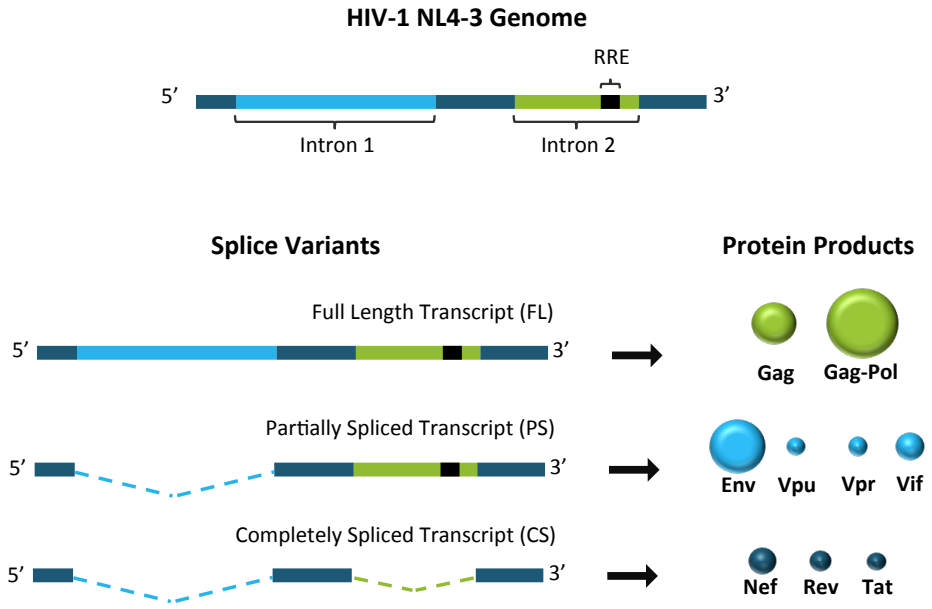


Figure 1.1: HIV genome and gene products. The nine-kilobase HIV genome is transcribed into the full-length primary transcript and spliced into three different classes of splice variants. The splice variants are then translated into nine viral proteins and polyproteins.

The virus utilizes these nine proteins and the protease products of the polyproteins, in addition to many proteins native to the host cell, to propagate and spread to other cells through a process called replication (Figure 1.2)[19,20]. HIV replication begins when the virus fuses with the host cell surface. The viral capsid enters the host cell and its protein and RNA contents are released upon capsid degradation. The HIV-RNA genome is then reverse transcribed into DNA using the virus-encoded protein, reverse transcriptase. This DNA is transported into the nucleus and the viral integrase enzyme integrates the viral genome into the host genome. With assistance from the viral regulatory factor Tat, the HIV genome is transcribed into the full-length transcirpt using the host cellular machinery. Early in this stage the transcript is generally completely spliced, all introns are removed, and exported into the cytoplasm via traditional RNA nuclear export pathways. The completely spliced transcript is then translated into the regulatory proteins Tat, Rev, and Nef. Nef is transported to the cell membrane where it facilitates viral infection and the Tat and Rev protein products are transported into the nucleus where they regulate the efficient transcription and splicing of the HIV transcripts. The presence of Rev in the nucleus improves the efficiency of production of the alternatively spliced variants of HIV, the unspliced transcript and the partially spliced transcript. Rev mediates the export of these two variants into the cytoplasm via a non-canonical pathway so as to avoid complete splicing or degradation in the nucleus. Once in the cytoplasm the partially spliced and full length transcripts are translated into the accessory proteins (Vif, Vpu, Vpr, and Env) and polyproteins (Gag and Gag-Pol), respecitively. The Gag protein then facilitates the packaging of full-length transcripts into virions for propagation of the virus to new cells where the process is repeated (Figure 1.2)[20]. The interaction of virus-encoded proteins with

the HIV RNA is integral to the success and efficiency of HIV replication. However, the virus has also evolved to highjack many host proteins to ensure efficient replication. These HIV RNA-host protein interactions are numerous and many of significance are described in the subsequent chapters.

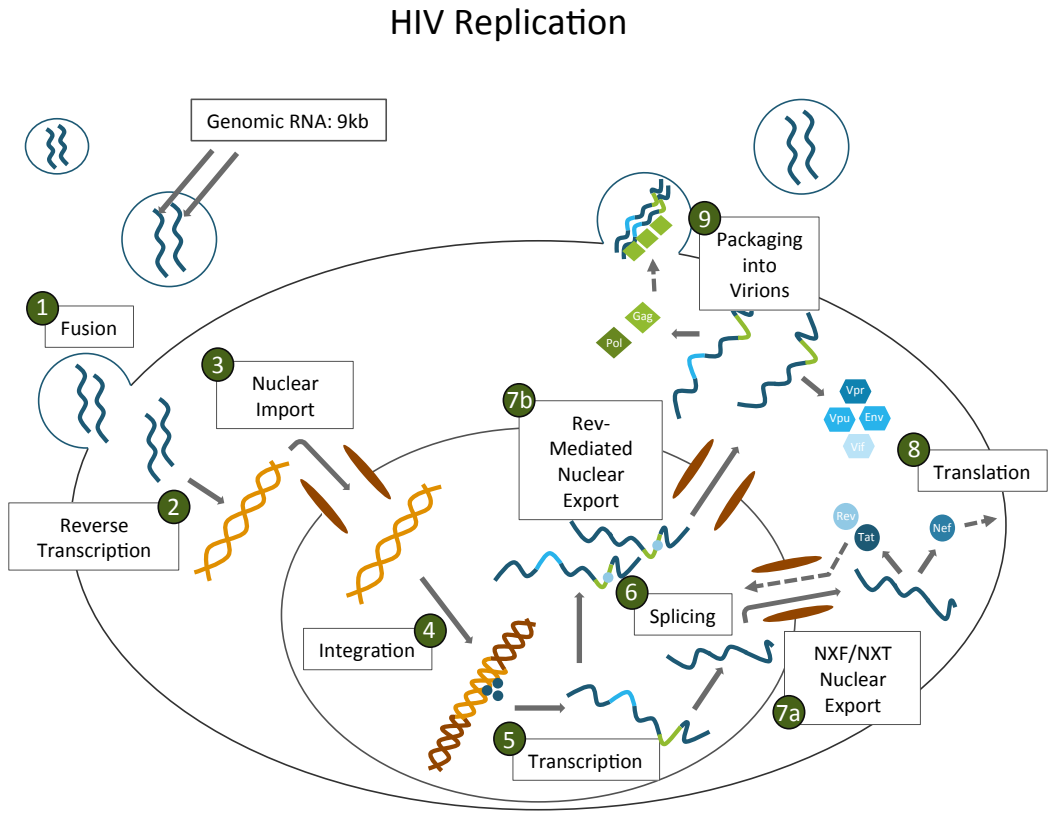


Figure 1.2: Graphic for HIV-1 replication. HIV replication begins with fusion of the virus to the outer cell membrane. Once the RNA genome is released into the cell cytoplasm it is reverse transcribed into DNA, imported into the nucleus, and integrated into the host's DNA. The integrated HIV-DNA is transcribed into full length RNA transcripts using the cell's machinery then either completely spliced and exported into the cytoplasm by canonical cellular pathways or the transcript is partially spliced or remains unspliced and is exported via the Rev-mediated nuclear export pathway. Once in the cytoplasm, all splice variants are translated into their respective proteins, however, the full-length transcript can also be packaged as the genome for nascent virions.

1.5 Long Non-Coding RNAs (lncRNAs)

The pervasion of high-throughput RNA-sequencing techniques has illuminated the prevalence of non-coding RNAs in the transcriptome of many organisms, particularly humans. This plethora of data has shown that only 1-2% of the genome produces protein products while 1000s of genomic sites produce RNAs with little capacity to code for protein[21]. These RNAs are referred to as non-coding RNAs (ncRNAs), and those >200 nucleotides in length are called long non-coding RNAs, or lncRNAs. Though many have been sequenced and identified, the mechanisms by which specific lncRNAs regulate gene expression are largely unknown. In general, however, several mechanisms have been proposed (Figure 1.3)[22,23]. Some lncRNAs regulate gene expression at the transcriptional level either by recruiting proteins to or evicting proteins from the chromatin. For example, lncRNAs can associate with the DNA and recruit histone modifiers to the nucleosomes to add modifications that promote or prevent transcription of a particular gene. In contrast, a lncRNA can act as a decoy and evict certain proteins from the gene promoter site to affect gene expression. lncRNAs can also act at the post-transcriptional level by acting as a scaffold to bring together proteins that function in concert, by acting as a sponge to attenuate the function of other proteins, or by recruiting proteins that stabilize RNA molecules and prevent their degradation[22,23].

It has been shown that lncRNA expression can be organism, tissue, and cell type specific. Furthermore, their mechanisms of action can be different depending on physiological conditions [24]. Many lncRNAs have been identified as markers for specific cancers; lncRNA MALAT1 has been shown to have various regulatory functions in lung cancer metastasis and

many other cancers[25], lncRNA SChLAP1 influences aggressive prostate cancer via chromatin remodeling[26], and lncRNA FAL1 has been shown to be relevant in ovarian cancer[27].

Extensive evidence has shown the impact of lncRNAs on the regulation or dysregulation of gene expression, therefore, the specific mechanisms by which lncRNAs act must be determined. This is no small task as the physiological context in which an individual lncRNA exists greatly impacts its expression and mode of action. Furthermore, the primary sequence of a lncRNA, alone, is not sufficient to predict its mode of action. We propose that a strategy for understanding lncRNA gene regulation is to identify the proteins with which a lncRNA interacts so as to infer its functional mechanism. Since lncRNAs often interact with protein partners in order to affect gene expression, the discovery of specific lncRNA-protein interactions is necessary and techniques for doing so are paramount.

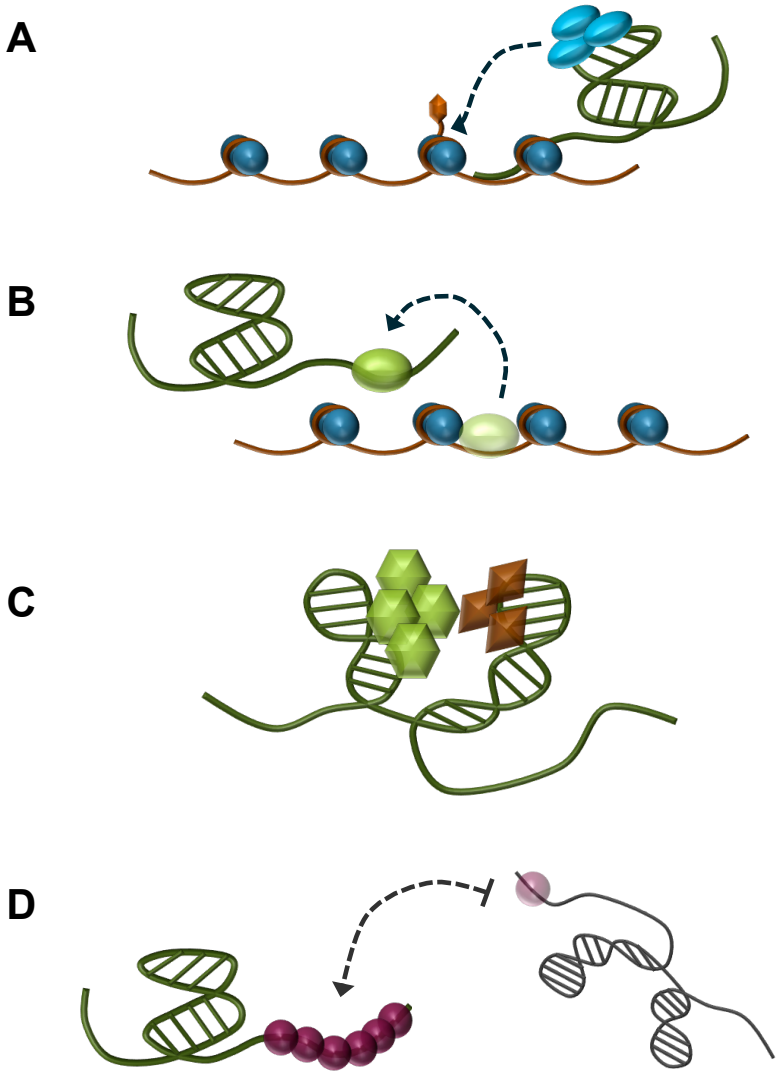


Figure 1.3: Examples of lncRNA regulatory mechanisms. A. lncRNAs can recruit proteins, such as histone modifiers, to the chromatin. The proteins can modify the nearby histones and promote or inhibit transcription of specific genes. B. lncRNAs can act as decoys for proteins that would otherwise bind to the chromatin and affect transcription. C. Some lncRNAs serve as scaffolds for proteins that work together to form a specific function. D. lncRNA can act as a "sponge" to bind proteins that would otherwise bind to other RNA targets to affect their regulation.

1.6 Technologies to Study RNA-Protein Interactions. The existing technologies for identifying specific RNA-protein interactions can be categorized as protein-centric or RNA-centric. Protein-centric technologies target a specific, known protein of interest and identify the associated RNAs; whereas, RNA-centric technologies target a specific RNA of interest and identify interacting proteins. Protein-centric technologies such as RNA-Immunoprecipitation (RIP) utilize an affinity agent (usually an antibody) specific to a known protein of interest to isolate it from cell lysate then identify the associated RNAs by RNA-sequencing, RT-PCR, or micro-arrays[28,29]. This strategy has elucidated many specific RNA-protein interactions; however, is limited to protein targets with suitable antibodies and requires previous knowledge of the identity of the protein targets. RNA-centric strategies can also leverage immunoprecipitation to identify RNA-protein interactions. One such strategy modifies an RNA-of-interest with specific nucleotide sequences that serve as MS2 bacteriophage-binding sites. These RNAs are incubated with and bound by flag-tagged MS2 coat proteins, the complexes are immunoprecipitated with anti-flag antibodies, and the proteins associated with the modified-RNA are identified by mass spectrometry. This variety of technology has been successfully implemented in various contexts [30,31]; however, it requires manipulation of the RNA sequence and therefore cannot be used to investigate unaltered cell-lines or tissues and could cause artifactual interactions.

Other RNA-centric approaches have been developed to study *in vivo* RNA-protein interactions, including: ChIRP-MS (comprehensive identification of RNA binding proteins by mass spectrometry)[32], CHART-MS (capture hybridization analysis of RNA targets and mass spectrometry)[33], and RAP-MS (RNA antisense purification and mass spectrometry)[34]. These strategies identify proteins that interact with a specific RNA by exposure of cells to an

RNA-protein cross-linker, capture of the target RNA by specific biotinylated capture oligonucleotides and, after elution from streptavidin-coated magnetic-beads, identification of the RNA-associated proteome by mass spectrometry[35]. These technologies are advantageous because they allow for the investigation of RNA-protein interactions under specific spatiotemporal and biological contexts and can target the native RNA of a particular cell line. These strategies have, thus far, identified the proteins that interact with lncRNAs including MALAT1, NEAT1, and Xist, and have enabled the discovery of RNA-protein interactions.

1.7 Development of Hybridization Purification of RNA-Protein Complexes and Mass Spectrometry (HyPR-MS). In the work presented here we demonstrate the development of a versatile strategy for discovery of *in vivo* RNA-protein interactomes called Hybridization Purification of RNA-protein complexes and Mass Spectrometry (HyPR-MS). HyPR-MS utilizes a biotinylated capture oligonucleotide that is complementary to a specific RNA of interest and streptavidin-coated magnetic beads to isolate RNA-protein complexes from cross-linked cells. The purified complexes are then processed for identification of the protein interactors via mass spectrometry. Though similar to the analogous technologies discussed above, HyPR-MS addresses some of the challenges that these technologies face by enabling multiplexing, to purify several specific RNAs from one cell culture, and the purification of different splice variants of the same primary RNA transcript. In doing so, HyPR-MS not only permits a deeper investigation into the complexities of RNA-protein interactions in gene expression, it permits more cost effective and labor efficient experiments. We demonstrate in Chapter 2 the development and application of HyPR-MS to discover the *in vivo* protein interactome of the HIV-1 full length, genomic RNA. To show

the relevance of our findings we performed functional validation on several of the identified HIV RNA-protein interactors to gain understanding of their effect on HIV replication. Next, the novel characteristics of HyPR-MS and its multiplexing capabilities are demonstrated. In Chapter 3, we investigate the protein interactomes of three lncRNAs of differing known general functions and discover their protein interactomes in a prostate cancer cell line. Finally, in Chapter 3, we adapted HyPR-MS to capture and isolate the three classes of HIV splice variants, the unspliced, partially spliced, and completely spliced transcripts.

1.8 References:

1. Eddy, S. R. (2001) Non-coding RNA genes and the modern RNA world, *Nature reviews. Genetics 2*, 919-929.

2. Seo, J., and Lee, K. J. (2004) Post-translational modifications and their biological functions: Proteomic analysis and systematic approaches, *J Biochem Mol Biol 37*, 35-44.

3. Bonasio, R., and Shiekhattar, R. (2014) Regulation of Transcription by Long Noncoding RNAs, *Annu Rev Genet 48*, 433-455.

4. Wang, Z., and Burge, C. B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code, *Rna 14*, 802-813.

5. Spector, D. L., and Lamond, A. I. (2011) Nuclear speckles, *Cold Spring Harbor perspectives in biology 3*.

6. Decker, C. J., and Parker, R. (2012) P-Bodies and Stress Granules: Possible Roles in the Control of Translation and mRNA Degradation, *Cold Spring Harbor perspectives in biology 4*.

7. Lee, S., Kopp, F., Chang, T. C., Sataluri, A., Chen, B., Sivakumar, S., Yu, H., Xie, Y., and Mendell, J. T. (2016) Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins, *Cell 164*, 69-80.

8. Blower, M. D. (2013) Molecular Insights into Intracellular RNA Localization, *Int Rev Cel Mol Bio 302*, 1-39.

9. Moore, M. J. (2005) From birth to death: the complex lives of eukaryotic mRNAs, *Science 309*, 1514-1518.

10. Jankowsky, E., and Harris, M. E. (2015) Specificity and nonspecificity in RNA-protein interactions, *Nature reviews. Molecular cell biology 16*, 533-544.

11. Conlon, E. G., and Manley, J. L. (2017) RNA-binding proteins in neurodegeneration: mechanisms in aggregate, *Genes & development 31*, 1509-1528.

12. Batista, P. J. (2017) The RNA Modification N6-methyladenosine and Its Implications in Human Disease, *Genomics, proteomics & bioinformatics 15*, 154-163.

13. Rosen, C. A. (1991) Regulation of HIV gene expression by RNA-protein interactions, *Trends in genetics : TIG 7*, 9-14.

14. Barresinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axlerblin, C., Vezinetbrun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983) Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune-Deficiency Syndrome (Aids), *Science 220*, 868-871.

15. (2016) Global AIDS Update - 2016, UN Joint Programme on HIV/AIDS (UNAIDS).

16. Brooks, D. G., and Zack, J. A. (2002) Effect of latent human immunodeficiency virus infection on cell surface phenotype, *J Virol 76*, 1673-1681.

17. Barouch, D. H., and Deeks, S. G. (2014) Immunologic strategies for HIV-1 remission and eradication, *Science 345*, 169-174.

18. Purcell, D. F., and Martin, M. A. (1993) Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity, *J Virol 67*, 6365-6378.

19. Mohammadi, P., Desfarges, S., Bartha, I., Joos, B., Zangger, N., Munoz, M., Gunthard, H. F., Beerenwinkel, N., Telenti, A., and Ciuffi, A. (2013) 24 Hours in the Life of HIV-1 in a T Cell Line, *Plos Pathog 9*.

20. Nisole, S., and Saib, A. (2004) Early steps of retrovirus replicative cycle, *Retrovirology 1*, 9.

21. Mattick, J. S. (2004) RNA regulation: a new genetics?, *Nature Reviews Genetics 5*, 316-323.

22. Wang, K. C., and Chang, H. Y. (2011) Molecular Mechanisms of Long Noncoding RNAs, *Molecular cell 43*, 904-914.

23. Schmitz, S. U., Grote, P., and Herrmann, B. G. (2016) Mechanisms of long noncoding RNA function in development and disease, *Cell Mol Life Sci 73*, 2491-2509.

24. Gutschner, T., and Diederichs, S. (2012) The hallmarks of cancer: a long non-coding RNA point of view, *RNA biology 9*, 703-719.

25. Gutschner, T., Hammerle, M., and Diederichs, S. (2013) MALAT1 -- a paradigm for long noncoding RNA function in cancer, *Journal of molecular medicine 91*, 791-801.

26. Prensner, J. R., Iyer, M. K., Sahu, A., Asangani, I. A., Cao, Q., Patel, L., Vergara, I. A., Davicioni, E., Erho, N., Ghadessi, M., Jenkins, R. B., Triche, T. J., Malik, R., Bedenis, R., McGregor, N., Ma, T., Chen, W., Han, S., Jing, X., Cao, X., Wang, X., Chandler, B., Yan, W.,

Siddiqui, J., Kunju, L. P., Dhanasekaran, S. M., Pienta, K. J., Feng, F. Y., and Chinnaiyan, A. M. (2013) The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex, *Nature genetics 45*, 1392-1398.

27. Hu, X., Feng, Y., Zhang, D., Zhao, S. D., Hu, Z., Greshock, J., Zhang, Y., Yang, L., Zhong, X., Wang, L. P., Jean, S., Li, C., Huang, Q., Katsaros, D., Montone, K. T., Tanyi, J. L., Lu, Y., Boyd, J., Nathanson, K. L., Li, H., Mills, G. B., and Zhang, L. (2014) A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer, *Cancer cell 26*, 344-357.

28. Niranjanakumari, S., Lasda, E., Brazas, R., and Garcia-Blanco, M. A. (2002) Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo, *Methods 26*, 182-190.

29. Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M., and Lee, J. T. (2010) Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq, *Molecular cell 40*, 939-953.

30. Lee, N., Pimienta, G., and Steitz, J. A. (2012) AUF1/hnRNP D is a novel protein partner of the EBER1 noncoding RNA of Epstein-Barr virus, *Rna-a Publication of the Rna Society 18*, 2073-2082.

31. Kula, A., Guerra, J., Knezevich, A., Kleva, D., Myers, M. P., and Marcello, A. (2011) Characterization of the HIV-1 RNA associated proteome identifies Matrin 3 as a nuclear cofactor of Rev function, *Retrovirology 8*.

32. Chu, C., Zhang, Q. C., da Rocha, S. T., Flynn, R. A., Bharadwaj, M., Calabrese, J. M., Magnuson, T., Heard, E., and Chang, H. Y. (2015) Systematic discovery of Xist RNA binding proteins, *Cell 161*, 404-416.

33. West, J. A., Davis, C. P., Sunwoo, H., Simon, M. D., Sadreyev, R. I., Wang, P. I., Tolstorukov, M. Y., and Kingston, R. E. (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites, *Molecular cell 55*, 791-802.

34. McHugh, C. A., Chen, C. K., Chow, A., Surka, C. F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., Sweredoski, M. J., Shishkin, A. A., Su, J., Lander, E. S., Hess, S., Plath, K., and Guttman, M. (2015) The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3, *Nature 521*, 232-236.

35. Simon, M. D. (2016) Insight into lncRNA biology using hybridization capture analyses, *Bba-Gene Regul Mech 1859*, 121-127.

2. ELUCIDATING THE *IN VIVO* INTERACTOME OF HIV-1 RNA BY HYBRIDIZATION CAPTURE AND MASS SPECTROMETRY.

This chapter is accepted for publication in *Scientific Reports*.

Rachel A. Knoener[1], Jordan T. Becker[2], Mark Scalf[1], Nathan M. Sherer[2], Lloyd M. Smith[*1,3]

[1]Department of Chemistry, University of Wisconsin, Madison, Wisconsin, United States
[2]McArdle Laboratory for Cancer Research and Institute for Molecular Virology, University of Wisconsin, Madison, Wisconsin, United States
[3]Genome Center of Wisconsin, University of Wisconsin, Madison, Wisconsin 53706, United States

2.1 ABSTRACT:

HIV-1 replication requires myriad interactions between cellular proteins and the viral unspliced RNA. These interactions are important in archetypal RNA processes such as transcription and translation as well as for more specialized functions including alternative splicing and packaging of unspliced genomic RNA into virions. We present here a hybridization capture strategy for purification of unspliced full-length HIV RNA-protein complexes preserved *in vivo* by formaldehyde crosslinking, and coupled with mass spectrometry to identify HIV RNA-protein interactors in HIV-1 infected cells. One hundred eighty-nine proteins were identified to interact with unspliced HIV RNA including Rev and Gag/Gag-Pol, 24 host proteins previously shown to bind segments of HIV RNA, and over 90 proteins previously shown to impact HIV replication. Further analysis using siRNA knockdown techniques against several of these proteins revealed significant changes to HIV expression. These results demonstrate the utility of the approach for the discovery of host proteins involved in HIV replication. Additionally, because this strategy only requires

availability of 30 nucleotides of the HIV-RNA for hybridization with a capture oligonucleotide, it is readily applicable to any HIV system of interest regardless of cell type, HIV-1 virus strain, or experimental perturbation.

2.2 INTRODUCTION:

Interactions between the viral unspliced RNAs of human immunodeficiency virus type 1 (HIV) and both virus- and host-encoded proteins play critical roles in viral replication. HIV produces over 40 splice variants from the primary HIV transcript[36,37], utilizing multiple splice donor and splice acceptor sites. These splice variants fall into three categories: unspliced (~9 kb), partially spliced (~4 kb), and completely spliced (~2kb). While the completely spliced transcripts pass through the classical nuclear export pathway and are translated into protein products in the cytoplasm, sequence motifs present in the partially spliced and unspliced transcripts lead to the use of alternative cellular pathways to effect both translation and virion assembly[38]. In order to understand these processes, it is critical to identify the proteins that interact with the various forms of the HIV RNAs.

The unspliced HIV transcript is of particular interest because it not only codes for the Gag and Gag-Pol polyproteins, it also serves as the viral genome of the virions that propagate infection. Typically, cellular transcripts that contain introns are either spliced or degraded while in the nucleus. The HIV unspliced transcript, however, interacts with host and viral proteins to prevent its splicing and degradation in the nucleus, promote its nuclear export, and enable its packaging into virions. For example, SR family proteins (proteins with stretches of serine (S) and arginine (R) amino acids) and heterogeneous nuclear

ribonucleoproteins (hnRNPs) interact with the splicing enhancers and splicing silencers coded in the transcript to permit alternative splicing[39,40]. Viral protein Rev binds to the Rev Response Element (RRE) of the transcript. This RNA-protein interaction helps the unspliced transcript evade surveillance mechanisms for degradation[39,41] and facilitates its export by the alternative Rev/CRM1-dependent nuclear export pathway[42]. Furthermore, Rev has been shown to recruit host proteins of diverse functions, including; nuclear matrix protein MATR3[31], nonsense-mediated decay protein UPF1[43], and RNA helicases DDX1 and MOV10[44,45]. Following nuclear export, the unspliced transcript can be incorporated into a variety of ribonucleoprotein complexes (RNPs). The binding of host factors such as DDX3 to the highly structured 5'-end of the unspliced transcript has been shown to facilitate translation[46,47]. Comparably, the binding of Staufen1, a double-stranded RNA binding protein, along with HIV-Gag, has been shown to influence the incorporation of the genome into virions[48]. HIV-Gag binds unspliced HIV RNA with sequence specificity via its nucleocapsid domain, leading to encapsidation of the HIV RNA[49]. As the eventual destiny of the unspliced HIV transcript is directly dependent on such interactions with both viral and host proteins, elucidating the identities and roles of these proteins is crucial to understanding the HIV replication process.

Extensive research, using various techniques, has revealed numerous host factors implicated in HIV replication. At least four genomic screens using RNA knockdown techniques to discover critical players in the HIV replication process in cells have been reported[50-53]. The RNA knockdown strategy utilizes small interfering RNAs (siRNAs) or small hairpin RNAs (shRNAs) to knockdown individual genes in cell culture and the effect on HIV infection is monitored[54]. The success of each knockdown requires that the levels of the target transcript

and resultant protein are sufficiently decreased in the cells and that the protein product's function is not replicated by an alternative protein. Despite these limitations, hundreds of genes have been shown in this way to play important roles in HIV replication. However, based on a meta-analysis of three of the four genome-wide RNA knockdown studies, reproducibility of the results is limited and the mechanisms by which most of the proteins act are still not understood[54]. Affinity purification serves as an alternative strategy for identification of pertinent host factors. In this strategy, a protein of interest is tagged with a ligand (e.g. Strep, Flag, GFP), beads coupled to a suitable affinity agent for the ligand are used to capture the desired protein from a complex mixture, and the accompanying proteins are identified by mass spectrometry or other techniques[55-57]. A recent example of such a study reported the assembly of viral and host protein networks from the affinity purification of 18 HIV protein products followed by mass spectrometric analysis[56]. A protein affinity purification strategy has also been used to target engineered HIV RNAs in cell lysate and capture associated proteins. Two such studies captured segments of HIV RNA modified with multiple MS2 bacteriophage binding sites, followed by mass spectrometric protein identification[31,58]. Kula et al. transfected cells with an HIV-1 derived vector containing truncated HIV RNA sequences modified with 24 MS2 bacteriophage-binding sites. Flag-tagged MS2 coat protein added to the cell lysate binds the MS2 binding sites and the complex is immunoprecipitated using anti-flag antibodies. This strategy revealed many previously known mRNA binders and regulators and also indicated a novel function in HIV replication for MATR3[31]. Similarly, Marchand et al. incubated SLS2-A7 HIV-1 RNA segments containing three MS2 binding sites in HeLa cell lysate. The resulting HIV RNPs were then affinity purified using an MS2-maltose binding protein (MBP) fusion protein and amylose beads. This study resulted in the discovery of a novel HIV-1 splicing regulator,

hnRNPK. While this RNA-centric affinity purification scheme is valuable for the discovery of RNA-protein interactions, it is limited to use on MS2-containing constructs of HIV RNA, rather than on native HIV RNA sequences.

Presented here is a versatile strategy for probing *in vivo* HIV RNA-protein interactions we call HyPR-MS: Hybridization Purification of RNA-protein-complexes followed by Mass Spectrometry. HyPR-MS is based upon the specific capture by nucleic acid hybridization of viral RNAs that have been subjected to *in vivo* formaldehyde crosslinking, and mass spectrometric identification of associated proteins. Related strategies have been reported for interrogating gene-specific DNA-protein interactions (PICh: proteomics of isolated chromatin[59]; HyCCAPP: hybridization capture of chromatin associated proteins for proteomics [60,61]) and lncRNA-protein interactions (RAP-MS: RNA antisense purification and mass spectrometry, ChIRP-MS: comprehensive identification of RNA binding proteins by mass spectrometry, CHART-MS: capture hybridization analysis of RNA targets and mass spectrometry)[62-64]. Briefly, the HyPR-MS strategy involves formaldehyde crosslinking of HIV infected cells to preserve RNA-protein interactions. The lysate is then incubated with a biotinylated DNA capture oligonucleotide specifically complementary to the unspliced viral RNA. The capture oligonucleotide hybridizes to the target RNA-protein complexes and the hybrid is isolated from the lysate using streptavidin coated magnetic beads. Following stringent washes the RNA-protein complexes are released from the beads and the sample is analyzed by mass spectrometry to identify associated proteins (Fig. 2.1a).

HyPR-MS is employed here to identify *in vivo*, unspliced full-length HIV RNA protein interactors. The term "interactors" in this study encompasses proteins that may interact

directly with the HIV RNA or proteins that may interact with a protein intermediate that in turn interacts directly with the HIV RNA. Since proteins often function in complexes of multiple proteins, the inclusion of the indirect interactors better elucidates the HIV RNA interactome. The virus encoded RNA binding proteins Rev and Gag/Gag-Pol were among the 189 proteins identified. Additionally, 24 host proteins previously shown to bind to constructs of sequences of HIV RNA and over 90 shown to interact with the RNA-binding viral proteins were identified. This considerable overlap of previously identified HIV-related host factors supports the ability of HyPR-MS to correctly identify RNA-interacting proteins. Many putative novel HIV RNA interactors were also identified including over 25 annotated RNA-binding proteins not previously known to bind HIV RNA. The GO annotations of the proteins identified, as well as the known HIV-related functions of some of the proteins, cover a broad range of RNA-processing steps. Finally, we demonstrate functional effects of several of these proteins using siRNA knockdown techniques followed by fluorescence microscopy to evaluate HIV expression in cells. These findings provide evidence that the HyPR-MS strategy can capture unspliced HIV RNA at various stages of the RNA's lifespan, and reveal important, functional, *in vivo* HIV RNA protein interactors.

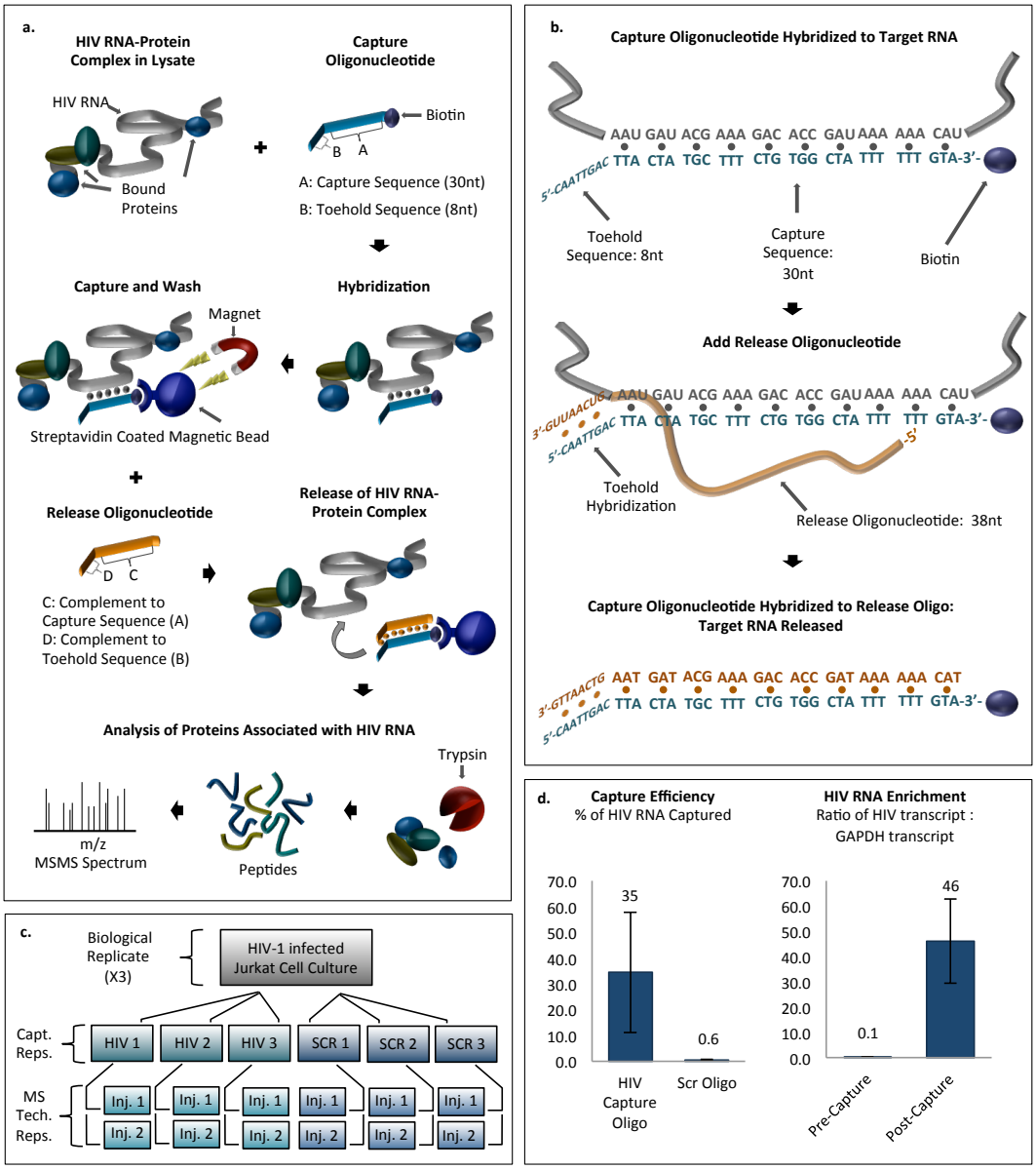


Figure 2.1: A. HyPR-MS uses hybridization capture for purification of HIV RNA-protein complexes from cell lysate for protein identification by mass spectrometry. B. A toehold-mediated capture and release strategy was implemented for sequence specific isolation of RNA-protein complexes. C. A total of three biological replicates were analyzed. Three capture replicates (Capt. Reps.) were conducted for each biological replicate and were each divided into two mass spectrometric analyses (MS Tech. Reps.). D. Capture efficiency and specificity for each capture was measured using a RT-qPCR assay specific to HIV RNA. RT-qPCR analysis of HIV RNA relative to GAPDH RNA in the capture and lysate samples measured enrichment of HIV RNA after capture. Error bars in graphs are standard deviations of measurements made for all three biological replicates.

2.3 RESULTS AND DISCUSSION:

2.3.1 HyPR-MS overview. Figure 2.1a shows a diagram of the HyPR-MS strategy. The strategy was applied here to HIV-infected Jurkat cells, a human T lymphocyte cell line that is widely used for the study of HIV infection[65]. Cells were grown in culture, infected with pseudotyped replication-deficient (Env-minus) NL4-3 HIV-1[66], and crosslinked with formaldehyde to stabilize protein-RNA interactions[67,68]. Cells are lysed with a detergent-containing buffer, briefly sonicated, and cellular debris removed by centrifugation. The cell lysate is then incubated with a biotinylated DNA capture oligonucleotide which contains a 30-nucleotide sequence specifically complementary to the unspliced viral RNA and an 8-nucleotide toehold sequence that is not complementary to the target RNA. The capture oligonucleotide hybridizes to the target RNA-protein complexes and the hybrid is isolated from the lysate using streptavidin-coated magnetic beads. Following stringent washes the RNA-protein complexes are released from the beads using a toehold release oligonucleotide[69]. This oligonucleotide is complementary to all 38 nucleotides of the capture oligonucleotide and is therefore thermodynamically favored to hybridize with it over the target RNA (which is only complementary to 30 nucleotides of the capture oligonucleotide), thereby releasing the HIV RNA-protein complex into solution (Fig. 2.1b). The sample is then trypsin digested and analyzed by mass spectrometry to identify associated proteins (Details in Supplementary Information and Methods).

2.3.2 Hybridization capture efficiency and specificity. RT-qPCR assays were employed to measure the yield and specificity of HIV RNA capture. Control experiments utilizing a scrambled capture oligonucleotide sequence provide a measure of hybridization specificity.

The scrambled oligonucleotide was designed to have the same number of nucleotides and approximately the same T_m as the capture oligonucleotide but without significant complementarity to the target transcript or any other transcript in the cells. The hybridization capture strategy was applied to three biological replicates of HIV infected Jurkat cells. A biological replicate is defined here as an independent cell culture and HIV infection followed by cross-linking and pelleting of the cells. Three capture oligonucleotide replicates and 3 scrambled oligonucleotide replicates were produced from each of the three biological replicates to give a total of 9 HIV-RNA oligonucleotide captures and 9 scrambled oligonucleotide captures. The captured and released material from each of these replicates was used mostly for two technical replicates by mass spectrometric analysis for protein identification (Fig. 2.1c), but a small aliquot (2% of total sample) was saved for verification by RT-qPCR of specificity and efficiency of capture.

Capture efficiency is defined here as a measure of the percentage of the total HIV RNA transcripts in the lysate aliquot that were hybridization captured and released from the beads. It was measured using the HIV RNA-specific qPCR assay to determine the amount of HIV RNA in the capture sample relative to the amount in the lysate prior to capture. The mean capture efficiency for the three biological replicates was approximately 35% (Fig. 2.1d). Capture specificity is defined here as the amount of HIV RNA in the specific oligonucleotide capture sample compared to that in the scrambled oligonucleotide capture sample. An average of 60-times more HIV RNA was obtained in the HIV capture samples than in the scrambled capture samples (Fig. 2.1d). We also evaluated the enrichment of the HIV RNA transcript relative to off-target transcripts. Using qPCR assays specific to HIV RNA and to GAPDH RNA (a ubiquitously expressed "housekeeping" gene[70]), we measured

the amount of each of these transcripts in both the HIV capture sample and the lysate prior to capture. The ratio of HIV:GAPDH, on average, in the capture sample was 46:1 whereas in the lysate sample the average ratio was 0.13:1. This gives an average enrichment factor of 350 for the three biological replicates (Fig. 2.1d). These results indicate that the capture procedure provides high specificity and efficiency of HIV transcript capture.

2.3.3 Mass spectrometry and statistical analysis of the data. The proteins from the HIV capture oligonucleotide and scrambled capture oligonucleotide samples were trypsin digested and analyzed by mass spectrometry. The resulting spectra were searched using MaxQuant[71,72] for protein identification and label-free relative quantitation[73] (details in Methods). The Perseus software program[74] was used to determine which proteins were significantly enriched in the HIV-capture compared to the scrambled-capture samples. In addition to a Student's T-test p-value, a "student's T-test test statistic" was calculated with the S_0=0.8. Proteins with a test statistic that pass a permutation-based 1% false discovery rate threshold are considered enriched in the HIV capture samples. The 189 proteins identified in this manner have a maximum p-value of 0.05 and a minimum fold change of 2.2.

2.3.4 Validation of HyPR-MS protein identifications. HyPR-MS was developed and utilized here in order to identify proteins that interact with unspliced HIV RNA, with a particular interest in those that play roles in HIV replication. Proteins found to interact fall into two major categories: those that have been previously identified as HIV RNA interactors, and those that have not. The former serve to support and validate the effectiveness of the approach, while the latter constitute new discoveries of possible

biological significance. To identify previously known interactors, we searched the HIV literature for proteins already shown to 1) interact directly with segments of HIV RNA, or 2) interact with viral proteins (Tat, Rev, Gag, Gag-Pol) that bind to HIV RNA. Two previous studies, Kula et al.[31] and Marchand et al.[58], were particularly valuable resources of known HIV RNA interactors. As described in the Introduction, both of these studies utilized engineered and truncated versions of HIV RNA to permit antibody-based affinity capture; this is in contrast to the HyPR-MS strategy employed here, in which the near-native full-length single-round infectious HIV RNA was captured by hybridization, with *in vivo* RNA-protein interactions stabilized by formaldehyde crosslinking. The study by Kula et al. identified 30 HIV RNA-interacting host proteins, while that of Marchand et al. yielded 42; 10 and 15 of the Kula and Marchand identifications (with 4 in common), respectively, were found in the present study as well. Additionally, three other proteins have been confirmed in three different studies to interact with HIV RNA[75-77]. In total, 24 host proteins, over 12% of those identified by HyPR-MS, had been previously identified as HIV RNA interactors using orthogonal isolation techniques.

In addition to the above host proteins, several virus-encoded proteins are known to interact with HIV-RNA: Tat, Rev, Gag, and Gag-Pol as well as Gag-Pol protease products integrase (IN) and reverse transcriptase (RT)[49,78-82]. These proteins enhance HIV-replication by interacting with HIV-RNA at multiple steps during the replication process. Polyproteins Gag and Gag-Pol are translated from the full length HIV-RNA and thus have a large portion of amino acid sequence in common in addition to sections of sequence specific to each polyprotein. HyPR-MS revealed peptides unique to proteins Rev and polyprotein Gag, as

well as several peptides that could be from Gag, Gag-Pol or their protein products providing further validation for the technology.

It has been shown that many RNA-binding proteins work in concert with other proteins to perform their functions. The use of formaldehyde as the crosslinking agent in HyPR-MS facilitates the identification of these co-factors because formaldehyde covalently links protein-protein interactors as well as RNA-protein interactors. To this point, it is satisfying that ABCE1 was identified in our study. ABCE1 interacts directly with Gag without an RNA intermediate[83]. The use of formaldehyde crosslinking preserves not just the protein - HIV RNA interaction (in this case Gag – HIV RNA), but also the broader protein-protein interactions (ABCE1 – Gag) of the HIV RNP. Extensive research has been conducted, using various biochemical techniques, to identify proteins that associate with viral-encoded RNA-binding proteins (RBPs) and this data has been compiled into various databases. We cross-referenced our protein list with the NCBI[84] and GPS-Prot[85] databases and found that 91 of the 189 proteins identified by HyPR-MS have been previously shown to interact with Tat, Rev, Gag, or Gag-Pol.

The analysis of the Gene Ontology (GO) annotations [86,87] for the 189 proteins further confirms the ability of HyPR-MS to identify RNA-associated proteins. Most significantly, 86 proteins are annotated to be RNA-binding. RNA-binding proteins with at least 4-fold enrichment in the HIV capture using HyPR-MS are presented in Table 2.1. Notably, many of these proteins have been previously shown to interact with HIV-RNA and/or to have functions in HIV replication. Over 90 proteins identified have one or more GO annotations representing involvement in various RNA-related processes. The representation of multiple

RNA processes in multiple cellular locations suggests that HyPR-MS is capable of capturing HIV-RNA protein complexes that form throughout the lifespan of the RNA.

The proteins identified by HyPR-MS to be unspliced HIV RNA interactors cover a broad range of cellular processes and functions. The list of 189 proteins was investigated for proteins that are known to affect HIV replication in general or at specific stages of the life cycle as well as for proteins that are novel HIV-RNA interactors. Figure 2.2 shows proteins that were identified in this study to interact with unspliced HIV RNA and were previously shown to affect HIV replication in general (ovals) or at specific stages of the replication cycle (rectangles). The proteins are grouped into RNA-process related categories based on each protein's known function in HIV replication (rectangles) or, if the specific function is not known, based on the protein's general GO annotated RNA related functions (ovals). Proteins that have previously been shown to interact with segments of HIV RNA, either *in vivo* or *in vitro*, are indicated with shaded ovals or rectangles. Additionally, lines connect proteins to the virus encoded RBPs (Tat, Rev, Gag, Gag-Pol) that they are known to interact with. This figure is not a comprehensive picture of all proteins identified by HyPR-MS nor of all proteins known to affect HIV replication; however, it provides a summary of perceived significant unspliced HIV RNA protein interactors identified in this study and their relevant interactions and functions. Below, we highlight some of these significant and potentially significant HIV RNA interactors.

Table 2.1: RNA-binding proteins identified to interact with HIV-RNA using HyPR-MS

Gene ID	Protein Name	Proposed Function in HIV Replication	Previous evidence of HIV-RNA binding?	Refs
Transcription / Regulation of Transcription				
FUBP1	Far upstream element-binding protein 1	Unknown	No	
HIST1H1B	Histone H1.5	Unknown	No	
HSPD1	60 kDa heat shock protein	Yes	No	74
ILF2	Interleukin enhancer-binding factor 2	Yes	Yes	24,74,113
NPM1	Nucleophosmin	Transcription	Yes	24,74,104
TRIM28	Transcription intermediary factor 1-beta	Early HIV Replication	Yes	24,74,91,128
Splicing / Splicesome				
DDX5	DEAD box protein 5	Nucleocytoplasmic Transport	Yes	24,56,74,127
EIF4A3	Eukaryotic initiation factor 4A-III	Yes	No	74
FUS	RNA-binding protein FUS	Yes	Yes	8,55
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	Splicing, Nucleocytoplasmic Transport	Yes	24,62,63,74,117
HNRNPC	Heterogeneous nuclear ribonucleoprotein C	Yes	No	74
HNRNPD	Heterogeneous nuclear ribonucleoprotein D	Splicing, Nucleocytoplasmic Transport	Yes	24,64,74
HNRNPK	Heterogeneous nuclear ribonucleoprotein K	Splicing	Yes	24,66,74
HNRNPR	Heterogeneous nuclear ribonucleoprotein R	Yes	Yes	24,66,74
LUC7L2	Putative RNA-binding protein Luc7-like 2	Yes	No	74
PCBP2	Poly(rC)-binding protein 2	Yes	Yes	43
RBM8A	RNA-binding protein 8A	Unknown	No	
SYNCRIP	Heterogeneous nuclear ribonucleoprotein Q	Splicing	Yes	24,74
Translation / Regulation of Translation				
CALR	Calreticulin	Unknown	No	
EEF1A1	Elongation factor 1-alpha 1	Early HIV Replication, Translation, RNA Packaging	Yes	24,74,112,123
EEF1D	Elongation factor 1-delta	Early HIV Replication, Translation	No	74,123
ELAVL1	ELAV-like protein 1 (HuR)	Translation	No	71-73
LRPPRC	Leucine-rich PPR motif-containing protein	Yes	No	74,121
NCL	Nucleolin	RNA Packaging/Budding	Yes	24,74,80
YBX3	Y-box-binding protein 3	Unknown	Yes	8
P-body / Stress Granule				
CAPRIN1	Caprin-1	Yes	Yes	8,74
MOV10	Putative helicase MOV-10	Nucleocytoplasmic Transport, RNA Packaging	Yes	8,11,74-77
PSMA6	Proteasome subunit alpha type-6	Transcription	No	129
YBX1	Nuclease-sensitive element-binding protein 1	Yes	Yes	41,74,93
Structural Molecule / Cytoskeleton				
ANXA1	Annexin A1	Unknown	No	
FLNB	Filamin-B	Unknown	No	
Gag/Gag-Pol	Gag/Gag-Pol polyprotein	Yes	No	15,46
RPL10A	60S ribosomal protein L10a	Unknown	No	
RPL11	60S ribosomal protein L11	Unknown	No	
RPL23	60S ribosomal protein L23	Unknown	No	
RPL6	60S ribosomal protein L6	Yes	No	74
RPSA	40S ribosomal protein SA	Unknown	No	
SPTBN1	Spectrin beta chain	Unknown	No	19,106
TUBA1B	Tubulin alpha-1B chain	Yes	Yes	8,17,106
Other				
ARF1	ADP-ribosylation factor 1	Yes	No	103
CCT6A	T-complex protein 1 subunit zeta	Unknown	No	
HSP90B1	Endoplasmin	Unknown	Yes	24
HSPA9	Stress-70 protein	Yes	No	74
HSPE1	10kDa heat shock protein	Unknown	No	
MDH2	Malate dehydrogenase	Unknown	No	
P4HB	Protein disulfide-isomerase	Unknown	No	
PDIA3	Protein disulfide-isomerase A3	Yes	No	120
PDIA4	Protein disulfide-isomerase A4	Unknown	No	
PEBP1	Phosphatidylethanolamine-binding protein	Unknown	No	
PPIA	Peptidyl-prolyl cis-trans isomerase A	Early HIV Replication	No	94,97
PRDX1	Peroxiredoxin-1	Unknown	No	
PRMT1	Protein arginine N-methyltransferase 1	Unknown	No	
STIP1	Stress-induced-phosphoprotein 1	Unknown	No	
YWHAE	14-3-3 protein epsilon	Unknown	No	
YWHAG	14-3-3 protein gamma	Unknown	No	
YWHAZ	14-3-3 protein zeta/delta	Unknown	No	

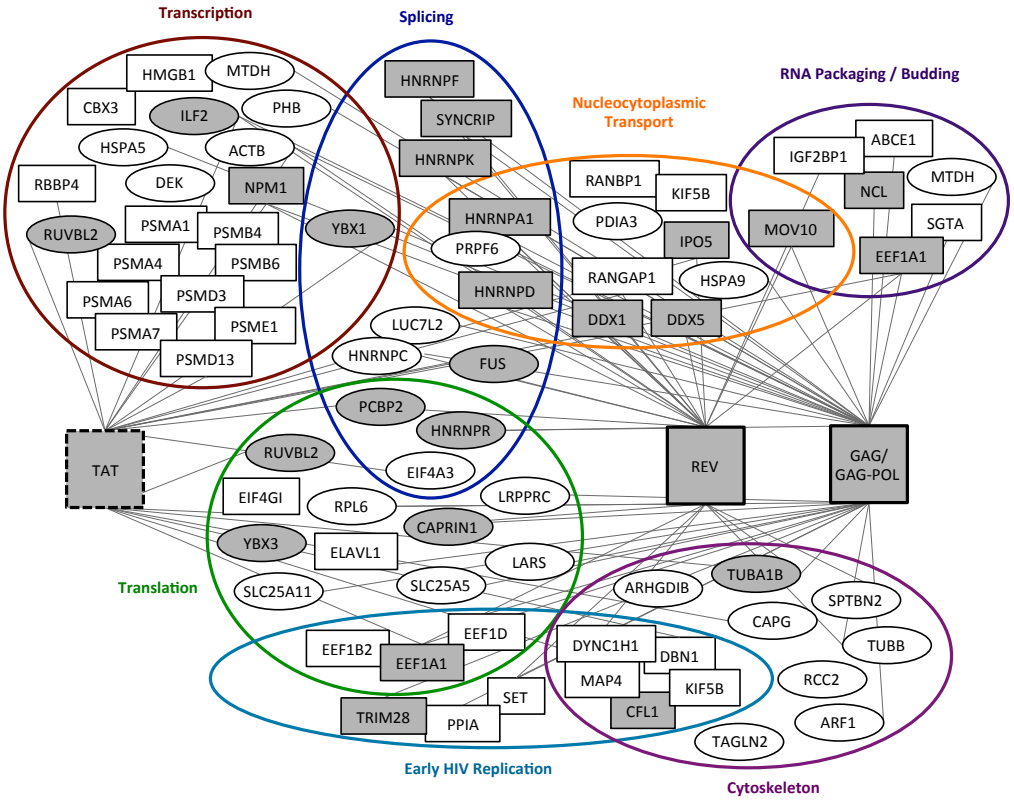


Figure 2.2: Unspliced HIV RNA protein interactor map. All proteins indicated were identified by HyPR-MS to interact with HIV unspliced RNA (with the exception of Tat). Proteins in rectangles have previously been shown to affect specific stages of HIV replication and are grouped accordingly (colored circles). Proteins in ovals have previously been shown to affect HIV replication but the stage of replication was not determined. These proteins are therefore grouped according to their RNA-process related GO annotation. Proteins shaded in grey have been previously identified to interact with segments of HIV RNA; unshaded proteins are novel unspliced HIV RNA interactors. Virus encoded HIV RNA binders are indicated in the large rectangles; and a line connects them to cellular proteins previously shown to interact with these viral proteins[10,11,16-19,24,41-43,54-57,62-64,66,71-77,80-82,85,91-127].

2.3.5 HyPR-MS identifies proteins involved in nuclear RNA processes. The unspliced HIV RNA, because it retains its introns, requires mechanisms to prevent its degradation and splicing prior to nuclear export. HyPR-MS identified the viral protein Rev, which is a nuclear export factor that binds to the RRE region of unspliced and partially spliced HIV transcripts and expedites their export into the cytoplasm[41]. In addition to Rev, HyPR-MS also identified RANBP1 and RANGAP1, host factors that are known to be involved in the Rev-dependent nuclear export of RRE-containing RNAs. RANBP1 and RANGAP1 work together to hydrolyze RanGTP to release Rev and its RNA cargo from the export machinery and into the cytoplasm[88]. Interestingly, an *in vitro* study found that PRPF6, another protein also identified by HyPR-MS, interacts with Rev bound to an RRE sequence in an RanGTP dependent manner, suggesting that PRPF6 has a role in HIV RNA nuclear export[89]. The discovery of PRPF6 as an *in vivo* unspliced HIV RNA interactor supports this hypothesis.

Several DEAD-box helicases have been shown to interact with Rev and to impact HIV replication[90]. Two such helicases, DDX1 and DDX5, were identified here as HIV RNA interactors. Previously, co-immunoprecipitation assays from HEK293 cells overexpressing Rev and DDX1 and from 293FT cell extracts expressing Rev and HA-tagged DDX5, confirmed the interaction of Rev with DDX1 and DDX5, respectively[90,91]. Additional studies investigated the impact of these interactions on HIV replication. Both helicases were shown to enhance Rev function using a Rev-dependent luciferase-based reporter plasmid (pDM628) transfected into 293FT cells; DDX5 was suggested to work synergistically with DDX3[90]. Also, siRNA knockdown of DDX1 resulted in restricted Rev function[90,91] while another study using siRNA knockdown proposed that DDX5 antagonizes the formation of DDX17 homodimers that are required for HIV replication[89]. The study presented here gives

a "snap-shot" of *in vivo* RNA-protein and protein-protein interactions, showing that both DDX1 and DDX5 interact either directly or indirectly with HIV RNA. The detection of these interactions was accomplished without modification or overexpression of RNA or protein in the cellular system, providing confidence that the interactions are not artifactual in nature.

Notably, one group of RNA-binding proteins is well represented in the HyPR-MS data: heterogeneous nuclear ribonucleoproteins (hnRNPs). hnRNPs play roles in multiple processes including alternative splicing[92], mRNA stability[93,94], and nuclear export[95]. HyPR-MS identified nine hnRNPs; eight (hnRNP A1, D, E2 (PCBP2), F, K, P2 (FUS), Q (SYNCRIP), and R) have previously been shown to bind to select segments of HIV RNA[31,58,77] and to have roles in HIV replication[96-100]. However, hnRNP C, to our knowledge, has not previously been shown to bind HIV-RNA but has been shown to bind to mRNAs with N^6-methyladenosine (m^6A) modifications to affect gene expression[101]. Interestingly, it has been shown recently that HIV-1 RNAs have multiple clusters of m^6A at the 3'-end and that their presence affects steady state levels of the viral mRNA expression[102]. We further evaluate the effect of HNRNP C on HIV replication using siRNA knock-down techniques described later.

2.3.6 HyPR-MS identifies proteins involved in cytoplasmic RNA processes. Once it has been exported from the nucleus to the cytoplasm, the unspliced HIV RNA is either translated into Gag and Gag-Pol polyproteins or incorporated into progeny virions. Like cellular mRNAs, viral RNA is assembled into either polysomes for translation or different ribonucleoprotein (RNP) complexes to regulate its translation. Processing bodies (P-bodies)

contain mRNAs that are destined for decay whereas stress granules (SGs) contain mRNAs that are stalled in translation initiation[103]. Unspliced HIV RNA can also be assembled into RNPs that are thought to transition the unspliced RNA into progeny virions[48]. These RNPs contain Staufen1 protein and unspliced HIV RNA (thus called SHRNPs) and appear to have a distinct function from other Staufen1 containing RNPs such as stress granules[48].

Many of the proteins identified by HyPR-MS are multifunctional and are known components of one or more of the above mentioned types of RNPs. Twenty-two proteins with GO annotations in translation, including translation initiation and elongation factors, tRNA ligases, and ribosomal proteins, were identified. Several components of stress granules were also identified, including translation initiation factors, small ribosomal subunit proteins and other translational regulators such as HuR (ELAVL1) and YBX3. HuR has multiple functions in RNA metabolism including regulation of alternative splicing, mRNA stability, and translation[104]. Specific to HIV RNA processing, HuR has been reported to regulate HIV RNA translation[105]; other studies debate whether HuR also influences reverse transcription[106,107]. Like HuR, YBX3 is annotated as a regulator of mRNA stability and translation; however, despite having been shown to bind to HIV RNA constructs[31] and native unspliced HIV RNA here, its role in HIV replication is still unknown.

Proteins associated with SHRNPs were also well represented in the data presented here. The composition of SHRNPs has been explored using tandem affinity purification and mass spectrometry of Staufen1 containing RNPs from cell lines expressing TAP-Staufen1 and HIV-1[108]. SHRNPs require unspliced HIV RNA for formation and incorporate Gag plus more than 200 cellular proteins[48,108]; 39 of which were also identified by HyPR-MS

(Supplementary Table S6). Staufen1, itself, was not identified, which could be due to various experimental factors involving crosslinking, protein abundance, or ionization efficiencies during mass spectrometric analysis. MOV10, a component of P-bodies, is a protein implicated in multiple functions affecting HIV replication including viral infectivity, virion production, reverse transcription, Gag proteolytic processing, and nuclear export[45,109-111]. It has been previously reported that MOV10 binds to engineered constructs of segments of HIV RNA[31], and now, HyPR-MS shows that MOV10 associates with native unspliced HIV RNA during HIV replication. Additional SHRNP protein components previously shown to affect HIV virion production were also identified by HyPR-MS. ABCE1 has been identified as a necessary factor for HIV-1 capsid formation[112,113] and NCL has been shown to complex with HIV RNA and Gag to enhance virion release[114]. Conversely, IGF2BP1 has been demonstrated to complex with Gag to block the formation of HIV-1 particles[115]. Protein Lyric (MTDH), though not identified as part of the SHRNP, has been shown to interact with Gag multimers and to be incorporated into HIV-1 virions, suggesting Lyric has an impact on virion infectivity[116]. This hypothesis is further informed by the finding here that Lyric is an unspliced HIV RNA interactor.

2.3.7 siRNA knockdown and fluorescence microscopy: Eight proteins identified using HyPR-MS (MOV10, YBX3, IGF2BP1, RPSA, HNRNPC, HSPD1, RBBP4, YWHAH) were evaluated for their influence on HIV replication using siRNA knockdown techniques. These proteins were selected for functional assessment for at least one of the following reasons: 1) the p-value comparing HIV capture samples and the scrambled capture samples was <0.05, 2) the total intensities of the peptides for the protein in the HIV capture sample was in the top quartile, or 3) the abundance of the protein in the HIV capture sample was statistically

higher than that in a polyA-RNA capture sample. Details regarding target selection for siRNA knockdown are presented in the Supplementary Information.

The effect of the knockdown of each target protein was evaluated by, first, transfecting 293T cells (stably-expressing YFP-APOBEC3G) in multi-well culture plates with siRNA pools specific for each target protein. In addition to cell cultures transfected with siRNAs against the individual target proteins, two controls were conducted. A sample was transfected with siGFP to knock down the stably expressed cellular YFP-APOBEC3G protein and demonstrate successful transfection conditions. Transfection with non-targeting siRNAs (siCTRL) was used as a negative control. After 48-hours the cells were transfected a second time with the same siRNAs followed by infection with a two-color fluorescent HIV-1 reporter virus (E-R-Gag-3xCFP mCherry/nef; Fig. 2.3a). This virus is similar to that used for infection of Jurkat cells for HyPR-MS analysis, with two notable changes: the Gag open-reading frame (ORF) contains three copies, in tandem, of CFP reporter and the nef ORF contains an mCherry reporter. Forty-eight hours post-infection the cells were fixed and the nuclei stained (4',6-diamidino-2-phenylindole, DAPI) and each knockdown-infection sample was imaged using fluorescence microscopy.

After DAPI-staining each multi-well plate was fluorescence-imaged using a Cytation 5 Imaging Reader to quantify cell viability following siRNA transfection. Knockdowns maintaining cell viability within +/- 1.5 standard deviations of the plate mean were considered acceptable. This resulted in the elimination of RPSA from further evaluation due to excessive cell death following siRNA knockdown.

HIV infection efficiency following siRNA knockdown was assessed using fluorescence microscopy imaging and quantification of CFP and mCherry expression in each sample. The knockdown efficiency of the target proteins was evaluated by western blot analysis showing greater than 87% knockdown of proteins MOV10, HNRNPC, HSPD1, and RBBP4 (Fig. 2.3c and Supplementary Figure S2.3). Figure 3b shows the CFP and mCherry intensities, normalized to DAPI-staining, for each of the successful siRNA knockdowns relative to the siCTRL knockdowns. The siGFP sample has a significant decrease in expression of CFP (and not mCherry) further supporting the efficacy of the siRNA knockdown in the samples. All four target proteins knocked down in 293T cells resulted in statistically significant changes in both CFP and mCherry expression upon HIV infection (Fig. 2.3b). The knockdown of three proteins (MOV10, HNRNPC, and HSPD1) resulted in a decrease of CFP production while the knockdown of RBBP4 resulted in an increase. Additionally, all four protein knockdowns (MOV10, HNRNPC, HSPD1, and RBBP4) resulted in decreases in mCherry production. The incubation of cells with siRNAs targeting the knockdown of proteins YBX3, IGF2BP1, YWHAH did not result in significant decreases in protein production per western blot analysis and also did not result in statistically significant changes in either CFP or mCherry expression.

We utilized siRNA knockdown of HIV-RNA binding proteins, infection with a dual fluorescence HIV clone (Fig. 2.3a), and fluorescence microscopy to demonstrate the efficacy of HyPR-MS while revealing potential function of such proteins in the HIV replication process. Four proteins analyzed in this way showed statistically significant changes in CFP and mCherry expression, suggesting their involvement in HIV replication. Of these four

proteins, three (MOV10, HNRNPC, and HSPD1) are known RNA-binding proteins (RBPs) though only MOV10 has been previously shown to bind to HIV-RNA specifically[31].

Interestingly, one protein, RBBP4, is not a known RBP; however, it is a known histone binding protein involved in chromatin remodeling, regulation of transcription and negative regulation of gene expression[117,118]. Additionally, RBBP4 has been identified as a component of the HIV pre-integration complex (PIC)[119] and to interact with the HIV transcription factor Tat. Our HyPR-MS capture data shows a much higher abundance of RBBP4 in the HIV capture samples relative to the mRNA capture samples. This suggests that RBBP4 does not interact ubiquitously with mRNAs in general but does interact, either directly or indirectly, with HIV RNA. Based on the above mentioned knowledge of RBBP4 function, it is possible that RBBP4 is not directly binding with the HIV RNA but is interacting through intermediate molecules such as Tat the PIC, or histones at the site of HIV RNA transcription.

The RBBP4 siRNA knockdown data presented here shows an increase in CFP expression and a decrease in mCherry expression; this is a proxy for showing an increase in expression of the full length HIV transcript (CFP) but a decrease in the expression of the completely spliced transcript (mCherry). Though additional experiments would be required to establish the function of RBBP4 in HIV replication, this data suggests that RBBP4 perhaps negatively regulates the expression of the full-length transcript but also could have a role in HIV splicing or in increasing the expression of the completely spliced HIV RNA. Furthermore, the identification of RBBP4 by HyPR-MS as an HIV-RNA interactor demonstrates how the use of formaldehyde crosslinking of RNA-protein complexes allows for the identification of

proteins that may not be in direct contact with the RNA but still play a role in HIV replication.

Many HNRNPs have been previously shown to interact with HIV-RNA or protein Rev to affect splicing and/or nuclear export during HIV replication. Many of these were identified here using HyPR-MS. We have also identified an additional HNRNP as an HIV-RNA interactor, HNRNPC; a protein generally involved in RNA stability[93] and known to bind N6-methyladenosine modified RNAs[101]. The siRNA knockdown of HNRNPC resulted in severe decreases in expression of both CFP (full-length HIV-RNA) and mCherry (completely spliced HIV RNA) supporting that HNRNPC is involved in maintenance of HIV RNA stability.

The knockdown of proteins MOV10 and HSPD1 each show a statistically significant decrease in expression of CFP that is greater than that of mCherry (Fig. 2.3b). Interestingly, both of these proteins have been previously shown to be components of the Staufen-1 HIV-RNP (SHRNP)[108]. This complex is thought to be, with protein Gag, a conduit for packaging HIV-RNA into virions because the depletion of Staufen1 in infected cells has been shown to reduce Gag protein levels and to deregulate the process of virion formation[48]. The knockdown of the two SHRNP components here results in the decrease of the HIV-RNA that codes for Gag protein, supporting this claim.

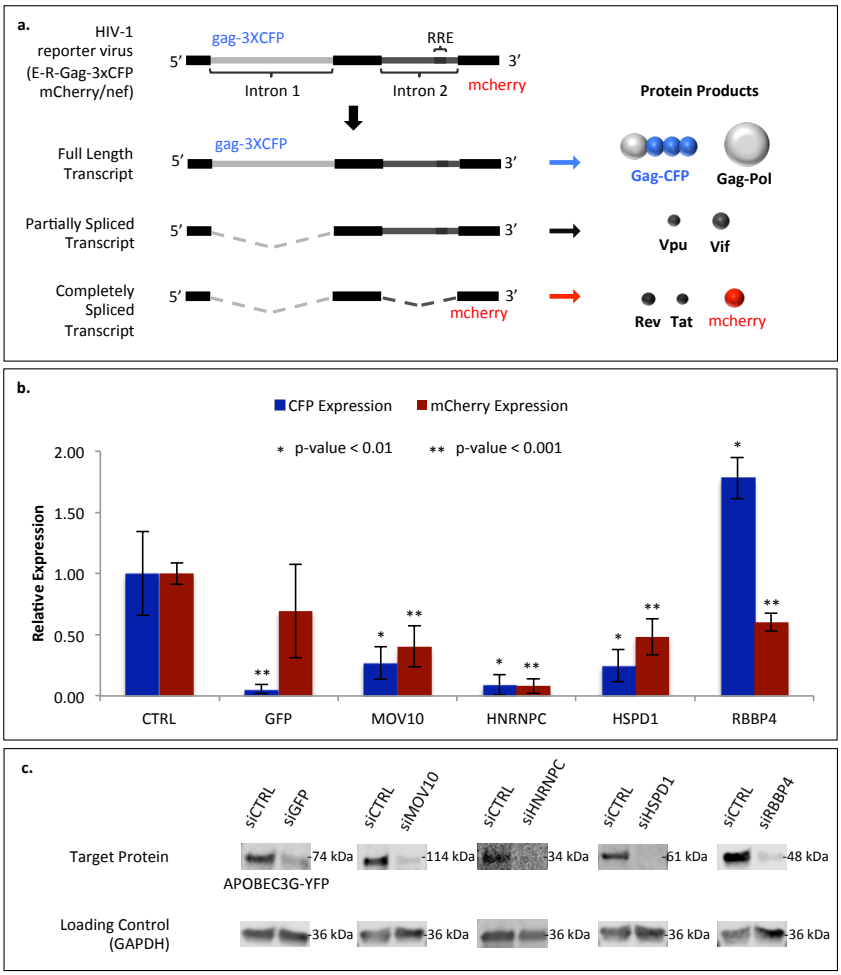


Figure 2.3: a. Diagram of the HIV-1 reporter virus used for functional evaluation of HIV RNA interactors. The RNA sequences coding for the fluorescent reporter proteins, CFP and mCherry, are postioned so that three CFP and one mCherry molecules are expressed following translation from the full length and the completely spliced transcripts, respectively. b. Quantitative analysis of fluorescence from reporter proteins following target protein siRNA knockdown and 48-hours post HIV infection. c. Western blot analysis of siRNA knocked-down target proteins. Demonstrates >90% knock down of each protein. Images and quantitative analysis of complete blots are in SI Figure S3.

2.3.8 Advantages of HyPR-MS. HyPR-MS identified many proteins with known functions in HIV unspliced RNA processing including alternative splicing, nuclear export, regulation of mRNA stability, translation, regulation of translation, and packaging into virions. This variety suggests that the strategy is capable of capturing unspliced HIV RNA at multiple stages throughout HIV replication. HyPR-MS also identified many proteins that do not yet have known functions in HIV replication, but have GO annotated functions in RNA processing, and in particular RNA-binding. These novel unspliced HIV RNA protein interactors are promising targets for future studies probing host factor functions in HIV replication.

Many techniques are available for interrogating HIV host factors, each unveiling new information about the virus and disease. A subset of such techniques includes those that identify proteins that associate with HIV RNA. These techniques successfully identified proteins from purified sub-genomic or reporter HIV RNAs from transfected, stably-transduced or transformed human cell lines (293T, U2OS, HeLa) using MS2 stem loop/coat protein affinity purification. In contrast, HyPR-MS purified near-native unspliced HIV RNAs from an infected T-cell line (Jurkat) and only relied on the sequence specificity of the capture oligonucleotides to extract RNA-protein complexes of interest. For development of HyPR-MS, multiple experimental iterations were required to optimize the many parameters affecting sample output. It was desirable to have a stable, reproducible, and readily obtained cell type for such studies: for this reason we chose to use the Jurkat T lymphocyte cell line, a widely used model for HIV infection studies[65]. The procedures developed and employed in the present study are quite general, and could readily be adapted to studies on primary cells.

The advantage of HyPR-MS is threefold: 1) the strategy does not require modification or overexpression of the HIV RNA nor any cellular proteins giving confidence that the interactions are not artifactual in nature; 2) the RNA-protein interactions are preserved *in vivo* by formaldehyde crosslinking prior to any perturbation of the cellular system giving confidence that the interactions are biologically relevant; and 3) only 30 nucleotides of the native HIV RNA sequence is required for purification of the complexes permitting the application of HyPR-MS nearly universally with regards to HIV-1 strain, cell line and cell type. The versatility of the HyPR-MS technique as well as its ability to capture unspliced HIV RNA throughout its lifespan suggests it will be a powerful tool for interrogation of *in vivo* HIV RNA-protein interactions under various experimental perturbations to gain insights into how HIV replication occurs and can be prevented. Although the focus of this study is the interactions of proteins with HIV RNAs, the technology itself is likely to be readily adapted to other RNA species such as mRNAs and lncRNAs. In fact, preliminary work in our laboratory applying HyPR-MS to both types of RNA targets is underway and will be presented in future publications.

2.4 METHODS:

2.4.1 Cell Culture, Virus Production, and Infections. HEK293T cells were cultured in DMEM media supplemented with 10% fetal bovine serum and 1% L-glutamine-penicillin-streptomycin at 37ºC in 5% CO_2. To generate single-round infectious HIV-1 virions, $2.5X10^6$ HEK293T cells were plated in 10 cm tissue-culture treated dishes in 10mL media. Each 10 cm dish was transfected using polyethylenimine (PEI) with 1μg of DNA plasmid expressing the G envelope glycoprotein from vesicular stomatitis virus (VSV-G) and 9μg of

plasmid DNA encoding the full-length NL4-3 molecular clone of HIV-1 bearing inactivating mutations in *env*, *vpr*, and expressing a Cyan Fluorescent Protein (CFP) reporter from the *nef* reading frame (HIV-1 E-R-CFP)[120-122]. At 24 hours post-transfection, old media was removed and replaced with 4mL fresh media. At 48 hours post-transfection, culture supernatants were harvested, filtered through a sterile 0.45μm syringe filter, aliquoted, and frozen at -80ºC. Infectivity of virus produced was determined by small-scale infection of Jurkat T-cells to quantify dose of viral inoculum required for large-scale infections. Jurkat T-cells (obtained from ATCC) were cultured in RPMI media supplemented with 10% fetal bovine serum and 1% L-glutamine-penicillin-streptomycin. Jurkat T-cells were expanded in 850 cm^2 tissue culture treated roller bottles rotated at 3 rotations per minute. A cell density of $1X10^6$ cells per mL of media was maintained by regular quantification. $300X10^6$ Jurkat T-cells were infected in a low volume (typically 25mL RPMI + 25mL viral inoculum in DMEM) for three hours rotating as above. Polybrene was added at a concentration of 10μg/mL to increase infectivity. After three hours, culture volume was increased to 300 mL using RPMI media. Infected cells were incubated and rotated as above for 45 hours. Successful infection was confirmed by visualizing CFP expression via epifluorescent microscopy. Typically >90% of cells were infected. Uninfected control cells were treated similarly (combined with 25mL DMEM media instead of viral inoculum). At this point, cells were concentrated via centrifugation at 1500rpm for 10 minutes and washed three times with PBS. Washed cells were cross-linked by resuspending in 0.25% formaldehyde (diluted in PBS) and incubated at room temperature for 10 minutes prior to centrifugation. Cross-linked cells were washed once with 1xPBS and then resuspended in 100mM Tris-HCl (pH 7.4) to quench cross-linking reactions for 10 minutes at room temperature. Cells were washed twice more in 1xPBS, pelleted by centrifugation, and frozen at -80ºC.

2.4.2 Special Considerations for HyPR-MS. All solutions used for cell lysis, hybridization, capture, release, and RT-qPCR were prepared using certified RNase free components. Water added to all solutions was nuclease free UltraPure distilled water (Invitrogen, 10977-015).

The amount of lysate, the concentration of capture-oligonucleotides, and the volume of streptavidin coated magnetic beads needed in each capture experiment for identification of proteins interacting with HIV RNA is conditional on the number of copies of the HIV RNA present in the infected cell culture. The capture oligonucleotide concentration and volume of streptavidin coated magnetic beads needed for optimal capture efficiency and specificity for each biological replicate was empirically determined. To do this, small-scale capture experiments using $5X10^5$ cells were performed using increasing amounts of capture oligonucleotides and streptavidin coated beads. The amount of HIV RNA captured was then measured using RT-qPCR (as described below) and the oligonucleotide concentration and bead volume producing the highest capture efficiency while maintaining a desirable capture specificity was scaled up for the large-scale capture experiments. The number of cells needed for a capture experiment for protein identification was then estimated and empirically confirmed by mass spectrometry. For the data presented here, $7.5X10^7$ HIV-1 infected Jurkat cells, 188 pmol of capture or scrambled oligonucleotides, 1.125 mL of streptavidin coated beads, and 188 nmol of release oligonucleotides were used for each capture sample.

2.4.3 Cell Lysis. Cells were resuspended on ice in lysis buffer (469mM LiCl, 62.5mM Tris HCl, pH 7.5, 1.25% LiDS, 1.25% Triton X-100, 12.5mM Ribonucleoside Vanadyl Complex, 12.5mM DTT, 125U/mL RNasin Plus, 1.25X Halt Protease Inhibitors) to a final cell concentration of $5X10^6$ cells/mL. Cells were lysed by frequent vortexing for 10 minutes, keeping the cells on ice between vortexes. The cell lysate was then sonicated on ice for 30 seconds with 4 seconds of rest between each 4-second sonication interval. The lysate was then centrifuged at 1000 g for 2 minutes at 4°C and the supernatant transferred to a new tube.

2.4.4 Hybridization and Capture. The lysate was diluted with nuclease free water so that the final component concentrations for hybridization are as follows: 375mM LiCl, 50mM Tris, 1% LiDS, 1% Triton X-100, 10mM RVC, 10mM DTT, 100U/mL RNasin Plus, 1X Halt Protease Inhibitors. The HIV-capture oligonucleotide or the scrambled oligonucleotide (Table S2) was added to 15mL of diluted lysate and the sample was incubated while nutating at 37°C for 3 hours. The predetermined volume of streptavidin coated magnetic Speedbeads (Thermo Fisher Scientific, 09981140) was washed 3 times with and resuspended in three volumes of bead wash buffer (375mM LiCl, 50mM Tris, 1% LiDS, 1% Triton X-100) prior to addition to each hybridization mixture. The bead capture mixture was then nutated at 37°C for 1 hour. Following incubation the beads were collected to the side of the tube using a magnet and the remaining lysate was removed. A volume of pre-warmed wash buffer (100mM LiCl, 50mM Tris, 0.2% LiDS, 0.2% Triton X-100, 37C) that was 5X the volume of beads used for capture in the sample was used to resuspend the beads and the beads were washed at 37°C for 15 minutes. The beads were then washed with a 5X volume of release buffer (375mM LiCL, 50mM Tris) at room temperature for 5 minutes.

2.4.5 Release from Beads. The beads were resuspended in release buffer (a volume 3X that of the beads used for capture) and the release oligonucleotide (Table S2) was added. The mixure was nutated at room temperature for 30 minutes followed by magnetic separation of the beads from the supernatant containing the released RNA-protein complexes. The solution was transferred to a new tube and divided into aliquots for RT-qPCR and mass spectrometric protein analysis.

2.4.6 RT-qPCR. Samples were incubated at 37°C with 1 mg/mL proteinase, 4mM $CaCl_2$, and 0.2% LiDS. The RNA was extracted with TriReagent (Sigma, T9424) per manufacturer's protocol and was then precipitated in 75% ethanol at -20°C for several hours. The tube was then centrifuged at 20,800 g to pellet the RNA and the RNA was then washed with 75% ethanol. The RNA pellet was resuspended in 15uL of nuclease free water and 10uL was used for reverse transcription (High Capacity cDNA Reverse Transcription Kit, Applied Biosystems) per the manufacturer's protocol. The resulting cDNA sample was analyzed with sequence-specific qPCR assays (Table S1) for quantitation of relevant transcripts.

2.4.7 eFASP. The protocol for protein preparation was adapted as follows from that described by Erde, J. et al. [123]. Amicon 50kDa MWCO filters (Millipore, UFC505096) and collection tubes were passivated by incubating overnight in 1% CHAPS and then rinsed thoroughly with mass spectrometry grade water. Each release sample was brought to 0.1% deoxycholic acid and 8M urea. The sample was passed through the filter in 500uL increments by centrifugation for 10 minutes at 14,000 g and the eluant was discarded. 400uL

of exchange buffer (8M urea, 0.1% DCA, 50mM Tris pH 7.5) was added to the filter and the tube was centrifuged at 14,000 g for 10 minutes. This was repeated for a total of 3 exchanges. 200uL of reducing buffer (8M urea, 20mM DTT) was added to the filter and the sample was incubated for 20 minutes at room temperature followed by centrifugation. 200uL of alkylation buffer (8M urea, 50mM iodoacetamide, 50mM ammonium bicarbonate) was then added to the sample, followed by incubation for 20 minutes at room temperature in the dark, and centrifugation at 14,000 g for 10 minutes. Finally, the sample was exchanged with 3 aliquots of 400uL of digestion buffer (1M urea, 50mM ammonium bicarbonate, 0.1% DCA) and resuspended in a final volume of 100uL of digestion buffer. Trypsin was added to the sample, the filter was transferred to a fresh, passivated collection tube, and the cap was sealed with parafilm followed by incubation overnight at 37°C for protein digestion. The filter-collection tube was centrifuged for 10 minutes at 14,000 g. 50uL of 50mM ammonium bicarbonate was added to the filter and centrifuged at 14,000 g for 10 minutes. This step was repeated once to ensure the collection of the entire peptide sample. The 200uL peptide sample was then brought to 1% TFA followed by addition of 200uL of ethyl acetate. The sample was vortexed for 1 minute then centrifuged at 15,800 g for 2 minutes. The top layer was aspirated and discarded and extraction with ethyl acetate was repeated 2 times. The aqueous layer was then dried using a Savant SVC-100H SpeedVac Concentrator and the sample resuspended in 150uL 0.1% TFA. For removal of salts from the sample a C18 solid-phase extraction pipette tip (OMIX C18, 100uL, Agilent Technologies) was first conditioned with 70% ACN, 0.1% TFA, and then equilibrated with 0.1% TFA. The peptide sample was then loaded onto the C18 solid phase by repeated passing of the 150uL sample over the cartridge. The OMIX pipette tip was then rinsed with 0.1% TFA 10 times followed by peptide elution in 150µL 70% ACN, 0.1% TFA. The

samples were then dried using the SpeedVac Concentrator and reconstituted in 95:5 H2O:ACN, 0.1% formic acid.

2.4.8 Mass Spectrometry of Peptides. The samples were analyzed using an HPLC-ESI-MS/MS system consisting of a high performance liquid chromatograph (nanoAcquity, Waters) set in line with an electrospray ionization (ESI) Orbitrap mass spectrometer (LTQ Velos, ThermoFisher Scientific). A 100 μm id X 365 μm od fused silica capillary micro-column packed with 20 cm of 1.7 μm-diameter, 130 Angstrom pore size, C18 beads (Waters BEH) and an emitter tip pulled to approximately 1 μm using a laser puller (Sutter Instruments) was used for HPLC separation of peptides. Peptides were loaded on-column with 2% acetonitrile in 0.1% formic acid at a flow-rate of 400nL/minute for 30 minutes. Peptides were then eluted at a flow-rate of 300 nL/minute over 120 min with a gradient from 2% to 30% acetonitrile, in 0.1% formic acid. Full-mass profile scans (300-1500 m/z) were performed in the FT orbitrap at a resolution of 60,000. The ten highest intensity parent ions were selected for MS/MS HCD scans at 42% relative collision energy, 7,500 resolution, and with a mass range starting at 100 m/z. Dynamic exclusion was enabled with a repeat count of two over a duration of 30 seconds and an exclusion window of 120 seconds. The Orbitrap raw files were analyzed using MaxQuant (version 1.5.3.30)[71] and searched with Andromeda[72] using the combined Uniprot[86] canonical protein databases for human and HIV-1 and supplemented with common contaminants (downloaded June 8, 2016). Samples were searched allowing for a fragment ion mass tolerance of 20 ppm and cysteine carbamidomethylation (static) and methionine oxidation (variable). A 1% false discovery rate for both peptides and proteins was applied. Up to two missed cleavages per peptide were allowed and at least two peptides were required for protein identification and

quantitation. Protein quantitation was achieved using the sum of the peptide peak intensities for each protein of each biological replicate and capture sample type. The peak intensities of HIV capture samples were normalized by the total peak intensity of all HIV capture samples and the same was done for scrambled capture samples.

2.4.9 Statistical Analysis of Protein Data. The intensities for each protein were then analyzed using the Perseus[74] companion software to MaxQuant for statistical differences between the 3 HIV capture and 3 scrambled capture biological replicates. Following log transformation of peak intensities, missing values were imputed with the width setting at 0.3 and the downshift set to 1.8. A permutation-based 1% FDR analysis of these values, with S_0 set to 0.8, produced 189 proteins statistically enriched in the HIV capture samples compared to the scrambled capture samples.

2.4.10 Cell culture, stable cell line, and HIV-1 reporter virus. Human 293T cells stably-expressing YFP-APOBEC3G were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum, 1% L-glutamine, and 1% penicillin-streptomycin. For all experiments, cells were maintained at 37°C and 5% CO_2 in a humidified incubator. The virus used was a two-color fluorescent HIV-1 reporter virus (E-R- Gag-3xCFP mCherry/nef) which expresses mCherry in the nef ORF and three copies of CFP, in tandem, between the matrix and capsid ORFs of Gag, in a similar but expanded manner as Mergener 1992, Muller 2004, Holmes 2015, and Hendrix 2015. This virus allows screening for early (mCherry; multiply-spliced genes) and late (CFP; unspliced gene products). Stocks of viral inoculum were produced in 293T cells by transfecting the E-R-Gag-3xCFP mCherry/nef with psPAX2 and VSV-G.

2.4.11 siRNA knockdown and infection. For siRNA knockdown cells were cultured in 48-well plates. A dilution of DarmaFECT transfection reagent in Opti-MEM was mixed with the appropriate siRNA pool also diluted in Opti-MEM and allowed to incubate for 20 minutes. This solution was then brought to 300uL with antibiotic free DMEM, mixed well, and used to replace the existing media in each well. The final, in-well concentration of each siRNA was 25nM. After four hours of incubation, the media was replaced with fresh media and the cells were incubated overnight. Approximately 24-hours post transfection the cells were split into two plates and a second siRNA transfection as described above was conducted an additional 24-hours later, in both plates. Again, four hours post transfection the siRNA containing media was replaced with fresh media. Additionally, in one plate, polybrene was added to the media followed by the HIV-1 reporter virus (E-R- Gag-3xCFP mCherry/nef). After 24 hours the media is exchanged for fresh media and 48-hours post infection the cells are washed with PBS and fixed for 12 minutes using 4% paraformaldehyde (PFA) in PBS then stored at 4C in PBS until imaged.

2.4.12 DAPI staining and fluorescence imaging. Fixed cells were permeabilized using 0.2% Triton X-100, and stained with DAPI (4',6'-diamidino-2-phenylindole4',6-diamidino-2-phenylindole). DAPI fluorescence was measured using a Cytation 5 Imaging Reader (Biotek Instruments, Inc.) operated by Gen5 software (v 2.07) using excitation/emission monochromator range (wavelengths in nanometers) 340 to 380/420 to 480 (DAPI). Cell number based on the relative DAPI signal is later used to normalize imaging data. Additional imaging experiments were performed on a Nikon Ti-Eclipse inverted wide-field epifluorescent deconvolution microscope (Nikon Corporation). Images were collected using

an Orca-Flash 4.0 C11440 (Hamamatsu Photonics) camera and Nikon NIS Elements software (v 4.20.03) using Nikon 4x/0.13 (Plan Apo) objective lense and the following excitation/emission filter set ranges (wavelengths in nanometers): 418 to 442/458 to 482 (CFP), 490 to 510/520 to 550 (YFP), 555 to 589/602 to 662 (mCherry). Images were processed and analyzed using FIJI/ImageJ2[124]. Results were obtained from two biological replicates, defined as cells treated with siRNAs and infected with virus on separate days.

2.4.13 Western blot analysis. For western blot analysis cells were prepared identically to those described in "siRNA knockdown and infection" except instead of paraformaldehyde fixation the cells were lysed in radioimmunoprecipitation assay (RIPA) buffer (10 mM Tris-HCl [pH 7.5], 150 mM NaCl, 1 mMEDTA, 0.1% SDS, 1% Triton X-100, 1% sodium deoxycholate) containing complete protease inhibitor cocktail (Roche). The cell lysates were then sonicated briefly and centrifuged for 10 minutes at 1000xg then boiled in dissociation buffer (62.5 mM Tris-HCl [pH 6.8], 10% glycerol, 2% sodium dodecyl sulfate [SDS], 10% beta-mercaptoethanol) at a 1:1 ratio. SDS-PAGE was performed on a 4% to 15% polyacrylamide gel, and the proteins were transferred onto a nitrocellulose membrane. The membranes were incubated with the appropriate antibodies (Supplementary Table S8) at a 1:1000 dilution overnight. Additionally, the gene encoding GAPDH, a housekeeping gene, was detected with an additional antibody in each membrane-incubation to normalize band density values to total cell protein values. Following washing of the membranes, secondary IRdye680- or IRdye800-conjugated antibodies (Supplementary Table S8) were incubated at 1:10000 dilution for 1.5 hours and used for quantitative immunoblotting with an Odyssey infrared scanner (Li-Cor Biosciences).

2.5 SUPPLEMENTARY INFORMATION:

The objective of HIV-HyPR-MS is to use hybridization capture of unspliced HIV RNA-protein complexes to identify proteins relevant to the process of HIV replication. Several challenges must be addressed to succeed in this goal.

1. Biological Relevance: *In vivo* RNA-protein interactions must be stabilized to withstand downstream experimental conditions.
2. Specificity of Capture: Unspliced HIV RNA-protein complexes must be specifically isolated from a complex lysate milieu.
3. Magnitude of Capture: High capture efficiency must be achieved so that a reasonable number of cells can be used for identification of proteins by mass spectrometry.
4. Determination of "true binders": Proteins that truly interact with the HIV RNA (true binders) must be differentiated from non-specific interactors using a combination of control experiments and statistical analysis strategies.

Below we describe the experimental strategies employed in HyPR-MS to overcome these challenges.

Culture, HIV-1 Infection, and Crosslinking of Jurkat T-cells.

Cell type selection and the HIV infection strategy are integral factors for obtaining both a sufficient magnitude of capture and biological relevance. Jurkat T-cells were selected because they are amenable to growth in large numbers as compared to primary peripheral blood mononuclear cells (PBMC) or purified CD4+ T-cells. Additionally, Jurkat T-cells largely recapitulate the cellular environment that HIV typically infects (CD4+ T-cells) ensuring that the hybridization capture results will be biologically relevant. HIV-1 infection was achieved

by using the replication incompetent NL4-3 molecular clone. The use of a molecular clone allowed us to know the genomic sequence, gene expression, and kinetics of the viral life cycle while also maintaining experimental reproducibility. Because the development of this multistep, complex technology requires several experimental iterations to establish the parameters optimal for success, Jurkat T-cells infected with the NL4-3 HIV molecular clone were ideal selections.

RNA-protein interactions in the cell are often dynamic, transient, and labile. These interactions must be stabilized so that biologically relevant interactions are preserved while non-specific interactions are disrupted during the HyPR-MS process. To achieve this the HIV infected cell culture is incubated with formaldehyde. Formaldehyde is a fast acting, zero-length, membrane permeable covalent cross-linker that only crosslinks molecules that are in contact with each other. Once covalently linked, RNA-protein and protein-protein interactions are preserved while non-specific interactions are disrupted using harsh salt and detergent concentrations experimentally downstream. This helps to ensure that proteins identified are either directly interacting with HIV RNA or are interacting with other proteins that directly interact with HIV RNA.

HIV Capture Oligonucleotide and Scrambled Capture Oligonucleotide Design and Placement. The capture oligonucleotide is a key design feature of HyPR-MS for obtaining specificity of capture. The oligonucleotide contains three elements: a 30-nucleotide (nt) complementary sequence, a 3'-biotin, and a 5' 8-nt non-complementary sequence (toehold). The complementary sequence is designed to specifically hybridize to the unspliced HIV RNA, the biotin then anchors the target RNA-protein complex to the streptavidin coated

magnetic beads, and the toehold sequence provides a mechanism for release of the HIV RNA-protein complex from the beads. The capture oligonucleotide must be designed so that it hybridizes to the unspliced HIV RNA but does not hybridize to any other transcript in the lysate, including the partially and completely spliced HIV RNA variants. The complementary sequence of the oligonucleotide was designed to be complementary to a region in a ~5kb sequence that is retained in the unspliced transcript but is spliced out of both the partially and the completely spliced transcripts (Supplementary Figure S2.1). The region of hybridization was further informed by the secondary structure of the full length HIV RNA. Watts, J, et. al. used Selective 2′-Hydroxyl Acylation Analyzed by Primer Extension, or SHAPE, to determine the structure of the HIV-1 genome in virions[125]. SHAPE determines the probability that a given nucleotide is not base-paired, or is single stranded. Based on this information, the capture oligonucleotide was designed to be complementary to a region that has a high probability to be single stranded and thus more likely to be accessible to the capture oligonucleotide. Finally, an online Basic Local Alignment Search Tool (BLAST) was used to determine that the capture oligonucleotide sequence is sufficiently specific to the target transcript so that wash conditions remove any off-target interactions. The complementary sequence of the capture oligonucleotide is a 30-nt DNA sequence with a T_m of ~68ºC at the conditions of hybridization (Supplementary Table S2.2). Following hybridization of the capture oligonucleotide to the HIV RNA, the streptavidin coated magnetic beads are added to the lysate. The biotin of the capture oligonucleotide then anchors the HIV RNA-protein complex to the beads and a magnet is used to pull the beads to the side of the tube so that the remaining lysate can be removed and the beads can be washed.

The capture oligonucleotide design incorporates a mechanism for release of the HIV RNA-protein complexes from the beads. We have developed a strategy that incorporates an 8-nt toehold sequence in the capture oligonucleotide for a programmable release from the beads[69]. The beads are resuspended in a release buffer and the 38-nt toehold release oligonucleotides are added to the solution. The release oligonucleotide is designed to be fully complementary to the entire 38-nt capture oligonucleotide. The 8-nt sequence on the capture oligonucleotide, which is not complementary to the target RNA, hybridizes with its complement on the release oligonucleotide. This gives the release oligonucleotide a "toehold" to hybridize with the remaining of the capture oligonucleotide. Since the target RNA is only complementary to 30-nt of the capture oligonucleotide and the release oligonucleotide is fully complementary, the interaction of the capture oligonucleotide with the release oligonucleotide is thermodynamically more favorable and the target RNA is released from the capture oligonucleotide into solution.

A second oligonucleotide, the scrambled oligonucleotide, was also designed to serve as a negative control capture. This scrambled oligonucleotide was designed to have the same number of nucleotides and approximately the same T_m as the capture oligonucleotide but does not have significant complementarity to the target transcript nor any other transcript in the cells. Capture experiments using the capture oligonucleotide and the scrambled oligonucleotide were performed in parallel.

HIV RNA Quantitation. Each capture sample was analyzed using RT-qPCR to determine the effectiveness of our technology for purifying HIV RNA-protein complexes. Determination of the capture efficiency and specificity and enrichment of HIV was achieved

using qPCR assays specific to HIV RNA and to GAPDH RNA (Supplementary Table S1). Since unspliced HIV RNA is reverse transcribed in the host cells during HIV replication, there is potential to capture the HIV DNA in addition to the HIV RNA. To ensure that the nucleic acid captured by HyPR-MS is, in fact, RNA, we analyzed the capture samples by RT-qPCR without reverse transcriptase added to the reverse transcription reactions. For these experiments, it should be noted that, the RNA/DNA purification step used a phenol-chloroform extraction instead of the usual Trizol extraction so that both RNA and DNA captured would be isolated from the capture sample. The polymerase used for qPCR does not bind and amplify RNA so any signal obtained in the no-reverse transcriptase reactions will be due to the presence of DNA. From this analysis it was determined that the presence of DNA is over 3 orders of magnitude lower than RNA in the capture samples ensuring proteins detected are associated with HIV RNA, not DNA (Figure S2.2).

Protein target selection for functional analysis. The statistical analysis used to identify HIV-RNA protein interactors (student's t-test with 1% permutation-based FDR) identified about 190 potential proteins for functional analysis. We sought to select several of these for analysis using three parameters: the p-value obtained from the student's t-test analyzing HIV capture compared to scrambled capture; the intensity or abundance of the protein in the HIV capture samples; and specificity of the HIV RNA-protein interaction by determining if that interaction is common among other cellular RNAs or if it is relatively unique to the HIV-RNA. First, the p-value parameter is obtained from the statistical analysis described above and is presented in Supplementary Table S3 and again in Supplementary Table S7 for this analysis. The primary purpose of this statistical test here is to detect differences in protein quantities between the HIV capture samples and the negative control (scrambled capture samples). This test accounts for the variability among biological

replicates for determining significant protein quantity differences between the HIV and scrambled capture samples. The second parameter seeks to eliminate proteins that might be present in the HIV capture samples in low abundances that may be near the limit of quantitation of the MS. Using peak intensities for determining relative peptide abundance between samples is a powerful tool that permits relative quantitation without laborious and expensive labeling techniques such as SILAC and TMT-tagging. Any quantitative method has a lower limit of quantitation that is typically determined using a calibration curve. However, each peptide varies in its properties, such as ionization efficiency, which would require that a calibration curve be implemented for each individual peptide to determine that peptide's limit of quantitation. This is not feasible since there are thousands of peptides in each sample. We sought to avoid selecting proteins with low abundances that may be near the limit of detection and thus have more uncertainty in their calculations. Therefore, for the purpose of selecting proteins for functional analysis, we calculated the sum of a protein's peptide intensities in the HIV capture samples and flagged proteins with a total protein intensity in the top quartile of total intensities for all proteins. Lastly, we sought to select for proteins that are not ubiquitous binders to mRNA but are somewhat selective for HIV RNA specifically. To do this we performed a HyPR-MS capture experiment using poly-dT capture oligonucleotides that hybridized with the common poly-A tail found on mature mRNAs. We determined the p-value using the student's t-test for proteins in the HIV capture samples and the total mRNA capture samples as well as the fold difference in protein intensity between the two sample types. These values are also presented in Supplementary Table S7. These three parameters combined with understanding from the literature of the known general function of the protein and its specific function with regards to HIV replication informed our selection of the proteins for functional analysis.

Western Blot Analysis: The blots shown in the main text are only cut-outs of the pertinent proteins targeted for siRNA knockdown. Figure S3 shows the full blot for each siRNA knockdown/western blot along with the ladder for indication of the protein sizes. Quantitation was done using ImageStudioLite to compare the intensity of fluorescence for the knock-down protein and the loading control protein for each western blot. The percentage of knockdown for each targeted protein was determined using these numbers and normalizing using the values obtained for GAPDH.

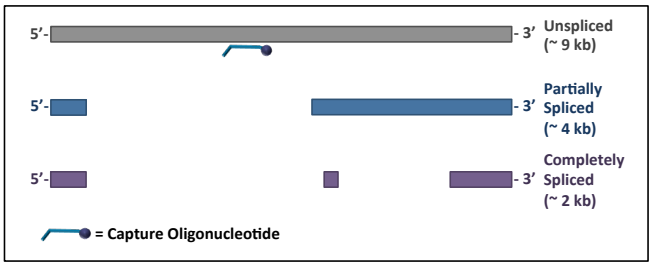


Figure S2.1: The capture oligonucleotide was designed to be complementary to 30 nucleotides of the full length transcript that are not present in the partially spliced and completely spliced variants.

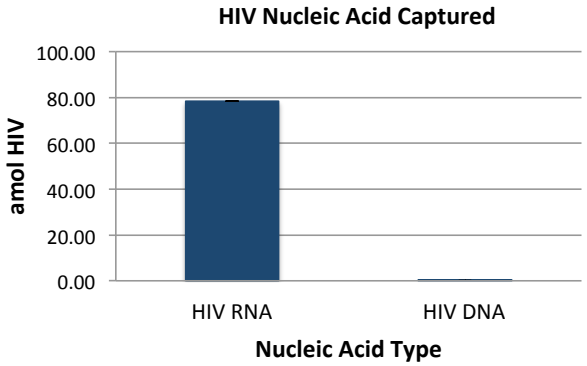
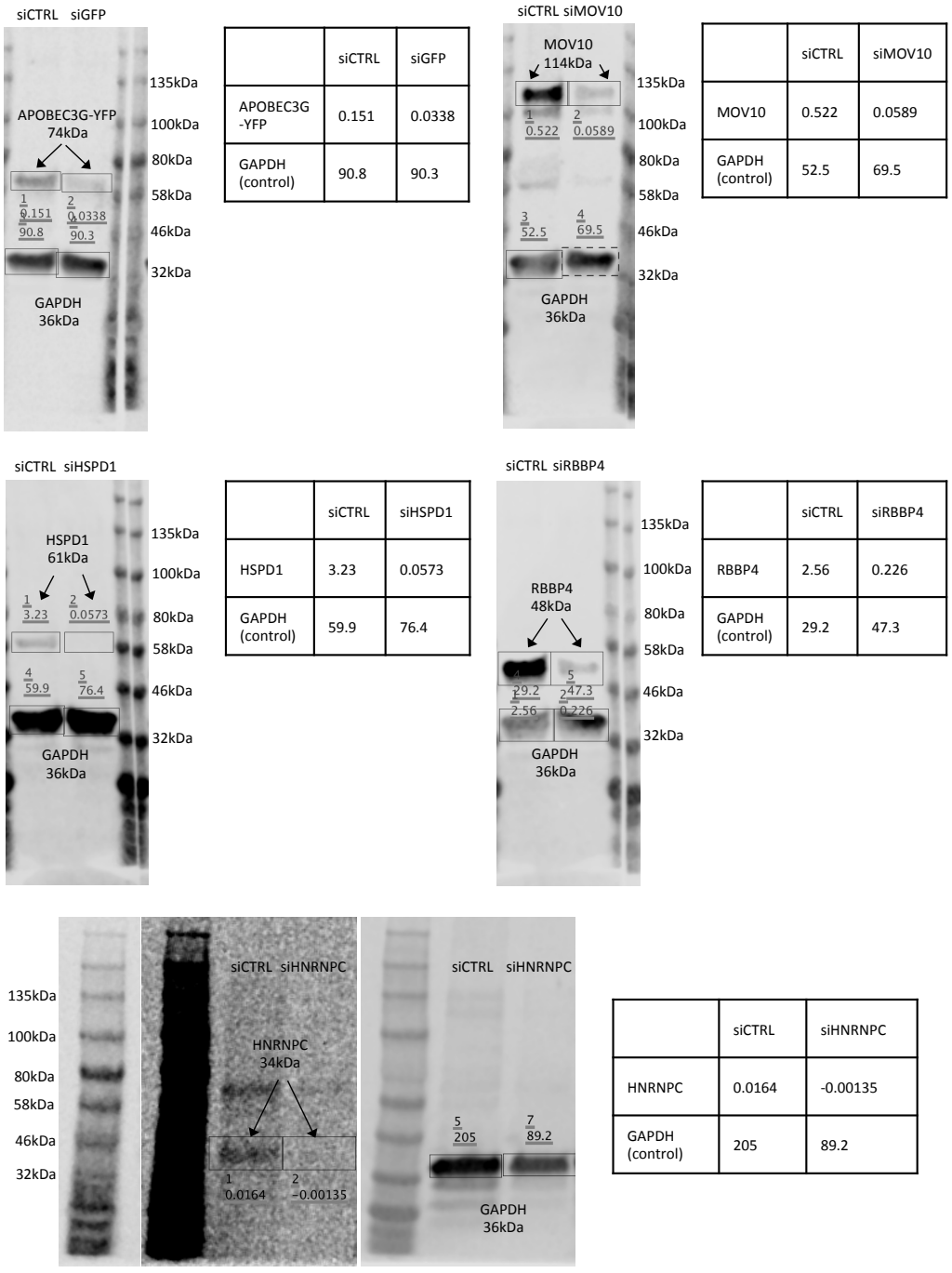


Figure S2.2: Amplification of HIV qPCR region to measure HIV-RNA (sample reverse transcribed then qPCR amplified) and to measure HIV DNA (sample not reverse transcribed then qPCR amplified). Demonstrates that large majority of HIV nucleic acids is in fact RNA, not DNA.



	siCTRL	siGFP
APOBEC3G -YFP	0.151	0.0338
GAPDH (control)	90.8	90.3

	siCTRL	siMOV10
MOV10	0.522	0.0589
GAPDH (control)	52.5	69.5

	siCTRL	siHSPD1
HSPD1	3.23	0.0573
GAPDH (control)	59.9	76.4

	siCTRL	siRBBP4
RBBP4	2.56	0.226
GAPDH (control)	29.2	47.3

	siCTRL	siHNRNPC
HNRNPC	0.0164	-0.00135
GAPDH (control)	205	89.2

Figure S2.3: Full image of each western blot including the target protein for siRNA knockdown, the loading control (GAPDH) and the ladder with sizes.

2.6 AUTHOR CONTRIBUTIONS:

Cell culture, virus production, and infection of cells for HyPR-MS was conducted by J.T.B and supported by NIH grant RO1AI110221 (PI: N.M.S). J.T.B received support from a National Science Foundation Graduate Research Fellow program (grant DGE-1256259), Research Competition Award from the UW—Madison Office of the Vice Chancellor of Research and Graduate Education, and a UW-Madison Graduate School Dissertation Completion Fellowship. HyPR-MS design and implementation, RT-qPCR, protein sample preparation, mass spectrometry data collection, siRNA knockdown with viral infection and fluorescence detection, data analysis and writing of manuscript were conducted by R.A.K. Western blot analysis was conducted by J.T.B and R.A.K. HPLC and mass spectrometer method development and maintenance were conducted by M.S. Work by R.A.K. and M.S. was funded by NIH grants 1P50HG004952 and R01CA193481 (PI: L.M.S). N.M.S. and L.M.S. provided advice and expertise throughout the project. We thank Dr Brian Frey for his comments on this manuscript, and Dr Michael Shortreed for guidance and support for the statistical analysis.

2.7 REFERENCES

1 Purcell, D. F. J. & Martin, M. A. Alternative splicing of human immunodeficiency virus type-1 mRNA modulates viral protein expression, replication, and infectivity. *Journal of virology* **67**, 6365-6378 (1993).

2 Ocwieja, K. E. *et al.* Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Research* **40**, 10345-10355, doi:10.1093/nar/gks753 (2012).

3 Kuzembayeva, M., Dilley, K., Sardo, L. & Hu, W. S. Life of psi: how full-length HIV-1 RNAs become packaged genomes in the viral particles. *Virology* **454-455**, 362-370, doi:10.1016/j.virol.2014.01.019 (2014).

4 Karn, J. & Stoltzfus, C. M. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harbor Perspectives in Medicine* **2**, doi:Artn A006916 10.1101/Cshperspect.A006916 (2012).

5 LeBlanc, J., Weil, J. & Beemon, K. Posttranscriptional regulation of retroviral gene expression: primary RNA transcripts play three roles as pre-mRNA, mRNA, and genomic RNA. *Wiley Interdisciplinary Reviews-Rna* **4**, 567-580, doi:10.1002/wrna.1179 (2013).

6 Felber, B. K., Hadzopoulou-Cladaras, M., Cladaras, C., Copeland, T. & Pavlakis, G. N. rev protein of human immunodeficiency virus type 1 affects the stability and transport of the viral mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 1495-1499 (1989).

7 Askjaer, P., Jensen, T. H., Nilsson, J., Englmeier, L. & Kjems, J. The specificity of the CRM1-Rev nuclear export signal interaction is mediated by RanGTP. *Journal of Biological Chemistry* **273**, 33414-33422, doi:Doi 10.1074/Jbc.273.50.33414 (1998).

8 Kula, A. *et al.* Characterization of the HIV-1 RNA associated proteome identifies Matrin 3 as a nuclear cofactor of Rev function. *Retrovirology* **8**, doi:Artn 60 10.1186/1742-4690-8-60 (2011).

9 Ajamian, L. *et al.* Unexpected roles for UPF1 in HIV-1 RNA metabolism and translation. *Rna-a Publication of the Rna Society* **14**, 914-927, doi:10.1261/rna.829208 (2008).

10 Edgcomb, S. P. *et al.* DDX1 is an RNA-dependent ATPase involved in HIV-1 Rev function and virus replication. *Journal of Molecular Biology* **415**, 61-74, doi:10.1016/j.jmb.2011.10.032 (2012).

11 Huang, F. *et al.* RNA helicase MOV10 functions as a co-factor of HIV-1 Rev to facilitate Rev/RRE-dependent nuclear export of viral mRNAs. *Virology* **486**, 15-26, doi:10.1016/j.virol.2015.08.026 (2015).

12 Lai, M. C. *et al.* Human DDX3 interacts with the HIV-1 Tat protein to facilitate viral mRNA translation. *Plos One* **8**, doi:ARTN e68665 10.1371/journal.pone.0068665 (2013).

13 Soto-Rifo, R., Rubilar, P. S. & Ohlmann, T. The DEAD-box helicase DDX3 substitutes for the cap-binding protein eIF4E to promote compartmentalized translation initiation of the HIV-1 genomic RNA. *Nucleic Acids Research* **41**, 6286-6299, doi:10.1093/nar/gkt306 (2013).

14 Abrahamyan, L. G. *et al.* Novel Staufen1 ribonucleoproteins prevent formation of stress granules but favour encapsidation of HIV-1 genomic RNA. *Journal of Cell Science* **123**, 369-383, doi:10.1242/jcs.055897 (2010).

15 Berkowitz, R. D., Luban, J. & Goff, S. P. Specific binding of human immunodeficiency virus type 1 Gag polyprotein and nucleocapsid protein to viral RNAs Detected by RNA mobility shift assays. *Journal of virology* **67**, 7190-7200 (1993).

16 Zhou, H. L. *et al.* Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host & Microbe* **4**, 495-504, doi:10.1016/j.chom.2008.10.004 (2008).

17 Yeung, M. L., Houzet, L., Yedavalli, V. S. R. K. & Jeang, K. T. A genome-wide short hairpin RNA screening of Jurkat T-cells for human proteins contributing to productive HIV-1 replication. *Journal of Biological Chemistry* **284**, 19463-19473, doi:10.1074/jbc.M109.010033 (2009).

18 Konig, R. *et al.* Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* **135**, 49-60, doi:10.1016/j.cell.2008.07.032 (2008).

19 Brass, A. L. *et al.* Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319**, 921-926, doi:10.1126/science.1152725 (2008).

20 Bushman, F. D. *et al.* Host cell factors in HIV replication: meta-analysis of genome-wide studies. *Plos Pathogens* **5**, doi:ARTN e1000437 10.1371/journal.ppat.1000437 (2009).

21 Engeland, C. E. *et al.* Proteome analysis of the HIV-1 Gag interactome. *Virology* **460**, 194-206, doi:10.1016/j.virol.2014.04.038 (2014).

22 Jager, S. *et al.* Global landscape of HIV-human protein complexes. *Nature* **481**, 365-370, doi:10.1038/nature10719 (2012).

23 Huang, F. *et al.* RNA helicase MOV10 functions as a co-factor of HIV-1 Rev to facilitate Rev/RRE-dependent nuclear export of viral mRNAs. *Virology* **486**, 15-26, doi:10.1016/j.virol.2015.08.026 (2015).

24 Marchand, V. *et al.* Identification of protein partners of the human immunodeficiency virus 1 tat/rev exon 3 leads to the discovery of a new HIV-1 splicing regulator, protein hnRNP K. *RNA biology* **8**, 325-342 (2011).

25 Dejardin, J. & Kingston, R. E. Purification of proteins associated with specific genomic loci. *Cell* **136**, 175-186, doi:10.1016/j.cell.2008.11.045 (2009).

26 Kennedy-Darling, J. *et al.* Discovery of chromatin-associated proteins via sequence-specific capture and mass spectrometric protein identification in Saccharomyces cerevisiae. *Journal of Proteome Research* **13**, 3810-3825, doi:10.1021/pr5004938 (2014).

27 Guillen-Ahlers, H. *et al.* HyCCAPP as a tool to characterize promoter DNA-protein interactions in Saccharomyces cerevisiae. *Genomics* **107**, 267-273, doi:10.1016/j.ygeno.2016.05.002 (2016).

28 Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404-416, doi:10.1016/j.cell.2015.03.025 (2015).

29 McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232-+, doi:10.1038/nature14443 (2015).

30 West, J. A. *et al.* The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Molecular Cell* **55**, 791-802, doi:10.1016/j.molcel.2014.07.012 (2014).

31 Dalgleish, A. G. *et al.* The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature* **312**, 763-767 (1984).

32 Page, K. A., Landau, N. R. & Littman, D. R. Construction and use of a human immunodeficiency virus vector for analysis of virus infectivity. *Journal of virology* **64**, 5270-5276 (1990).

33 Moller, K., Rinke, J., Ross, A., Buddle, G. & Brimacombe, R. The use of formaldehyde in RNA-protein cross-linking studies with ribosomal subunits from Escherichia coli. *Eur J Biochem* **76**, 175-187 (1977).

34 Solomon, M. J. & Varshavsky, A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* **82**, 6470-6474 (1985).

35 Kennedy-Darling, J., Holden, M. T., Shortreed, M. R. & Smith, L. M. Multiplexed programmable release of captured DNA. *Chembiochem* **15**, 2353-2356, doi:10.1002/cbic.201402343 (2014).

36 Barber, R. D., Harmer, D. W., Coleman, R. A. & Clark, B. J. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiological Genomics* **21**, 389-395, doi:10.1152/physiolgenomics.00025.2005 (2005).

37 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**, 1367-1372, doi:10.1038/nbt.1511 (2008).

38 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805, doi:10.1021/pr101065j (2011).

39 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* **13**, 2513-2526, doi:10.1074/mcp.M113.031591 (2014).

40 Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature methods* **13**, 731-740, doi:10.1038/nmeth.3901 (2016).

41 Mu, X., Li, W., Wang, X. & Gao, G. YB-1 stabilizes HIV-1 genomic RNA and enhances viral production. *Protein Cell* **4**, 591-597, doi:10.1007/s13238-013-3011-3 (2013).

42 Mu, X. *et al.* HIV-1 exploits the host factor RuvB-like 2 to balance viral protein expression. *Cell Host Microbe* **18**, 233-242, doi:10.1016/j.chom.2015.06.018 (2015).

43 Woolaway, K., Asai, K., Emili, A. & Cochrane, A. hnRNP E1 and E2 have distinct roles in modulating HIV-1 gene expression. *Retrovirology* **4**, 28, doi:10.1186/1742-4690-4-28 (2007).

44 Dingwall, C. *et al.* HIV-1 tat protein stimulates transcription by binding to a U-rich bulge in the stem of the TAR RNA structure. *EMBO J* **9**, 4145-4153 (1990).

45 Heaphy, S. *et al.* HIV-1 regulator of virion expression (Rev) protein binds to an RNA stem-loop structure located within the Rev response element region. *Cell* **60**, 685-693 (1990).

46 Khorchid, A., Halwani, R., Wainberg, M. A. & Kleiman, L. Role of RNA in facilitating Gag/Gag-Pol interaction. *Journal of virology* **76**, 4131-4137, doi:10.1128/JVI.76.8.4131-4137.2002 (2002).

47 Kessl, J. J. *et al.* HIV-1 Integrase binds the viral RNA genome and is essential during virion morphogenesis. *Cell* **166**, 1257-+, doi:10.1016/j.cell.2016.07.044 (2016).

48 Abbondanzieri, E. A. *et al.* Dynamic binding orientations direct activity of HIV reverse transcriptase. *Nature* **453**, 184-U182, doi:10.1038/nature06941 (2008).

49 Klein, K. C. *et al.* HIV Gag-leucine zipper chimeras form ABCE1-containing intermediates and RNase-resistant immature capsids similar to those formed by wild-type HIV-1 Gag. *Journal of virology* **85**, 7419-7435, doi:10.1128/JVI.00288-11 (2011).

50 Fu, W. *et al.* Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Research* **37**, D417-D422, doi:10.1093/nar/gkn708 (2009).

51 Fahey, M. E. *et al.* GPS-Prot: A web-based visualization platform for integrating host-pathogen interaction data. *Bmc Bioinformatics* **12**, doi:Artn 298 10.1186/1471-2105-12-298 (2011).

52 Bateman, A. *et al.* UniProt: a hub for protein information. *Nucleic Acids Research* **43**, D204-D212, doi:10.1093/nar/gku989 (2015).

53 Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic acids research* **45**, D183-D189, doi:10.1093/nar/gkw1138 (2017).

54 Cullen, B. R. HIV-1 auxiliary proteins: making connections in a dying cell. *Cell* **93**, 685-692, doi:Doi 10.1016/S0092-8674(00)81431-2 (1998).

55 Naji, S. *et al.* Host cell interactome of HIV-1 Rev includes RNA helicases involved in multiple facets of virus production. *Molecular & Cellular Proteomics* **11**, doi:Artn M111.015313 10.1074/Mcp.M111.015313 (2012).

56 Yasuda-Inoue, M., Kuroki, M. & Ariumi, Y. Distinct DDX DEAD-box RNA helicases cooperate to modulate the HIV-1 Rev function. *Biochemical and Biophysical Research Communications* **434**, 803-808, doi:10.1016/j.bbrc.2013.04.016 (2013).

57 Fang, J. H. *et al.* A DEAD box protein facilitates HIV-1 replication as a cellular co-factor of Rev. *Virology* **330**, 471-480, doi:10.1016/j.virol.2004.09.039 (2004).

58 Huelga, S. C. *et al.* Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep* **1**, 167-178, doi:10.1016/j.celrep.2012.02.001 (2012).

59 Shetty, S. Regulation of urokinase receptor mRNA stability by hnRNP C in lung epithelial cells. *Mol Cell Biochem* **272**, 107-118 (2005).

60 Woo, K. C. *et al.* Circadian amplitude of cryptochrome 1 is modulated by mRNA stability regulation via cytoplasmic hnRNP D oscillation. *Mol Cell Biol* **30**, 197-205, doi:10.1128/MCB.01154-09 (2010).

61 Mili, S., Shu, H. J., Zhao, Y. & Pinol-Roma, S. Distinct RNP complexes of shuttling hnRNP proteins with pre-mRNA and mRNA: candidate intermediates in formation and export of mRNA. *Mol Cell Biol* **21**, 7307-7319, doi:10.1128/MCB.21.21.7307-7319.2001 (2001).

62 Damgaard, C. K., Tange, T. O. & Kjems, J. hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. *Rna-a Publication of the Rna Society* **8**, 1401-1415, doi:10.1017/S1355838202023075 (2002).

63 Najera, I., Krieg, M. & Karn, J. Synergistic stimulation of HIV-1 Rev-dependent export of unspliced mRNA to the cytoplasm by hnRNP A1. *Journal of Molecular Biology* **285**, 1951-1964, doi:Doi 10.1006/Jmbi.1998.2473 (1999).

64 Lund, N. *et al.* Differential effects of hnRNP D/AUF1 isoforms on HIV-1 gene expression. *Nucleic Acids Research* **40**, 3663-3675, doi:10.1093/nar/gkr1238 (2012).

65 Vincendeau, M., Nagel, D., Brenke, J. K., Brack-Werner, R. & Hadian, K. Heterogenous nuclear ribonucleoprotein Q increases protein expression from HIV-1 Rev-dependent transcripts. *Virology Journal* **10**, doi:Artn 151 10.1186/1743-422x-10-151 (2013).

66 Hadian, K. *et al.* Identification of a heterogeneous nuclear ribonucleoprotein-recognition region in the HIV Rev protein. *Journal of Biological Chemistry* **284**, 33384-33391, doi:10.1074/jbc.M109.021659 (2009).

67 Liu, N. *et al.* N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* **518**, 560-564, doi:10.1038/nature14234 (2015).

68 Kennedy, E. M. *et al.* Posttranscriptional m(6)A editing of HIV-1 mRNAs enhances viral gene expression. *Cell Host Microbe* **19**, 675-685, doi:10.1016/j.chom.2016.04.002 (2016).

69 Beckham, C. J. & Parker, R. P bodies, stress granules, and viral life cycles. *Cell Host & Microbe* **3**, 206-212, doi:10.1016/j.chom.2008.03.004 (2008).

70 Hinman, M. N. & Lou, H. Diverse molecular functions of Hu proteins. *Cellular and Molecular Life Sciences* **65**, 3168-3181, doi:10.1007/s00018-008-8252-6 (2008).

71 Rivas-Aravena, A. *et al.* The Elav-like protein HuR exerts translational control of viral internal ribosome entry sites. *Virology* **392**, 178-185, doi:10.1016/j.virol.2009.06.050 (2009).

72 Lemay, J. *et al.* HuR interacts with human immunodeficiency virus type 1 reverse transcriptase, and modulates reverse transcription in infected cells. *Retrovirology* **5**, doi:Artn 47 10.1186/1742-4690-5-47 (2008).

73 Ahn, J. *et al.* The RNA binding protein HuR does not interact directly with HIV-1 reverse transcriptase and does not affect reverse transcription in vitro. *Retrovirology* **7**, doi:Artn 40 10.1186/1742-4690-7-40 (2010).

74 Milev, M. P., Ravichandran, M., Khan, M. F., Schriemer, D. C. & Mouland, A. J. Characterization of Staufenl ribonucleoproteins by mass spectrometry and biochemical analyses reveal the presence of diverse host proteins associated with human immunodeficiency virus type 1. *Frontiers in Microbiology* **3**, doi:Artn 367 10.3389/Fmicb.2012.00367 (2012).

75 Burdick, R. *et al.* P body-associated protein Mov10 inhibits HIV-1 replication at multiple stages. *Journal of virology* **84**, 10241-10253, doi:10.1128/JVI.00585-10 (2010).

76 Furtak, V. *et al.* Perturbation of the P-body component Mov10 inhibits HIV-1infectivity. *Plos One* **5**, doi:ARTN e9081 10.1371/journal.pone.0009081 (2010).

77 Abudu, A. *et al.* Identification of molecular determinants from Moloney Leukemia Virus 10 homolog (MOV10) protein for virion packaging and anti-HIV-1 activity. *Journal of Biological Chemistry* **287**, 1220-1228, doi:10.1074/jbc.M111.309831 (2012).

78 Zimmerman, C. *et al.* Identification of a host protein essential for assembly of immature HIV-1 capsids. *Nature* **415**, 88-92, doi:Doi 10.1038/415088a (2002).

79 Dooher, J. E., Schneider, B. L., Reed, J. C. & Lingappa, J. R. Host ABCE1 is at plasma membrane HIV assembly sites and its dissociation from gag is linked to subsequent events of virus production. *Traffic* **8**, 195-211, doi:10.1111/j.1600-0854.2006.00524.x (2007).

80 Ueno, T. *et al.* Nucleolin and the packaging signal, psi, promote the budding of human immunodeficiency virus type 1 (HIV-1). *Microbiology and Immunology* **48**, 111-118 (2004).

81 Zhou, Y. D., Rong, L. W., Lu, J. H., Pan, O. & Liang, C. Insulin-like growth factor II mRNA binding protein 1 associates with gag protein of human immunodeficiency virus type 1, and its overexpression affects virus assembly. *Journal of virology* **82**, 5683-5692, doi:10.1128/JVI.00189-08 (2008).

82 Engeland, C. E. *et al.* The cellular protein Lyric interacts with HIV-1 Gag. *Journal of virology* **85**, 13322-13332, doi:10.1128/JVI.00174-11 (2011).

83 Verreault, A., Kaufman, P. D., Kobayashi, R. & Stillman, B. Nucleosome assembly by a complex of CAF-1 and acetylated histones H3/H4. *Cell* **87**, 95-104 (1996).

84 Skowyra, D. *et al.* Differential association of products of alternative transcripts of the candidate tumor suppressor ING1 with the mSin3/HDAC1 transcriptional corepressor complex. *The Journal of biological chemistry* **276**, 8734-8739, doi:10.1074/jbc.M007664200 (2001).

85 Raghavendra, N. K. *et al.* Identification of host proteins associated with HIV-1 preintegration complexes isolated from infected CD4(+) cells. *Retrovirology* **7**, doi:Artn 66 10.1186/1742-4690-7-66 (2010).

86 Adachi, A. *et al.* Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *Journal of virology* **59**, 284-291 (1986).

87 Connor, R. I., Chen, B. K., Choe, S. & Landau, N. R. Vpr is required for efficient replication of human immunodeficiency virus type 1 in mononuclear phagocytes. *Virology* **206**, 935-944, doi:Doi 10.1006/Viro.1995.1016 (1995).

88 Becker, J. T. & Sherer, N. M. Subcellular localization of HIV-1 gag-pol mRNAs regulates sites of virion assembly. *Journal of virology*, doi:10.1128/JVI.02315-16 (2017).

89 Erde, J., Loo, R. R. O. & Loo, J. A. Enhanced FASP (eFASP) to increase proteome coverage and sample recovery for quantitative proteomic experiments. *Journal of Proteome Research* **13**, 1885-1895, doi:10.1021/pr4010019 (2014).

90 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature methods* **9**, 676-682, doi:10.1038/nmeth.2019 (2012).

91 Allouch, A. *et al.* HIV-1 acetylated integrase is targeted by KAP1 (TRIM28) to inhibit viral integration. *Retrovirology* **6** (2009).

92 Anderson, I. *et al.* Heat shock protein 90 controls HIV-1 reactivation from latency. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E1528-E1537, doi:10.1073/pnas.1320178111 (2014).

93 Ansari, S. A. *et al.* Interaction of YB-1 with human immunodeficiency virus type 1 Tat and TAR RNA modulates viral promoter activity. *Journal of General Virology* **80**, 2629-2638 (1999).

94 Braaten, D. & Luban, J. Cyclophilin A regulates HIV-1 infectivity, as demonstrated by gene targeting in human T cells. *Embo Journal* **20**, 1300-1309, doi:Doi 10.1093/Emboj/20.6.1300 (2001).

95 Caillet, M. *et al.* Rab7A is required for efficient production of infectious HIV-1. *Plos Pathogens* **7**, doi:ARTN e1002347
10.1371/journal.ppat.1002347 (2011).

96 Callahan, M. A. *et al.* Functional interaction of human immunodeficiency virus type 1 Vpu and Gag with a novel member of the tetratricopeptide repeat protein family. *Journal of virology* **72**, 5189-5197 (1998).

97 De Iaco, A. & Luban, J. Cyclophilin A promotes HIV-1 reverse transcription but its effect on transduction correlates best with its effect on nuclear entry of viral cDNA. *Retrovirology* **11**, doi:Artn 11
10.1186/1742-4690-11-11 (2014).

98 DeBoer, J. *et al.* Alterations in the nuclear proteome of HIV-1 infected T-cells. *Virology* **468**, 409-420, doi:10.1016/j.virol.2014.08.029 (2014).

99 DeBoer, J., Madson, C. J. & Belshan, M. Cyclophilin B enhances HIV-1 infection. *Virology* **489**, 282-291, doi:10.1016/j.virol.2015.12.015 (2016).

100 Dharan, A. *et al.* KIF5B and Nup358 cooperatively mediate the nuclear import of HIV-1 during infection. *Plos Pathogens* **12**, doi:ARTN e1005700
10.1371/journal.ppat.1005700 (2016).

101 du Chene, I. *et al.* Suv39H1 and HP1 gamma are responsible for chromatin-mediated HIV-1 transcriptional silencing and post-integration latency. *Embo Journal* **26**, 424-435, doi:10.1038/sj.emboj.7601517 (2007).

102 Dunn, S. J. *et al.* Identification of cell surface targets for HIV-1 therapeutics using genetic screens. *Virology* **321**, 260-273, doi:10.1016/j.virol.2004.01.010 (2004).

103 Faure, J. *et al.* ARF1 regulates Nef-induced CD4 degradation. *Curr Biol* **14**, 1056-1064, doi:10.1016/j.cub.2004.06.021 (2004).

104 Gadad, S. S. *et al.* HIV-1 infection induces acetylation of NPM1 that facilitates Tat localization and enhances viral transactivation. *Journal of Molecular Biology* **410**, 997-1007, doi:10.1016/j.jmb.2011.04.009 (2011).

105 Gallo, D. E. & Hope, T. J. Knockdown of MAP4 and DNAL1 produces a post-fusion and pre-nuclear translocation impairment in HIV-1 replication. *Virology* **422**, 13-21, doi:10.1016/j.virol.2011.09.015 (2012).

106 Gaudin, R., de Alencar, B. C., Arhel, N. & Benaroch, P. HIV trafficking in host cells: motors wanted! *Trends in Cell Biology* **23**, 652-662, doi:10.1016/j.tcb.2013.09.004 (2013).

107 Gautier, V. W. *et al.* In vitro nuclear interactome of the HIV-1 Tat protein. *Retrovirology* **6**, doi:Artn 47
10.1186/1742-4690-6-47 (2009).

108 Gordon-Alonso, M. *et al.* Actin-binding protein Drebrin regulates HIV-1-triggered actin polymerization and viral infection. *Journal of Biological Chemistry* **288**, 28382-28397, doi:10.1074/jbc.M113.494906 (2013).

109 Hearps, A. C. & Jans, D. A. HIV-1 integrase is capable of targeting DNA to the nucleus via an importin alpha/beta-dependent mechanism. *Biochemical Journal* **398**, 475-484, doi:10.1042/BJ20060466 (2006).

110 Jarboui, M. A. *et al.* Nucleolar protein trafficking in response to HIV-1 Tat: rewiring the nucleolus. *Plos One* **7**, doi:ARTN e48702
10.1371/journal.pone.0048702 (2012).

111 Le Sage, V., Cinti, A., Valiente-Echeverria, F. & Mouland, A. J. Proteomic analysis of HIV-1 Gag interacting partners using proximity-dependent biotinylation. *Virology Journal* **12**, doi:Artn 138
10.1186/S12985-015-0365-6 (2015).

112 Li, D. S., Wei, T., Abbott, C. M. & Harrich, D. The unexpected roles of eukaryotic translation elongation factors in RNA virus replication and pathogenesis. *Microbiology and Molecular Biology Reviews* **77**, 253-266, doi:10.1128/MMBR.00059-12 (2013).

113 Li, Y. & Belshan, M. NF45 and NF90 bind HIV-1 RNA and modulate HIV gene expression. *Viruses-Basel* **8**, doi:Artn 47
10.3390/V8020047 (2016).

114 Lingappa, J. R., Dooher, J. E., Newman, M. A., Kiser, P. K. & Klein, K. C. Basic residues in the nucleocapsid domain of gag are required for interaction of HIV-1 Gag with ABCE1 (HP68), a cellular protein important for HIV-1 capsid assembly. *Journal of Biological Chemistry* **281**, 3773-3784, doi:10.1074/jbc.M507255200 (2006).

115 Liu, Y., Belkina, N. V. & Shaw, S. HIV infection of T cells: actin-in and actin-out. *Science Signaling* **2**, doi:ARTN pe23
10.1126/scisignal.266pe23 (2009).

116 Lukic, Z., Dharan, A., Fricke, T., Diaz-Griffero, F. & Campbell, E. M. HIV-1 uncoating is facilitated by dynein and kinesin 1. *Journal of virology* **88**, 13613-13625, doi:10.1128/JVI.02219-14 (2014).

117 Marchand, V. *et al.* A janus splicing regulatory element modulates HIV-1 tat and rev mRNA production by coordination of hnRNP A1 cooperative binding. *Journal of Molecular Biology* **323**, 629-652, doi:10.1016/S0022-2836(02)00967-1 (2002).

118 Rojas-Araya, B., Ohlmann, T. & Soto-Rifo, R. Translational control of the HIV unspliced genomic RNA. *Viruses-Basel* **7**, 4326-4351, doi:10.3390/v7082822 (2015).

119 Santos, S., Obukhov, Y., Nekhai, S., Bukrinsky, M. & Iordanskiy, S. Virus-producing cells determine the host protein profiles of HIV-1 virion cores. *Retrovirology* **9**, doi:Artn 65 10.1186/1742-4690-9-65 (2012).

120 Schweitzer, C. J., Jagadish, T., Haverland, N., Ciborowski, P. & Belshan, M. Proteomic analysis of early HIV-1 nucleoprotein complexes. *Journal of Proteome Research* **12**, 559-572, doi:10.1021/pr300869h (2013).

121 Schweitzer, C. J. *et al.* Knockdown of the cellular protein LRPPRC attenuates HIV-1 infection. *Plos One* **7**, doi:ARTN e40537 10.1371/journal.pone.0040537 (2012).

122 Thierry, S. *et al.* High-mobility group box 1 protein induces HIV-1 expression from persistently infected cells. *Aids* **21**, 283-292, doi:Doi 10.1097/Qad.0b013e3280115b50 (2007).

123 Warren, K. *et al.* Eukaryotic elongation factor 1 complex subunits are critical HIV-1 reverse transcription cofactors. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 9587-9592, doi:10.1073/pnas.1204673109 (2012).

124 Watanabe, T. *et al.* The hematopoietic cell-specific Rho GTPase inhibitor ARHGDIB/D4GDI limits HIV Type 1 replication. *Aids Research and Human Retroviruses* **28**, 913-922, doi:10.1089/aid.2011.0180 (2012).

125 Widera, M. *et al.* An intronic G run within HIV-1 intron 2 is critical for splicing regulation of vif mRNA. *Journal of virology* **87**, 2707-2720, doi:10.1128/JVI.02755-12 (2013).

126 Yan, N., Cherepanov, P., Daigle, J. E., Engelman, A. & Lieberman, J. The SET complex acts as a barrier to autointegration of HIV-1. *Plos Pathogens* **5**, doi:ARTN e1000327 10.1371/journal.ppat.1000327 (2009).

127 Zhou, X. X. *et al.* DDX5 facilitates HIV-1 replication as a cellular co-factor of Rev. *Plos One* **8**, doi:ARTN e65040 10.1371/journal.pone.0065040 (2013).

128 Allouch, A. *et al.* The TRIM family protein KAP1 inhibits HIV-1 integration. *Cell Host & Microbe* **9**, 484-495, doi:10.1016/j.chom.2011.05.004 (2011).

129 Lassot, I. *et al.* The proteasome regulates HIV-1 transcription by both proteolytic and nonproteolytic mechanisms. *Molecular Cell* **25**, 369-383, doi:10.1016/j.molcel.2006.12.020 (2007).

3. HyPR-MS FOR MULTIPLEXED DISCOVERY OF THE MALAT1, NEAT1, AND NORAD lncRNA PROTEIN INTERACTOMES

This chapter is near submission to *Molecular Cell*.

Rachel A. Knoener[1,6,7], Michele Spiniello[1,7], Maisie I. Steinbrink[1,2], Bing Yang[3], Anthony J. Cesnik[1], Katherine E. Buxton[1], Mark Scalf[1], David F. Jarrard[2,3,4], Lloyd M. Smith[1,5,8]

[1] Department of Chemistry; University of Wisconsin – Madison; Madison, Wisconsin 53706; United States
[2] Molecular and Environmental Toxicology; University of Wisconsin – Madison; Madison, Wisconsin 53706; United States
[3] Department of Urology; University of Wisconsin School of Medicine and Public Health; Madison, Wisconsin 53705; United States
[4] Carbone Comprehensive Cancer Center; University of Wisconsin – Madison; Madison, Wisconsin 53792; United States
[5] Genome Center of Wisconsin; University of Wisconsin – Madison; Madison, Wisconsin 53706; United States
[6]Lead Contact
[7]These authors contributed equally
[8]Senior author

3.1 ABSTRACT:

RNA-protein interactions are integral to the regulation of gene expression. The protein interactomes of individual RNAs vary temporally, spatially, and within different physiological contexts. These factors in conjunction with the enormous variation in RNA and protein functions make the global acquisition of individual RNA-protein interactomes a lofty endeavor. Although techniques have been forthcomimg for discovery of the protein interactomes of specific RNAs, they are largely laborious, costly, and accomplished singly in individual experiments. We developed HyPR-MS, a multiplexing strategy for discovery of RNA-protein interactomes that utilizes a workflow that reduces design times and increases efficiencies to discover the protein interactomes of multiple RNAs, concurrently. Presented here is the application of HyPR-MS to simultaneously and selectively isolate the lncRNAs

MALAT1, NEAT1, and NORAD and elucidation their distinct protein interactomes. This platform provides a necessary tool for efficient and cost-effective elucidation of the complex and diverse population of RNA-protein interactomes.

3.2 INTRODUCTION:

RNA-protein interactions are crucial to a wide variety of biological processes involved in gene expression including transcription, nuclear organization, splicing, and translation [1]. The highly coordinated interplay between RNAs and proteins is vital to proper cell physiology, and its dysregulation causes many human diseases. Knowledge of the RNA-binding proteome in normal and diseased states is thus essential to the understanding of RNA biology and for the development of efficient therapies [2].

Numerous experimental strategies have been developed for the study of RNA-protein interactions and can be categorized as protein-centric or RNA-centric technologies. Protein-centric strategies isolate RNA-protein complexes using immunoprecipitation of a target protein followed by identification of the associated RNAs, primarily by RNA-sequencing. This strategy has been employed to map the binding locations of various proteins throughout the transcriptome [3]. RNA-centric methods have a different aim: to reveal the proteins associated with a target RNA. *In vitro* RNA-centric methods use synthetic RNA molecules as bait to isolate binding proteins from cell lysate but without regard to the physiological RNA-protein interactions [4]. *In vivo* RNA-centric approaches, like ChIRP-MS (comprehensive identification of RNA binding proteins by mass spectrometry) [5], CHART-

MS (capture hybridization analysis of RNA targets and mass spectrometry) [6], and RAP-MS (RNA antisense purification and mass spectrometry) [4] have become gold standards in the field because of their ability to reveal RNA-interacting proteins under specific spatiotemporal and biological contexts. The commonality of these *in vivo* strategies includes the exposure of cells to a protein crosslinker, capture of the target RNA by specific biotinylated capture oligonucleotides and, after elution, identification of the RNA-associated proteome by mass spectrometry [7]. These technologies have been used to identify the interacting proteomes of MALAT1 and NEAT1 lncRNAs in Michigan Cancer Foundation-7 (MCF7) cells and Xist lncRNAs in various cell types [4-6]. However, several unaddressed challenges with these technologies reduce their widespread application, including: (a) the capture oligonucleotide (CO) design limitations and (b) the costly and labor-intensive procedure [2].

Over the last two decades, lncRNAs have increasingly been studied for their role in cancer development and progression [8]. In particular, the dysregulation of lncRNAs MALAT1 and NEAT1 has been observed in a large spectrum of cancers, specifically in prostate cancer [9-12]. Both MALAT1 and NEAT1 are abundant, ubiquitously expressed, and nuclear-localized; they also share several interacting proteins, suggesting complementary roles in addition to their own unique functions [6,13,14]. MALAT1 (metastasis-associated lung adenocarcinoma transcript 1), also known as NEAT2 (nuclear-enriched abundant transcript 2), is an active regulator of alternative splicing through association with protein factors such as SRSF1, SC35, and SRSF3 [13]. It also interacts in a tissue-specific manner with other nucleic acids and proteins such as TDP-43 and AGO2 to control various aspects of gene expression, cell growth, synapse formation, and cell cycle [13,15]. NEAT1 (nuclear enriched abundant

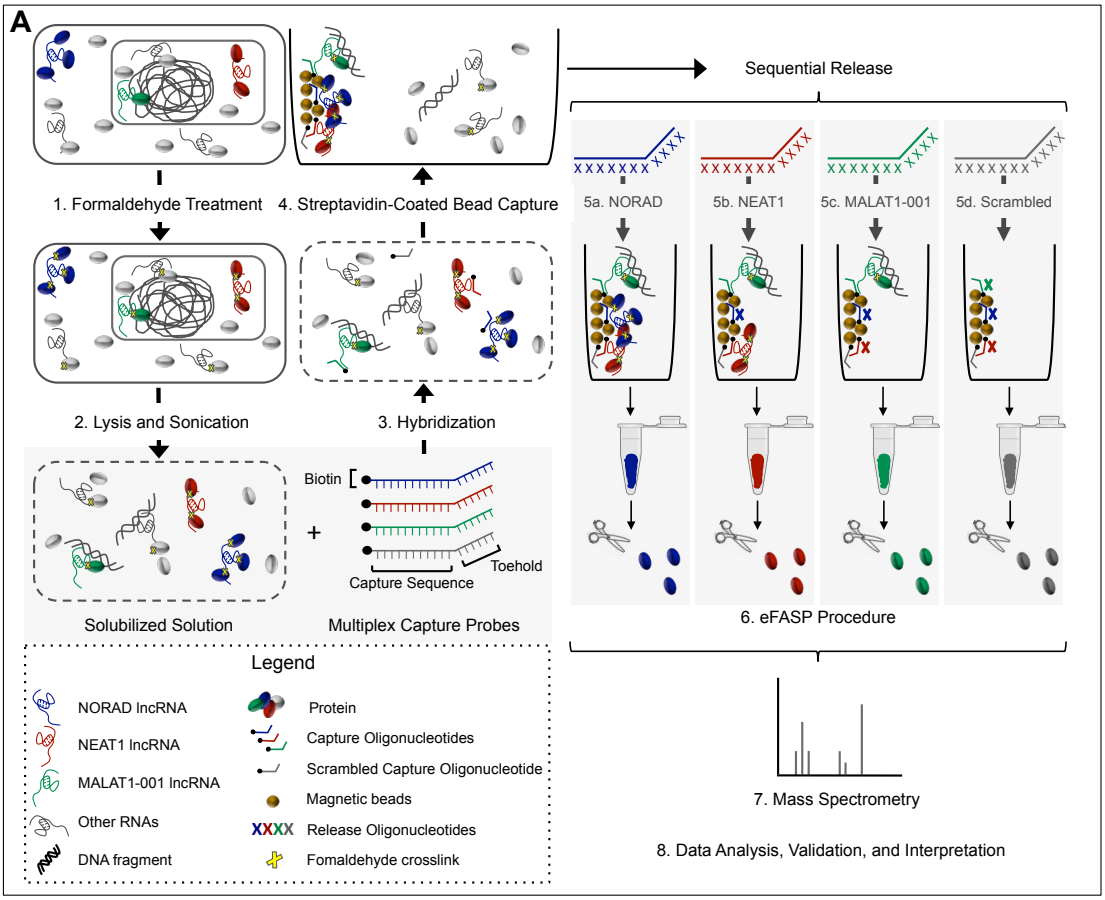
transcript 1), also known as Men (multiple endocrine neoplasia), is a necessary component of paraspeckles and its removal causes the disintegration of this nuclear body [16]. It is present as two isoforms; NEAT1_2 is essential for paraspeckle biogenesis, interacting with known paraspeckle components such as NONO and SFPQ while NEAT1_1 is later recruited with PSPC1 among other proteins to complete paraspeckle formation [10,14]. While the function of paraspeckles is unclear, they have been shown to be involved in cell differentiation, alternative splicing, and the cellular response to stress through sequestration of transcripts and proteins [17-20]. NEAT1 also acts independently of paraspeckles, regulating microRNA levels and epigenetic gene expression by sequestering RNAs using sponge-like activity, and through direct interaction with the chromatin [10].

NORAD (non-coding RNA activated by DNA damage) or LINC00657 is a highly expressed lncRNA in many cancers and ubiquitously present in normal tissue, suggesting a relevant cellular function [21,22]. NORAD has been called the "defender of the genome" due to its preservation of chromosomal stability; therefore, its dysregulation may implicate it in tumorigenesis [23]. Several manuscripts also discuss the molecular mechanism of NORAD regarding the induction of hypoxia in endothelial cells and other stress pathways [24,25]. Considering this, elucidation of the NORAD *in vivo* interacting proteome has high potential to be biologically important [23]. Currently, the documented primary NORAD interactors are the translational regulators Pumilio 1 and Pumilio 2 (PUM1 and PUM2) [22,24,26]. NORAD acts as a molecular decoy for PUM1 and PUM2 via several repetitive binding motifs within its sequence allowing the binding of at least fifteen Pumilio proteins per molecule. Several other NORAD-interactors, such as XRN2, IGF2BP1/2/3, PABPN1 have been hypothesized by

Tichon and colleagues to interact with NORAD based on *in vitro* RNA pulldown in a human osteosarcoma cell line (U2OS) [22].

In the present work we describe HyPR-MS (Hybridization Purification of RNA-protein-complexes followed by Mass Spectrometry) as a multiplexed *in vivo* RNA-centric method and apply it to discover, concurrently, the RNA-protein interactomes of three lncRNAs in the human prostate cancer cell line, PC3 [27]. Briefly, HyPR-MS utilizes biotinylated capture oligonucleotides specifically complementary to RNAs of interest to capture the target RNA-protein complexes from formaldehyde-crosslinked PC3 cell lysate. Following isolation of the capture oligonucleotide-target RNA hybrid using streptavidin-coated magnetic-beads, the RNA is released from the beads using a sequence-specific release oligonucleotide strategy that permits selective isolation of multiple RNA targets from a single lysate sample (Figure 1). Finally, the proteins associated with the RNA target are purified then trypsin digested and analyzed using mass spectrometry (MS). Here, HyPR-MS is applied to identify the protein interactomes of three lncRNAs, MALAT1-001 (specifically, the full length variant), NEAT1, and NORAD, in PC3 cells. The functions and interacting proteins of MALAT1 and NEAT1 have been the objects of numerous studies using various cell lines and therefore serve here, in part, as controls for evaluating the performance of the HyPR-MS technology. However, while many RNA-protein interactions may be common between different cell and tissue types, many may also be different due to variations in lncRNA, protein, and proteoform expression or in cellular contexts. [13,28-30]. Additionally, HyPR-MS is applied here to specifically capture only one splice variant of the 17 known variants of MALAT1 [31], allowing for the assignment of the protein interactome to only the full length MALAT1 transcript. Using HyPR-MS, 127, 94, and 415 interacting proteins were identified for

MALAT1-001, NEAT1, and NORAD, respectively. Of these proteins, several have been previously identified; these include SRSF and PRPF proteins for MALAT1-001 [9,32], NONO and SFPQ for NEAT1 [33], and PUM1 for NORAD [22,24]. The depth of coverage provided by HyPR-MS is demonstrated by discovery of many novel interactors with functions relating to known documented features of their associated lncRNAs, including several histone modifiers and transcriptional regulators found to interact with MALAT1 and regulators of gene expression associated with NORAD. The novel design workflow of HyPR-MS achieves high efficiency, specificity, and multiplexing capability; and the results shown here demonstrate the ability of HyPR-MS to discern the unique protein interactomes of multiple RNA species, simultaneously, from one lysate sample.



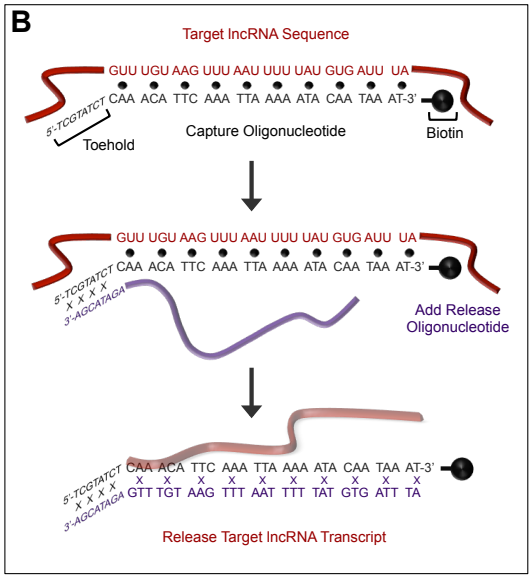


Figure 3.1: Multiplexed HyPR-MS Overview. A. Formaldehyde crosslinked cells are lysed and sonicated to solubilize the nucleic acid contents. Biotinylated capture oligonucleotides (COs), complementary to the RNA targets are added and allowed to hybridize. The resulting CO-RNA-protein complexes are combined with streptavidin-coated magnetic beads to immobilize and isolate the complexes. Once the beads are washed and resuspended in buffer, release oligonucleotides complementary to the COs of one target are added sequentially to elute each RNA-protein target species independently. Proteins from each resulting solution are purified and analyzed by mass spectrometry. B. The toehold mediated capture and release strategy allows for multiplexed isolation of RNA-protein complexes. The COs are designed so that 25-30 nucleotides (nt) are complementary to the RNA target and 8 nt are not complementary and remain unhybridized. Elution of the RNA-protein complexes occurs after adding the release oligonucleotide that is fully complementary to the CO. The capture and release oligonucleotide hybridization is thermodynamically more favorable than that of the CO and RNA target and the RNA-protein complexes are released into solution. An individual target RNA can then be released from the beads without disturbing the interactions of other RNA targets with the beads.

3.3 DESIGN:

HyPR-MS Overview and Strategy. While sharing the basic structure of previously developed RNA-centric strategies, HyPR-MS improves on their limitations through key differences in (a) capture oligonucleotide design (b) crosslinking and solubilization and (c) specific, multiplexed release of RNA-protein complexes from the beads (Tables 3.1 and S3.1). These alterations reduce the cost and labor of elucidating specific RNA-associated proteomes by improving efficiency and permitting multiplexing.

Many intertwined factors contribute to achieving the desired efficiency of purifying specific RNAs from a complex lysate while preserving *in vivo* RNA-protein interactions. Not least among them is the balance of crosslinking conditions and lysate solubilization. While formaldehyde crosslinking, *in vivo,* preserves a snap-shot of RNA-protein and protein-protein interactions, the parameters used for crosslinking can have great effect on the number of proteins crosslinked to the RNA of interest and thus the solubility of the crosslinked molecules as well as the accessibility of the RNA to hybridization with the capture oligonucleotide. Furthermore, the extent of crosslinking influences the optimal sonication parameters that allow for the solubilization of chromatin and other large biomolecules while still maintaining low fragmentation of the target RNA-protein complexes. These parameters can also ultimately affect the overall cost of the experiment; an increase in efficiency reduces the total amount of materials needed to capture sufficient protein for MS analysis. Another major contributor to the cost effectiveness of HyPR-MS is the design and implementation of the capture oligonucleotides. Using a publicly available program, M-fold [34], to predict the likely secondary structure of the target RNA, though likely less predictive than alternatively

used RNase H or SHAPE methods for determining accessible RNA regions, it is also much less time consuming and financially burdensome. Also, compared to the RNA tiling strategy that can require more than 40 capture oligonucleotides to capture a single RNA target, this strategy provides sufficient efficiency of capture so that only 2-3 capture oligonucleotides are needed (Figure S3.1A) thus reducing costs and allowing for the purification of specific isoforms of an RNA target. Finally, a significant design attribute of HyPR-MS is the use of a toe-hold capture and release strategy which allows for the capture and purification of multiple RNA species from a single cell lysate. This provides not only the experimental advantage of maintaining the same background for all RNA species captured, but also greatly reduces cost and time requirements. This strategy is highlighted in Figure 3.1B and Table S3.1.

Each of these steps was designed to improve upon the bottlenecks in current RNA-centric strategies for identification of RNA-binding proteins; the capture oligonucleotide (CO) design limitations, high cost of experiments, and lengthy experiment time requirements. Details on reasoning for each of these parameters are provided in Table S3.1.

A=Advantage L=Limitation	CHART	CHIRP	RAP-MS	HyPR-MS
Capture Oligonucleotide Design; # of Oligonucleotides	RNAse H assay 1 A: Know available sites for hybridization L: High time and money requirement; location bias	Full RNA tiling 43 A: No location bias L: Expensive; no isoform option	Full RNA tiling 142 A: No location bias L: Expensive; no isoform option	M-fold; Software 2-3 A: Less location bias, money, and time; isoform option L: Some potential location bias
Crosslinking Conditions	3% formaldehyde 30 minutes A: Preserves protein-protein interactions; efficient; widely used L: Potential fragmentation; reversal over time	3% formaldehyde 30 minutes A: Preserves protein-protein interactions; efficient; widely used L: Potential fragmentation; reversal over time	UV crosslink A: Specific to RNA-protein interactions; irreversible L: Low efficiency, therefore high false negatives	1% formaldehyde 10 minutes A: Preserves protein-protein interactions; efficient; widely used; less time and money than CHART and CHIRP; Parameters selected to optimize solubilization and hybridization stages of HyPR-MS L: Potential fragmentation; reversal over time
Lysate Solubilization; DNA fragment size	Sonication; 2-10 kb	Sonication; 200-500 bp	Sonication; unspecified	Sonication; 6 kb
Hybridization Conditions	1.3M urea, 800mM salt	10% formamide, 500mM salt	4M urea, 500mM LiCl	375mM LiCl
Elution Strategy	RNase H Digestion L: Can not multiplex	10 minutes at 65C in biotin elution buffer L: Can not multiplex	RNase and DNase Digestion L: Can not multiplex	Toehold-mediated release A: Able to multiplex to capture and isolate multiple RNA targets from one cell culture thus reducing financial and time costs as well as experimental variability.

Table 3.1: Comparison of Technologies.

3.4 RESULTS:

The HyPR-MS strategy was utilized to identify the protein interactomes of MALAT1, NEAT1, and NORAD in three biological replicates of PC3 cells. Each capture experiment used 1 x 10^8 cells and two to three capture oligonucleotides (Table S3.2 and Figure 3.2A) to meet the minimum capture efficiency threshold. This threshold was calculated so that, assuming one protein molecule per lncRNA molecule, a sufficient number of copies of a given protein would be present for mass spectrometric detection (Figure S3.1). For each replicate, small aliquots of the purified capture samples as well as lysate and post-release bead samples were analyzed using RT-qPCR to determine capture and release efficiencies, capture specificity, and fold enrichment of each lncRNA target.

3.4.1 Capture Efficiency. The capture efficiency (CE) of each lncRNA target was measured using three to four target specific qPCR assays that amplify regions dispersed along the length of the RNA target (Figure 3.2A and Table S3.3). The average CEs for MALAT1-001, NEAT1, and NORAD are approximately 8, 20, and 28%, respectively (Figure 3.2B). The capture efficiencies for all qPCR-measured regions range from 1 to 12% for MALAT1-001, 3-40% for NEAT1, and 17-35% for NORAD (Figure S3.2A). The CEs measured at different regions of each target are all above the estimated necessary capture efficiency thresholds. Notably, all CEs, with the exception of the 5'-NEAT1 region, are several-fold higher than their respective thresholds. This result suggests that even proteins with low-occupancy on the lncRNA could be detected by MS. Similarly, the relatively high capture efficiencies suggest that HyPR-MS can be applied to study the interacting proteomes of less abundant RNAs while still using a reasonable number of cells. To test this, the

HyPR-MS strategy was applied to capture the c-Myc mRNA, which is present in <100 copies per cell in human myelogenous leukemia cells (K562). Preliminary data shows that capture of c-myc from 1 x 10^8 K562 cells is sufficient for mass spectrometric detection of proteins (data not shown). This indicates a broad utility for HyPR-MS to identify the protein interactome of a variety of RNA targets while limiting the experimental cost.

3.4.2 Capture Specificity. The capture specificity of HyPR-MS is a measure of the capacity to capture the target lncRNA while avoiding the capture of non-target RNA and DNA molecules. Optimization of this parameter is crucial to the accurate identification of an RNA's interacting proteome. Capture specificity depends not only on the unique complementarity of the CO to the target RNA but also on the conditions in which they hybridize, the crosslinking and sonication parameters, the bead wash conditions and the release strategy. These conditions were optimized in concert to provide selectivity of lncRNA capture relative to non-target RNA and DNA. Capture specificity for each target is calculated using RT-qPCR to compare the amount of the target RNA in its corresponding capture sample to that in the capture sample of a different RNA target. The capture oligonucleotide specific to each lncRNA target captures 10-fold more of that lncRNA than does the oligonucleotide specific to a different lncRNA. For example, the NORAD capture sample contains nearly 47-fold more copies of the NORAD transcript than does the MALAT1-001 capture sample. Capture specificity was also demonstrated by conducting a capture experiment using a scrambled capture oligonucleotide. The scrambled CO was designed so that it is not complementary to any sequence in the PC3 transcriptome or genome and has a melting temperature (Tm) similar to that of the targeted COs (Table S3.2). The amount of each lncRNA target in its corresponding capture sample relative to

that in the scrambled capture sample is at least 10-fold higher for MALAT1, NEAT1, and NORAD (Figure 3.2D). These capture specificity results not only confirm the specificity of HyPR-MS but also support the use of the scrambled oligonucleotide capture sample as a negative control for evaluation of the lncRNA protein interactomes. As a broad measure of the quality and power of the HyPR-MS method, fold enrichment for each target is calculated compared to GAPDH RNA (an abundant housekeeping gene). To do this, the ratio of target-to-GAPDH in the captured sample over target-to-GAPDH in lysate is found using RT-qPCR (Figure 3.2E). The fold enrichment of each target is greater than 100-fold relative to GAPDH for all three target lncRNAs, supporting the results of the capture specificity and further suggesting that each captured sample is unique and independent of the others. Finally, we ensured that the capture samples were devoid of any DNA that corresponds in sequence to the lncRNA captured. Data in Figure S3.2B shows that all capture samples contain insignificant amounts of the analogous DNA sequences and thus the levels of any associated proteins are inconsequential. Taken together these data support the specificity of each capture, ensuring the uniqueness of the associated interacting proteome, and meet or surpass the performance of other RNA capture strategies [6].

3.4.3 Release Efficiency. The levels of capture efficiency and capture specificity in part rely in part on the specific and complete release of each individual lncRNA target from its hybrid interaction with the capture oligonucleotide-bead complex. This is accomplished by using a sequence-specific toehold release oligonucleotide to displace the capture oligonucleotide, thus releasing only the target RNA into solution. The release efficiency is measured using RT-qPCR analysis of the RNA released from the beads and of the RNA remaining on the beads following release. The percent release efficiencies from each capture

oligonucleotide range from 40-80% (Figure 3.2F). This result is sufficient, as is demonstrated by the level of overall capture efficiency and capture specificity; however, further optimization of the release parameters is desirable for future studies.

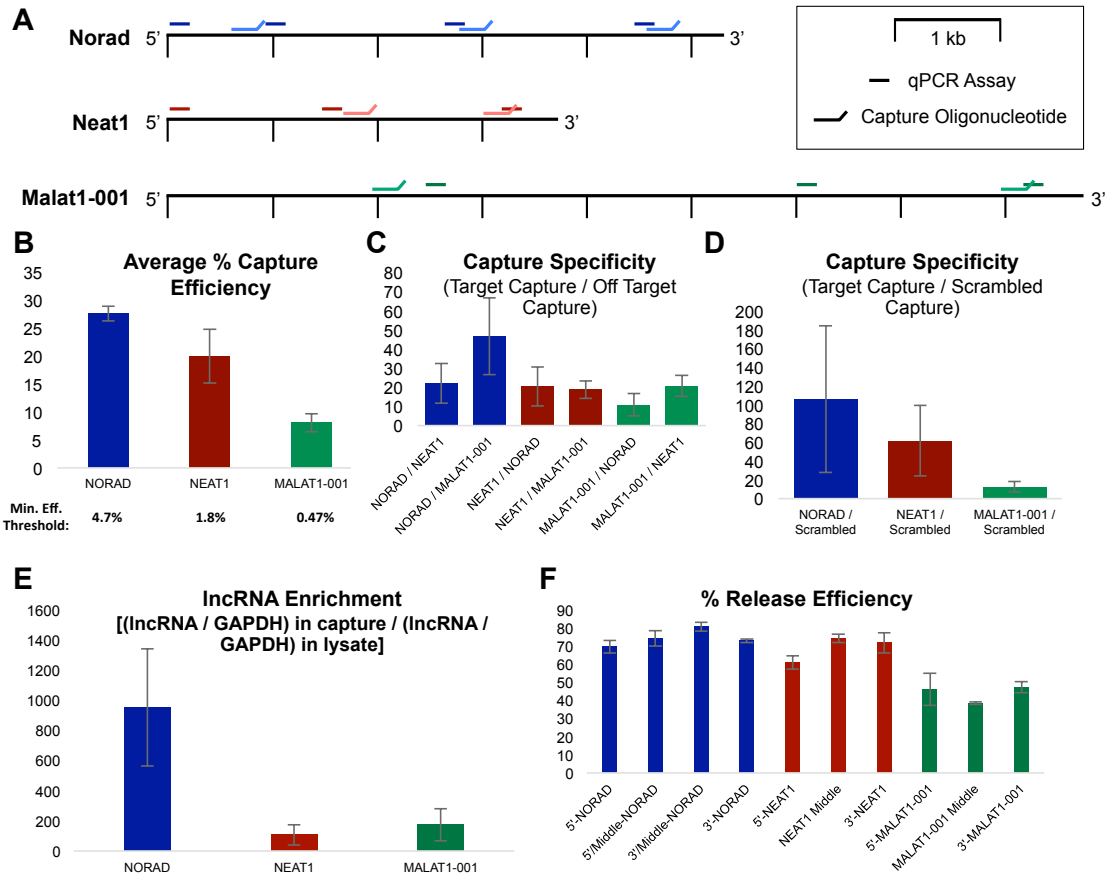


Figure 2. HyPR-MS Efficiencies and Specificities. A. Relative locations on the target RNA for CO hybridization and qPCR amplification. B. Average capture efficiency from all qPCR assays for each lncRNA target. C, D. Specificity determined from amount of target transcript captured using the designated COs relative to the amount of the same lncRNA captured using off target COs or scramble oligonucleotides. E. The fold enrichment of each target lncRNA. Calculated from the ratio of target RNA to housekeeping gene, GAPDH RNA in the capture sample compared to the lysate input sample. F. Efficiency of releasing each target from the magnetic beads with each CO-release oligonucleotide pair.

3.4.4 lncRNA Protein Interactome. The proteins isolated from the capture samples, as well as an aliquot of the complex lysate used as the input material for each capture, were analyzed by mass spectrometry. The resulting spectra for all three biological replicates were searched and quantified using the MaxQuant [35] platform. Statistical analysis as described in the method section was used to determine the proteins enriched in each capture sample relative to the scrambled oligonucleotide control and input samples (Details in SI). This analysis yielded 127, 94, and 415 proteins in the MALAT1, NEAT1, and NORAD interactomes, respectively. The interactome proteins for each lncRNA target were evaluated using STRING [36], an open-source software for identifying protein-protein interactions as well as calculation of GO (gene ontology) term enrichment. The enriched GO terms for each of the three interactomes are summarized in Table 2 and discussed below.

3.4.5 Each lncRNA-Protein Interactome is Enriched for Distinct Gene Ontology Terms. Each lncRNA interactome is highly enriched for RNA-binding proteins (p-values between 2E-19 and 6E-51) giving evidence that HyPR-MS identifies proteins interacting with specific RNAs. Also significant is that the enriched proteins in each interactome localized to the cellular components associated with the known locations of the lncRNAs themselves. Both MALAT1 and NEAT1 lncRNAs have been shown to localize to the nucleus, [32,37] and here both have interactomes enriched for proteins with nuclear localization, but not for proteins with cytoplasmic localization. Conversely, NORAD has been shown to localize primarily to the cytoplasm and the NORAD interactome data here show significant enrichment for cytoplasmic proteins (8E-27); however, they are also enriched for proteins with nuclear localization (3E-18). These nuclear proteins could be indicative of a novel

nuclear function for NORAD or could represent general RNA-processing functions required by all RNAs.

The MALAT1 interactome is enriched for components of the spliceosomal complex, while both MALAT1 and NEAT1 are enriched for nuclear bodies called nuclear speckles and paraspeckles (Table 3.2). These nuclear bodies are not sites of active transcription but are thought to be locations for the assembly and storage of splicing machinery [38]. Since NEAT1 and MALAT1 have been shown to have a role in the regulation of alternative splicing in different cell lines [13,17,32] and NEAT1 is a primary structural component of paraspeckles, these enrichments provide further support for the ability of HyPR-MS to correctly reveal specific RNA-interacting proteins.

The MALAT1 and NORAD interactomes are both enriched for proteins associated with nuclear export of RNA, although the proteins that fall into this category differ for each lncRNA. For example, the NORAD interactome is enriched for several proteins of the nuclear pore complex (NPC), which is the conduit for RNA export through the nuclear membrane. This is not surprising as NORAD is localized to the cytoplasm and therefore must exit the nucleus. Interestingly, XPO1 (or CRM1) was also enriched in the NORAD capture samples. This protein is required for the nucleocytoplasmic export of only a subset of proteins and RNAs and its overexpression has been implicated in poor prognosis of many cancer types [39]. In contrast, the annotated nuclear export proteins enriched in the MALAT1 interactome include proteins of the TREX complex and its associated proteins including SRSFs. The TREX complex links mRNA processing steps from transcription to splicing to nuclear export and its depletion in cells can cause genome instability leading to cancer [40].

These TREX and SRSF proteins may be functioning to co-transcriptionally splice and export MALAT1 into the cytoplasm or could be sequestered with MALAT1 in the nucleus in a nuclear body.

All three lncRNA interactomes in the prostate cancer cell line (PC3) are enriched for extracellular vesicle proteins (Table 3.2). Extracellular vesicles are involved in cell-to-cell communication and it has been shown that their lncRNA content can be altered by the tumor microenvironment [41]. Both MALAT1 and NEAT1 have previously been discovered to be present in extracellular vesicles produced by PC3 cells [42]. In addition, MALAT1 is known to be enriched in extracellular vesicles originating from cervical carcinoma and breast cancer cells [43] and NEAT1 is associated with extracellular vesicles in HepG2 human liver cancer (HepG2) and cholangiocarcinoma (Mz-ChA-1) cell lines [44]. These data are in accordance with enrichment of extracellular vesicles in the MALAT1 and NEAT1 interactomes. Interestingly, the NORAD interactome is more enriched for extracellular vesicles (p-value 2E-19) than are MALAT1 or NEAT1 (p-values 8E-6 and 3E-7, respectively). To our knowledge, NORAD has not been detected in extracellular vesicles produced by PC3 cells and this finding suggests that the lncRNA NORAD could be a significant component of such vesicles.

Category	Sub-category	GO Term	(-)log p-value		
			MALAT1	NEAT1	NORAD
Cellular Component	Nuclear Localization	Nucleoplasm	17.5	2.8	17.0
		Nuclear Speckles	30.3	10.0	7.4
		Spliceosomal Complex	15.4	NE	3.5
		Paraspeckles	2.0	2.4	NE
		Minichromosome Maintenance	NE	NE	2.4
	Cytoplasmic Localization	Cytoplasm	NE	NE	26.1
		Extrasomal Vesicle	5.1	6.5	18.7
Molecular Function	RNA-binding	RNA-binding proteins	50.6	18.8	50.2
		RS domain binding	1.9	NE	NE
Biological Process	Transcription	DNA-templated	2.1	NE	NE
		From RNA polymerase II promoter	1.8	NE	NE
	RNA splicing	mRNA splicing via spliceosome	30.7	3.2	8.0
		regulation of mRNA splicing	15.2	NE	NE
	RNA Transport	mRNA transport	8.1	NE	3.0
		RNA export from nucleus	7.5	NE	4.5
	Translation	Regulation of translation	NE	NE	8.6
	Others	Post-transcriptional regulation of gene expression	NE	NE	12.8
		Regulation of response to stress	NE	NE	2.9
		Mitotic cell cycle	NE	NE	6.7
		Regulation of RNA stability	NE	NE	2.0
		Regulation of gene silencing	NE	NE	1.8

Table 3.2: Gene Ontology Term Enrichment for Each lncRNA Interactome.

3.4.6 Hierarchical Clustering into a Heatmap Shows Protein Enrichment Differences in lncRNA Captures.

A clustering algorithm and heatmap visualization strategy was employed to interpret the lncRNA protein interactomes. A matrix of log_2-transformed, mean-centered protein intensities for each lncRNA biological replicate was compiled for all proteins enriched in at least one species of lncRNA capture. A correlation-uncentered, centroid-linkage hierarchical clustering algorithm [45] grouped the proteins based on the similarities between each protein's intensity-profile across the three lncRNA capture species and TreeView software [46] was used to visualize these similarities and differences. Figure 3.3A is the resulting heatmap; the nine columns represent the biological replicates for the three lncRNA captures and each row shows how the intensity of a protein compares to the mean intensity for all capture samples and biological replicates; a brighter red or yellow data-point indicates a greater difference from the mean intensity value. This strategy allows us to not only use the scrambled capture and input samples to determine which proteins are enriched in the lncRNA captures but also provides an additional layer of analysis that compares each capture to the different species of lncRNA capture. As can be seen in Figure 3.3A, this strategy is especially beneficial for identifying groups of proteins that are specifically enriched in one or two capture types.

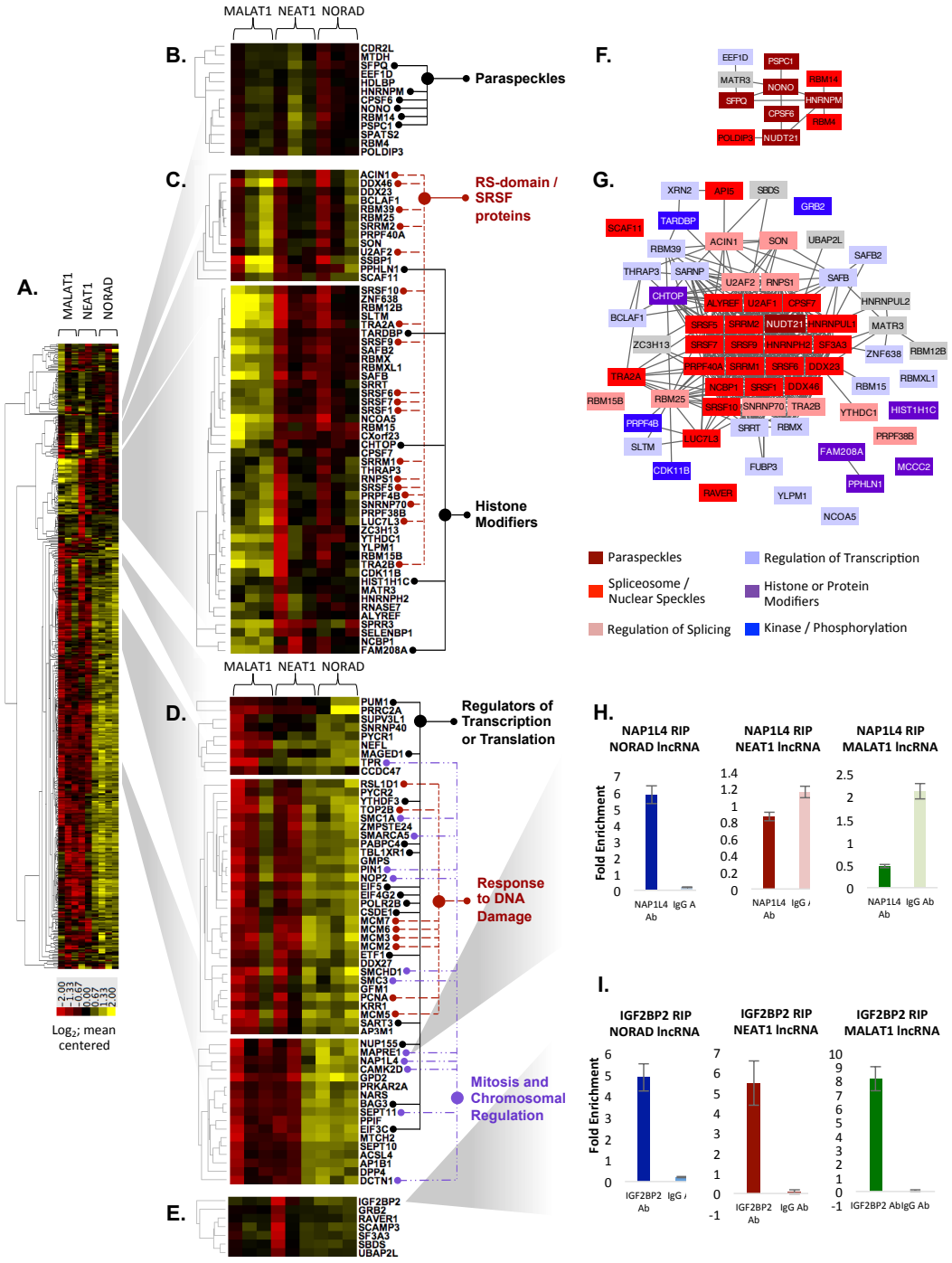


Figure 3.3: Hierarchical Clustering and Heat Map Visualization of lncRNA Interactomes. A. Heat map of all proteins enriched in the capture samples of at least one lncRNA species. B-D. Clusters contain proteins with emphasized enrichment in a particular lncRNA capture type. Highlighted in these clusters are proteins with known functions related to the function of the lncRNA they were found to interact with. E-F. STRING analysis of the proteins some clusters shows that often proteins with similar intensity profiles have known biological interactions. G. RIP-qPCR validation of select proteins confirms interaction with the lncRNAs they were found to interact with via HyPR-MS.

3.4.7 HyPR-MS Identifies Known and Novel lncRNA Interactors. The clustering algorithm subdivides the heatmap in Figure 3.3A into sections consisting of the proteins that are elevated in an individual lncRNA capture relative to the other captures. For example, the heatmap shows the clustering of over 40 proteins that are more abundant in the MALAT1 and NEAT1 captures than in the NORAD captures, over 70 proteins and 250 proteins that are more abundant in the MALAT1 and NORAD captures, respectively, relative to the opposing lncRNA captures, and over 50 proteins that are approximately equally abundant in all three lncRNA interactomes. Examples of annotated clusters demonstrating these trends are shown in Figures 3.3B through 3.3E.

NEAT1 is a core component of paraspeckles; satisfyingly, the GO term enrichment analysis for NEAT1 shows enrichment for paraspeckle protein components. Figure 3.3B shows a cluster of proteins with abundances elevated in NEAT1 and MALAT1 captures that includes six paraspeckle components. This cluster also includes seven additional proteins that share a similar intensity profile with the paraspeckle components. STRING analysis using experimental and database evidence for software settings of the proteins in this cluster shows that nine of the 13 proteins in the cluster are known to interact (Figure 3.3E). The remaining proteins in the cluster are transcriptional regulators or have known direct interaction with transcriptional regulators thereby making them potential candidates for further analysis for their role in lncRNA mechanisms.

Figure 3.3C shows clusters containing proteins with greater abundance in the MALAT1 capture relative to the NEAT1 and NORAD captures and Figure 3.3G shows the known interactions of the proteins in these clusters according to STRING analysis. Twenty proteins

enriched in the MALAT1 capture samples have been previously shown to interact with MALAT1 [6,13] with 15 of those grouped into the clusters shown in Figure 3.3C. Also among these clusters and enriched in the MALAT1 interactome are many nuclear speckle and splicing factors, including SR-splicing factors which contain RS-domains rich in serine (S) and arginine (R) residues. This is consistent with one proposed mode of action for MALAT1 that suggests MALAT1 regulates alternative splicing of endogenous RNAs by localizing SR splicing factors to nuclear speckles [32]. Notably, SR protein function is known to be affected by their phosphorylation states [32]; this makes the proteins with kinase or phosphorylation regulatory activities (CDK11, PRPF4B, GRB2, TARDBP) clustered among the MALAT1 interactors intriguing candidates for further functional analysis.

The clusters in Figure 3.3C also include many proteins involved in transcriptional regulation. Interesting among these are histone modifiers, including components of the HUSH complex (PPHLN1 and FAM208A) known for gene silencing through methylation of H3K9me3, CHTOP, an associate of the methylosome complex that methylates H4R3, and HIST1H1C which trimethylates H3K27 (Figures 3.3C and 3.3G). These findings, in addition to other transcriptional regulators identified by HyPR-MS (Figure 3.3G), are congruent with previous works that have demonstrated that lncRNAs in general [47], and MALAT1 specifically [6], localize to a subset of genomic sites and recruit proteins that affect gene expression.

The functions and modes of action of nuclear lncRNAs such as MALAT1 contrast with those of cytoplasmic lncRNAs like NORAD. Cytoplasmic lncRNAs are less well understood; however, they are thought to affect the activity or expression of the RNAs or proteins that they interact with. Two recent studies specifically investigated the interaction of

NORAD, a lncRNA whose expression is induced upon DNA damage [24], with PUM1 and PUM2, proteins that bind a specific subset of mRNAs and promote their degradation [22,24]. Both studies found that the abundant NORAD lncRNA acts as a molecular decoy capable of binding many molecules of these PUM proteins. These interactions in turn prevent the PUM proteins from negatively regulating the expression of their usual RNA targets. Satisfyingly, HyPR-MS identified PUM1 among the >400 NORAD interactors. Lee et al. also conducted NORAD RNA knockdown experiments to determine the effects of decreased levels of NORAD on gene expression. While the genes of 193 PUM protein targets were affected, the expression of over 1000 other genes were also affected [24], suggesting that NORAD has additional modes of action aside from PUM protein sequestration.

Because PUM1 was identified by HyPR-MS and binds to RNAs to regulate the expression of specific genes, we further searched our data for additional regulators of gene expression through RNA binding. Figure 3.3D shows three clusters containing protein with elevated abundances in the NORAD captures and highlights the transcription and translation regulators among these clusters. Among these are proteins with various known modes of action on specific subsets of RNA. For example, protein EIF3C is not only required for initiation of translation, it has also been found to bind to certain RNA stem-loop configurations to regulate specifically RNAs involved in cell growth control [48]. YTHDF3 binds to N^6-methyladenosine-modified RNA and has been found to work with YTHDF1 to promote translation or with YTHDF2 to promote RNA decay [49].

IGF2BP2 was enriched in all three of the lncRNA captures presented here with a higher relative enrichment in the NORAD samples (Figure 3.3E). IGF2BP2 binds to and recruits target RNAs to cytoplasmic mRNPs thus affecting the rate of translation and decay of the RNAs [50]. We selected IGF2BP2 as a target for validation using RNA-immunoprecipitation followed by qPCR analysis (RIP-qPCR). Figure 3.3I shows that following immunoprecipitation of IGF2BP2, all three lncRNAs were enriched relative to the negative controls, supporting that IGF2BP2 binds to MALAT1, NEAT1 and NORAD. Since MALAT1 and NEAT1 are nuclear localized and NORAD is cytoplasmically localized, the modes of action for their interactions with IGF2BP2 are likely different. Previous works have shown interaction of IGF2BP2 with other lncRNAs. For example, the interaction of IGF2BP2 with the lncRNA MyoD inhibits the translation of n-Ras and c-Myc, in this way regulating skeletal muscle differentiation [51,52]. In another study conducted on mesenchymal glioblastoma multiforme (GBM), the interaction of lncRNA HIF1A-AS2 with IGF2BP2 was shown to be fundamental to the expression of mRNA targets with pro-oncogenic roles, such as HMGA1 [53]**.** Finally, IGF2BP proteins are able to destabilize lncRNAs in cancer models, regulating them post-transcriptionally, as shown by the interaction between IMP1 and the lncRNA HULC (highly up-regulated in liver cancer) in hepatocellular carcinoma model [54]. Given its biological role in normal and tumoral models and its established association with lncRNAs, it is likely that the interaction between IGF2BP2 and MALAT1-001, NEAT1, and NORAD, specifically in PC3 cells and possibly in other systems, points to new molecular pathways to be ascertained by functional experiments.

The above mentioned NORAD-RNA knockdown study conducted by Lee et al. also showed that the absence of NORAD resulted in chromosomal instability and a high

occurrence of abnormal chromosome numbers in the cells [24]. Interestingly, GO term analysis for NORAD HyPR-MS results showed enrichment of proteins with the response to DNA damage, the mitotic cell cycle and the minichromosome maintenance complex (MCM) (Table 3.2). Figure 3D highlights many proteins identified to function in response to DNA damage correlating with the previous finding that NORAD expression itself is induced upon DNA damage. Among these are five components of the MCM which is involved in the regulation of DNA replication, ensuring its occurrence only once per cell [55]. NORAD has previously been implicated in the control of DNA replication through interaction with Pumilio proteins; the observation of its interaction with the MCM lends additional support to its role in this process. Also striking are the numerous proteins associated with mitosis and chromosomal regulation identified in the NORAD interactome (Figure 3.3D). One such protein, also clustered with PUM1 in the heatmap but only enriched in NORAD capture samples, is NAP1L4. NAP1L4 is described as a histone chaperone for its ability to bind core and linker histones and to transfer them from the cytoplasm to the nucleus where they are assembled into nucleosomes [56]. NAP1L4 localization and function is regulated by its phosphorylation state; the phosphorylated protein localizes in the cytoplasm in complex with histones during the G1 phase, while its dephosphorylated form initiates its transport into the nucleus at the start of S phase. At this phase, NAP1L4 is involved in chromatin assembly, a process coupled to DNA replication and DNA repair [57]. This is particularly interesting in relation to the finding of chromosomal abnormalities as a result of NORAD knockdown in cells. We validate the interaction of NAP1L4 with NORAD, and not with MALAT1 and NEAT1, using RIP-qPCR. Figure 3.3H shows the enrichment of NORAD transcripts upon immunoprecipitation of NAP1L4 compared to the negative control, but not the enrichment of MALAT1 or NEAT1. Considering NORAD is already described as a molecular decoy

for the PUM1 and PUM2 proteins [24], it is possible that it may have a similar role with regards to NAP1L4 or be involved in the regulation of the phosphorylation state of NAP1L4 as MALAT1 is for SR proteins (see MALAT1-001 interacting proteome section). Future studies are necessary to test these hypotheses.

3.4.8 HyPR-MS Identifies Epi-Factors and Prostate Cancer Markers. Six, two, and twenty-one of the MALAT1, NEAT1, and NORAD interacting proteins, respectively, are 'epi-factors' with involvement in epigenetic regulation of gene expression [58]. These findings agree with the defined involvement of lncRNAs in epigenetic mechanisms of gene expression regulation (Holoch and Moazed, 2015) and with the already ascertained epigenetic functions of MALAT1 and NEAT1 as previously discussed. The 22 epi-factors identified in the NORAD interactome suggest that NORAD, too, may function in epigenetic mechanisms for regulation of gene expression. Four protein interactors indentified in all three lncRNA interactomes have a known role in prostate tumorigenesis (AZGP1, S100A8, RBM25, SFN) and two are involved in the pathogenesis of cancer or tumors in general (LTF, SERPINB3) [58]. This information can be useful to elucidate the molecular mechanism used by these onco-factors, providing a starting point for future studies.

3.5 DISCUSSION

HyPR-MS was designed in response to several unaddressed challenges that have prevented the widespread application of *in vivo* RNA-centric methods for interactome discovery. While maintaining the basic structure of previous strategies we reworked each step to reduce the time and cost of experiments and empower multiplexing for concurrent analysis of several RNA species.

Compared with other strategies, HyPR-MS exhibits expedient and reliable oligonucleotide design via utilization of the publicly available M-fold software. The use of a multi-oligonucleotide capture strategy provides a high capture efficiency along the full length of each RNA target while still providing the flexibility to investigate the specific variants of an RNA. Less stringent crosslinking treatments and solubilization steps reduce target RNA fragmentation and improves efficiencies for capture. Finally, the toehold-mediated capture and release strategy permits the study of several targets within the same cell culture preparation, guaranteeing reduced costs and time requirements and reduced background and sample variability.

We utilized HyPR-MS to investigate the protein interactomes of MALAT1-001, NEAT1 and NORAD in the PC3 cell line. The functions and many interacting proteins of MALAT1 and NEAT1 have been previously studied in alternative cell lines and, in part, serve to provide markers for establishing the efficacy of the HyPR-MS method. In addition, the analysis of MALAT1 and NEAT1 protein interactomes in PC3 cells presented here provides novel discoveries to further the understanding of their function. The identification of the *in vivo*

NORAD interactome, a much less extensively studied lncRNA, adds additional depth to the pool of knowledge surrounding its function.

The centerpiece of the HyPR-MS strategy, the toehold-mediated release, allows for the isolation of multiple RNA targets and for extensive cost and time reductions but also provides the means for robust analysis downstream. Because the different RNA targets are isolated from one cell culture, the background for each sample is nearly identical. Additionally, because one can directly compare the interactomes of different RNAs, it can be ascertained which proteins are unique to or extensively enriched in a particular RNA capture sample. We demonstrate one strategy for assessing these data by using mean-centered normalization followed by clustering and heatmap visualization of proteins determined to be enriched in one or more species of lncRNA isolation. The strategy of using the scrambled capture oligonucleotide samples and lysate input samples for determination of protein enrichment was substantiated by validation of IGF2BP2 and NAP1L4, proteins that were assessed to interact with all three lncRNA species or only the NORAD lncRNA, respectively. The RIP-qPCR assays affirmed these specific conclusions and also validated that the data analysis strategy used was sound.

We have demonstrated that HyPR-MS is a powerful tool for identifying RNA interacting proteomes by reducing the technological complexity, cost, and time of the procedure and establishing its quality and flexibility. Preliminary proteomic data suggest that HyPR-MS is able to identify the interacting proteome of RNAs with less than 100 copies per cell while still using a reasonable number of cells. Looking forward, the new applications for HyPR-MS include comparisons between the interacting proteomes of individual RNA targets in

different physiological conditions, the study of less abundant RNAs, and the identification of post-translational modifications of RNA interactomes for a deeper understanding of RNA biology in physiological and pathological conditions.

3.6 LIMITATIONS:

HyPR-MS is the only known strategy that is able to concurrently analyze the interacting proteomes of multiple RNA targets from the same cell culture. However, the extent to which the multiplexing capabilities of HyPR-MS can be exploited requires further analysis. Additionally, the capture efficiency, release efficiency, and mass spectrometry sensitivity will define the number of cells required to discover the interactomes of low abundance RNAs. Preliminary data collected in our lab for identification of the protein interactome of c-Myc mRNA, an RNA with reportedly less than 100 copies per cell in K562 cells, suggests that HyPR-MS is applicable to RNAs of low abundance. However, the lowest reaches of HyPR-MS have yet to be tested.

3.7 SUPPLEMENTARY INFORMATION:

Table S1: Comparison of existing RNA-capture technologies to HyPR-MS.
(In reference to the Design section and Table 1; A=Advantage, L=Limitiation)

	CHART (West et al., 2014)	ChIRP (Chu et al., 2015)	RAP-MS (McHugh et al., 2015)	HyPR-MS
Capture Oligonucleotide Design	**RNAse H Assay** A: Able to design the capture oligonucleotide to hybridize to the optimal location of the target RNA using empirical evidence of where the single stranded regions are. L: Expensive and laborious. Using only one oligonucleotide for capture can result in a bias towards capturing fragments of the RNA that are close to the site of hybridization.	**Full RNA Tiling** A: Avoid location bias caused from using only one capture oligonucleotide by using many oligonucleotides that hybridize along the full length of the target RNA. L: Expensive. Unable to isolate different isoforms of the RNA because the full sequence is hybridized to the oligonulceotides.	**Full RNA Tiling** A: Avoid location bias caused from using only one capture oligonucleotide by using many oligonucleotides that hybridize along the full length of the target RNA. L: Expensive. Unable to isolate different isoforms of the RNA because the full sequence is hybridized to the oligonulceotides.	**Secondary-structure prediction software** A: Mfold software (Zuker et al., 2003) uses minimum free energy estimations to predict nucleotide regions with the highest probability to be single stranded. 2-3 oligonucleotides complementary to these regions produces sufficient capture efficiency along the length of the target RNA. Faster and less expensive than RNAse H Assay and Full RNA Tiling. Allows for isolation of isoforms. Reduced location bias than using only one oligonucleotide for capture. L: Secondary structure is only a prediction, not empirically determined.
Crosslinking Conditions	**3% formaldehyde, 30 min.** A: Identification of direct and indirect protein-RNA interactions critical to RNA processing (Schmitz et al., 2016). Applied with success in many applications such as PICh and HyCCAPP. L: Formaldehyde may cause single- and double-stranded breaks in nucleic acids (Grafstrom et al., 1984, Kondo et al., 1985) resulting in the target RNA potentially being fragmented. This is problematic if only one oligonucleotide is used for capture.	**3% formaldehyde, 30 min.** A: Identification of direct and indirect protein-RNA interactions critical to RNA processing (Schmitz et al., 2016). Applied with success in many applications such as PICh and HyCCAPP. L: Formaldehyde may cause single- and double-stranded breaks in nucleic acids (Grafstrom et al., 1984, Kondo et al., 1985) resulting in the target RNA potentially being fragmented. This is not problematic for Full RNA Tiling capture strategies.	**UV-crosslinking** A: Crosslinks only direct RNA-protein interactions which could be desirable in some applications. Milder solubilization and less target fragmentation risk. L: For discovery and understanding of RNA interactomes it is beneficial to find indirect interactors as their impact on processing is significant. (Schmitz et al., 2016). Low efficiency of crosslinking which could correspond to excessive false negatives.	**1% formaldehyde, 10 min.** A: Same advantages as those for CHART and ChIRP however the limited exposure to formaldehyde here reduces the risk of RNA fragmentation and also reduces the amount of sonication necessary for nucleic acid solubilization downstream which also reduces RNA fragmentation. This is critical for the use of only a few capture oligonucleotides to capture the full RNA sequence without using the Full RNA Tiling capture strategy.

	CHART (West et al., 2014)	ChIRP (Chu et al., 2015)	RAP-MS (McHugh et al., 2015)	HyPR-MS
Lysate Solubilization and DNA Fragment Size	**Sonication; 2-10 kilobases** A: Sonication fragments and solubilizes the DNA so that lncRNAs that are associated with the chromatin can be isolated from the full genome. L: Excessive sonication can cause target RNA fragmentation (see above).	**Sonication; 200-500 bases** A: Sonication fragments and solubilizes the DNA so that lncRNAs that are associated with the chromatin can be isolated from the full genome. L: Excessive sonication can cause target RNA fragmentation (see above).	**Sonication; unspecified** A: Sonication fragments and solubilizes the DNA so that lncRNAs that are associated with the chromatin can be isolated from the full genome. L: Excessive sonication can cause target RNA fragmentation (see above).	**Sonication; about 6 kilobases** A: Sonication fragments and solubilizes the DNA so that lncRNAs that are associated with the chromatin can be isolated from the full genome. Milder sonication is sufficient for HyPR-MS, because of reduced crosslinking, thus preserving the target RNA integrity. L: Excessive sonication can cause target RNA fragmentation (see above).
Elution Strategy	**RNAse H** L: only one RNA target per 10^8 cells.	**Heat and Biotin Elution** L: only one RNA target per 10^8 cells.	**RNAse and DNAse Digestion** L: only one RNA target per 10^8 cells.	**Toehold-Mediated Release** Allows for isolation of several RNA targets from the same 10^8 cells. Dramatically decreasing the time, cost, and background variability between each RNA target captured. See Figure 1B for mechanism.

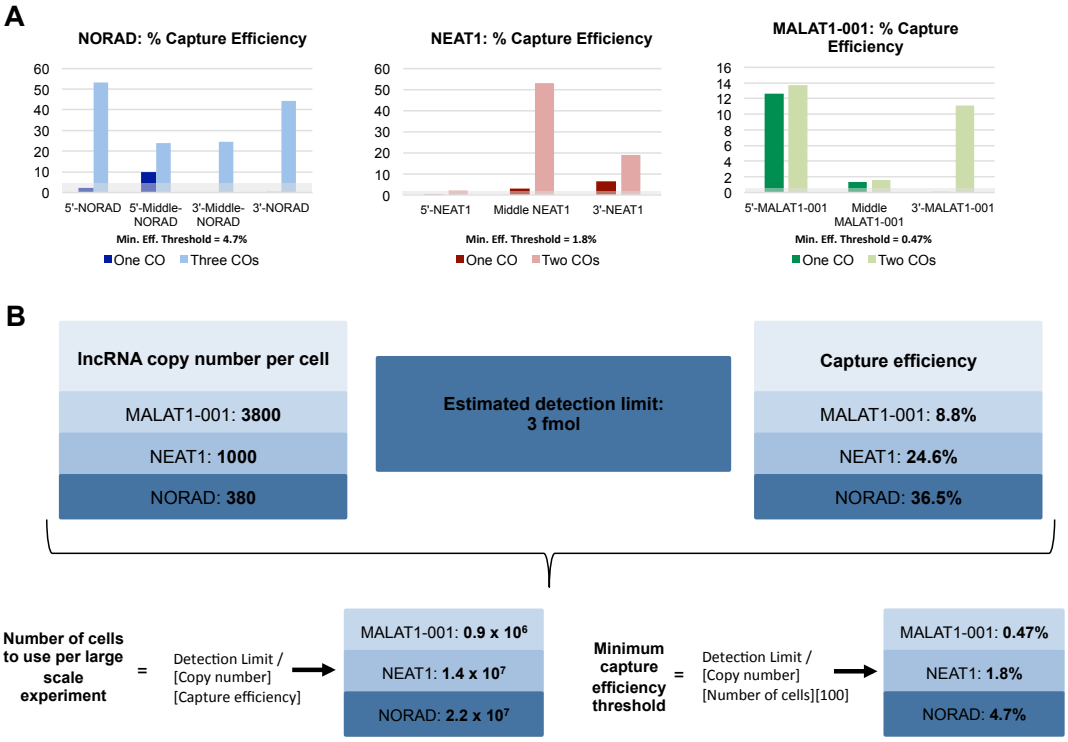


Figure S1: Determination of capture efficiencies and cell numbers required for protein detection. (In reference to the Design section and Figure 2A and 2B) A. Two to three COs were designed (Figure 2A) for each target and empirically tested for efficiency of capture. Capture efficiency (CE) is defined here as the percentage of the target lncRNA transcripts that were captured and released from the beads relative to the amount available for capture in the lysate sample. The amount of target in each sample was determined using multiple qPCR assays (Table S2) that amplify various regions along each target lncRNA; near the 5'-end, towards the middle, and near the 3'-end (Figure 2A). The capture efficiencies for each lncRNA target were determined using only one CO and using multiple COs. The results here show that when using only one CO per RNA target the qPCR amplification declines at regions further from the location of the CO. The data shows that using 2 or 3 capture oligonucleotides reduces this effect and that the capture efficiencies at each location along the RNA exceed the calculated minimum efficiency thresholds for each target. B. The number of cells necessary to obtain sufficient protein for one capture experiment was determined using the estimated copy number of each lncRNA transcript, the estimated detection limit for a protein in a complex biological mixture, and the capture efficiency of each lncRNA as determined from small scale experiments. Then, using the determined cell number, a minimum capture efficiency threshold was calculated for each target. This threshold is used to ensure that the actual capture efficiencies calculated from each qPCR measured region of the large scale capture experiments is sufficient to detect by MS even low abundance proteins. These efficiencies are indicated in the graphs in panel A as grey boxes. The multiple-CO captures provide sufficient copies of each target lncRNA to surpass the capture efficiency threshold established above, while still allowing the isolation of specific isoforms of a given lncRNA.

Table S2: Capture Oligonucleotide (CO) and Release Oligonucleotide (RO) Sequences and Concentrations Used. (In reference to the Results section and Figure 2A)

Target	Capture Oligonucleotide Sequence 5'-3' (toehold in red)	[CO] (in 25mL), Melting Temp (°C)	Release Oligonucleotide Sequence 5'-3' (toehold in red)	[RO] (in 25mL)
5`-MALAT1 - 001	TCGTATCTCAATATTTTCA TTTTCTATCTTGTTTCTAT	0.003 uM (60.5°C)	ATAGAAACAAGATAGAAA ATGAAAATATTGAGATAC GA	0.3 uM
3`-MALAT1 - 001	TCGTATCTCACCCAGCAT TACAGTTCTT	0.003 uM (59.0°C)	AAGAACTGTAATGCTGG GTGAGATACGA	0.3 uM
5`-NEAT1	TCGTATCTAAGTTATTTC ATCAGGCTAAGAA	0.014 uM (60.2°C)	TTCTTAGCCTGATGAAAT AACTTAGATACGA	1.4 uM
3`-NEAT1	TCGTATCTCAAACATTCA AATTAAAAATACAATAAAT	0.00714 uM (59.0°C)	ATTTATTGTATTTTTAATT TGAATGTTTGAGATACGA	0.714 uM
5`-NORAD	TCGTATCTAGCTTTTTCA TATTATATACACAG	0.00714 uM (56.8°C)	CTGTGTATATAATATGAA AAAGCTAGATACGA	0.714 uM
NORAD Middle	TCGTATCTAATGGTTATA TCTGATAGTGTCTT	0.00714 uM (58.8°C)	AAGACACTATCAGATATA ACCATTAGATACGA	0.714 uM
3`-NORAD	CAACTGTCACATCAATGG CTATCAAAATGTAAATAT GG	0.00714 uM (65.5°C)	CCATATTTACATTTTGATA GCCATTGATGTGACAGTT G	0.714 uM
Scrambled	CAACTGTCGCGTCTTTAT TTAGTTTACTCTTGATTGT T	0.003 uM (64.5°C)	AACAATCAAGAGTAAACT AAATAAAGACGCGACAGT TG	0.3 uM

Table S3: Primer and Probe Sequences for qPCR Assays. (In reference to Results section and Figure 2A).

Target	Forward 5'-3'	Probe 5'-3'	Reverse 5'-3'
5`-MALAT1-001	CAGGATTCCAGGAACCAGTG	CTAGGACTGAGGAGCAAGCGAGC	TTCCTATCTTCACCACGAACTG
MALAT1-001 Middle	GAACGAATGTAACTTTAAGGCAGG	TCCAGGCACATGGCAATAGAGGC	GATCATAATCTCCCACCTGTCTAAG
3`-MALAT1-001	ACGTATTGTTTTCTCAGGTTTTGC	AAAGATGCTGGTGGTTGGCACTC	GATTTGAACCCCGTCCTGG
5`-NEAT1	GCCTCCGGTCATACTAGTTTTG	CCTTGTAGATGGAGCTTGCAGATGGA	AGGTGGGTAGGTGAGAGG
NEAT1-Middle	CACCTAAAATCAGTTTGGAAAACAAG	CTCTCCCCACAATCCCCATCCC	AGGTGGGTAGGTGAGAGG
3`-NEAT1	CACCTAAAATCAGTTTGGAAAACAAG	CTCTCCCCACAATCCCCATCCC	ACATGTAGTAAAGGCACCTCG
5`-NORAD	TGGCTGTGCCCAGACCTT	CCACGGCCGCCATTAGTC	CAGCGAACCTCTCTTTCCCACCC
5`/Middle-NORAD	CACGTTTGTTAAGTGGGTTAGATG	ACATGGAGCTGGAAGACCTGAGAAG	AATATGACCAGTCTAGCATAGAACC
3`/Middle-NORAD	GACACGTGCCTATATCCATCAG	CCTTCCAACTCCTCTCCACCACC	CTTCTAAATACGAACATTCTGGTCTAG
3`-NORAD	TTGTTAAGCCACCTCTGAGC	TGCCAACCTAATGAACAAGTCCTGACA	CCTGTATAATTCCTTCTGCCCC

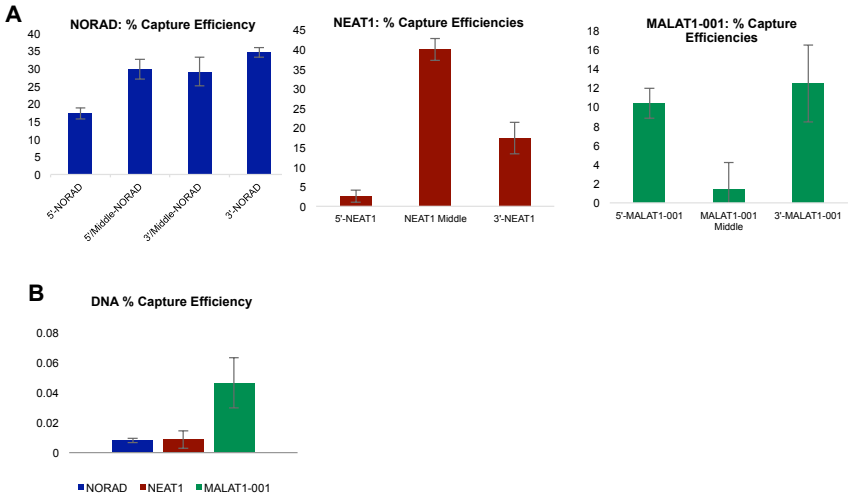


Figure S2: (In reference to Figure 2) **A. Capture efficiencies (CE) at each qPCR reagion for each target lncRNA.** It should be noted that there are differences in the CE measurement at different regions of the same lncRNA. These differences could be due to the secondary structure or protein occupancy at the region of CO complementarity in the lncRNA resulting in fewer captured molecules of the target RNA. The differences in measurements could also be due to variable efficiencies of the RT-qPCR assays in complex samples. Though the cause is undetermined, each qPCR-amplified region of each lncRNA exceeds the capture efficiency threshold, thus suggesting that proteins bound near those regions are captured at levels detectable by mass spectrometry. The variable captured efficiency observed is also an artifact of experimental design, which took into account not only the capture efficiency but also the cost of each experiment. More specifically, the capture efficiency of each capture oligonucleotide is dependent on the concentration of CO and target RNA in the hybridization reaction. Preliminary data showed that: 1) the optimal concentration of each CO is different for each lncRNA target and 2) within a certain concentration range, increasing concentrations of CO in the hybridization reaction increases the CE (data not shown). However, the number of streptavidin coated beads needed, and thus a large portion of the cost of the experiment, also increases as the number of biotinylated-CO increases. For this reason, the concentration of each CO added to the capture experiment was determined to ensure that the CE would surpass the calculated CE threshold, but not necessarily to maximize the CE (see Table S2 for concentrations used). **B. DNA capture efficiencies (CE)**. DNA CE is calculated for each target to test HyPR-MS capture specificity; each CO can hybridize not only the target lncRNA but also the analogous DNA sequence, meaning any proteins interacting with that region of the genome could be falsely identified as RNA interactors. As with the lncRNA CE (described above), DNA CE is found by measuring the number of DNA copies corresponding to the target RNA sequence in the captured sample and comparing it to that in the lysate. Based on DNA CE values that are consistently lower than 0.1%, the small total DNA copy number, and the mass spectrometry sensitivity (approximately 3 femtomole, or 1.8 x 10^9 copies), any contamination by DNA-interacting proteins is not able to affect the specificity of the RNA captures.

3.8 AUTHOR CONTRIBUTIONS

HyPR-MS development and optimization was initially performed by R.A.K with additional development and optimization conducted by M.Spiniello. and M.I.S. Study design and target selection was initiated by R.A.K, K.E.B. with assistance from A.J.C. and additional target selected by M.Spiniello and M.I.S. Cell cultures were maintained and supplied by B.Y. Implementation of HyPR-MS in PC3 cells, RT-qPCR experiments and data analysis, protein sample preparation, and RIP-qPCR analysis were conducted by M.Spiniello and M.I.S. Protein sample preparation and mass spectrometer maintenance and development were conducted by M.Scalf. R.A.K., M. Spiniello, and M.I.S. conducted data interpretation and manuscript writing. D.F.J. and L.M.S. provided support and advice during development and implementation. Special acknowledgement goes to Dr. Michael Shortreed and Dr. Brian Frey for early discussions of experimental design and input in data analysis.

3.9 REFERENCES:

1 Khalil, A. M. & Rinn, J. L. RNA-protein interactions in human health and disease. *Seminars in cell & developmental biology* **22**, 359-365, doi:10.1016/j.semcdb.2011.02.016 (2011).

2 Marchese, D., de Groot, N. S., Lorenzo Gotor, N., Livi, C. M. & Tartaglia, G. G. Advances in the characterization of RNA-binding proteins. *Wiley interdisciplinary reviews. RNA* **7**, 793-810, doi:10.1002/wrna.1378 (2016).

3 McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome biology* **15**, 203, doi:10.1186/gb4152 (2014).

4 McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232-236, doi:10.1038/nature14443 (2015).

5 Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404-416, doi:10.1016/j.cell.2015.03.025 (2015).

6 West, J. A. *et al.* The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Molecular cell* **55**, 791-802, doi:10.1016/j.molcel.2014.07.012 (2014).

7 Simon, M. D. Insight into lncRNA biology using hybridization capture analyses. *Bba-Gene Regul Mech* **1859**, 121-127, doi:10.1016/j.bbagrm.2015.09.004 (2016).

8 Martens-Uzunova, E. S. *et al.* Long noncoding RNA in prostate, bladder, and kidney cancer. *European urology* **65**, 1140-1151, doi:10.1016/j.eururo.2013.12.003 (2014).

9 Gutschner, T. *et al.* The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer research* **73**, 1180-1189, doi:10.1158/0008-5472.CAN-12-2850 (2013).

10 Lo, P. K., Wolfson, B. & Zhou, Q. Cellular, physiological and pathological aspects of the long non-coding RNA NEAT1. *Frontiers in biology* **11**, 413-426, doi:10.1007/s11515-016-1433-z (2016).

11 Wilusz, J. E. Long noncoding RNAs: Re-writing dogmas of RNA processing and stability. *Biochimica et biophysica acta* **1859**, 128-138, doi:10.1016/j.bbagrm.2015.06.003 (2016).

12 Yoshimoto, R., Mayeda, A., Yoshida, M. & Nakagawa, S. MALAT1 long non-coding RNA in cancer. *Biochimica et biophysica acta* **1859**, 192-199, doi:10.1016/j.bbagrm.2015.09.012 (2016).

13 Gutschner, T., Hammerle, M. & Diederichs, S. MALAT1 -- a paradigm for long noncoding RNA function in cancer. *Journal of molecular medicine* **91**, 791-801, doi:10.1007/s00109-013-1028-y (2013).

14 Zhang, Y., Yang, L. & Chen, L. L. Life without A tail: New formats of long noncoding RNAs. *Int J Biochem Cell B* **54**, 338-349, doi:10.1016/j.biocel.2013.10.009 (2014).

15 Wu, Y., Huang, C., Meng, X. & Li, J. Long Noncoding RNA MALAT1: Insights into its Biogenesis and Implications in Human Disease. *Current pharmaceutical design* **21**, 5017-5028 (2015).

16 Chujo, T., Yamazaki, T. & Hirose, T. Architectural RNAs (arcRNAs): A class of long noncoding RNAs that function as the scaffold of nuclear bodies. *Biochimica et biophysica acta* **1859**, 139-146, doi:10.1016/j.bbagrm.2015.05.007 (2016).

17 Cooper, D. R. *et al.* Long Non-Coding RNA NEAT1 Associates with SRp40 to Temporally Regulate PPARgamma2 Splicing during Adipogenesis in 3T3-L1 Cells. *Genes* **5**, 1050-1063, doi:10.3390/genes5041050 (2014).

18 Fox, A. H. & Lamond, A. I. Paraspeckles. *Cold Spring Harbor perspectives in biology* **2**, a000687, doi:10.1101/cshperspect.a000687 (2010).

19 Naganuma, T. *et al.* Alternative 3'-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *The EMBO journal* **31**, 4020-4034, doi:10.1038/emboj.2012.251 (2012).

20 Yamazaki, T. & Hirose, T. The building process of the functional paraspeckle with long non-coding RNAs. *Frontiers in bioscience* **7**, 1-41 (2015).

21 Ashouri, A. *et al.* Pan-cancer transcriptomic analysis associates long non-coding RNAs with key mutational driver events. *Nature communications* **7**, 13197, doi:10.1038/ncomms13197 (2016).

22 Tichon, A. *et al.* A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nature communications* **7**, 12209, doi:10.1038/ncomms12209 (2016).

23 Ventura, A. NORAD: Defender of the Genome. *Trends in genetics : TIG* **32**, 390-392, doi:10.1016/j.tig.2016.04.002 (2016).

24 Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164**, 69-80, doi:10.1016/j.cell.2015.12.017 (2016).

25 Michalik, K. M. *et al.* Long noncoding RNA MALAT1 regulates endothelial cell function and vessel growth. *Circulation research* **114**, 1389-1397, doi:10.1161/CIRCRESAHA.114.303265 (2014).

26 Mak, W., Fang, C., Holden, T., Dratver, M. B. & Lin, H. An Important Role of Pumilio 1 in Regulating the Development of the Mammalian Female Germline. *Biology of reproduction* **94**, 134, doi:10.1095/biolreprod.115.137497 (2016).

27 Kaighn, M. E., Narayan, K. S., Ohnuki, Y., Lechner, J. F. & Jones, L. W. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). *Investigative urology* **17**, 16-23 (1979).

28 Smith, L. M., Kelleher, N. L. & Consortium for Top Down, P. Proteoform: a single term describing protein complexity. *Nature methods* **10**, 186-187, doi:10.1038/nmeth.2369 (2013).

29 Yeger-Lotem, E. & Sharan, R. Human protein interaction networks across tissues and diseases. *Frontiers in genetics* **6**, 257, doi:10.3389/fgene.2015.00257 (2015).

30 Zhu, J. *et al.* Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-seq. *Scientific reports* **6**, 28400, doi:10.1038/srep28400 (2016).

31 Aken, B. L. *et al.* The Ensembl gene annotation system. *Database : the journal of biological databases and curation* **2016**, doi:10.1093/database/baw093 (2016).

32 Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell* **39**, 925-938, doi:10.1016/j.molcel.2010.08.011 (2010).

33 Sasaki, Y. T., Ideue, T., Sano, M., Mituyama, T. & Hirose, T. MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 2525-2530, doi:10.1073/pnas.0807899106 (2009).

34 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **31**, 3406-3415 (2003).

35 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* **13**, 2513-2526, doi:10.1074/mcp.M113.031591 (2014).

36 Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research* **45**, D362-D368, doi:10.1093/nar/gkw937 (2017).

37 Clemson, C. M. *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Molecular cell* **33**, 717-726, doi:10.1016/j.molcel.2009.01.026 (2009).

38 Hung, T. & Chang, H. Y. Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA biology* **7**, 582-585 (2010).

39 Ishizawa, J., Kojima, K., Hail, N., Jr., Tabe, Y. & Andreeff, M. Expression, function, and targeting of the nuclear exporter chromosome region maintenance 1 (CRM1) protein. *Pharmacology & therapeutics* **153**, 25-35, doi:10.1016/j.pharmthera.2015.06.001 (2015).

40 Heath, C. G., Viphakone, N. & Wilson, S. A. The role of TREX in gene expression and disease. *The Biochemical journal* **473**, 2911-2935, doi:10.1042/BCJ20160010 (2016).

41 Sato-Kuwabara, Y., Melo, S. A., Soares, F. A. & Calin, G. A. The fusion of two worlds: non-coding RNAs and extracellular vesicles--diagnostic and therapeutic implications (Review). *International journal of oncology* **46**, 17-27, doi:10.3892/ijo.2014.2712 (2015).

42 Ahadi, A., Khoury, S., Losseva, M. & Tran, N. A comparative analysis of lncRNAs in prostate cancer exosomes and their parental cell lines. *Genomics data* **9**, 7-9, doi:10.1016/j.gdata.2016.05.010 (2016).

43 Qian, Z., Shen, Q., Yang, X., Qiu, Y. & Zhang, W. The Role of Extracellular Vesicles: An Epigenetic View of the Cancer Microenvironment. *BioMed research international* **2015**, 649161, doi:10.1155/2015/649161 (2015).

44 Takahashi, K., Yan, I. K., Wood, J., Haga, H. & Patel, T. Involvement of extracellular vesicle long noncoding RNA (linc-VLDLR) in tumor cell responses to chemotherapy. *Molecular cancer research : MCR* **12**, 1377-1387, doi:10.1158/1541-7786.MCR-13-0636 (2014).

45 de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453-1454, doi:10.1093/bioinformatics/bth078 (2004).

46 Saldanha, A. J. Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**, 3246-3248, doi:10.1093/bioinformatics/bth349 (2004).

47 Bonasio, R. & Shiekhattar, R. Regulation of transcription by long noncoding RNAs. *Annual review of genetics* **48**, 433-455, doi:10.1146/annurev-genet-120213-092323 (2014).

48 Lee, A. S. Y., Kranzusch, P. J. & Cate, J. H. D. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature* **522**, 111-U292, doi:10.1038/nature14267 (2015).

49 Shi, H. *et al.* YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA. *Cell research* **27**, 315-328, doi:10.1038/cr.2017.15 (2017).

50 Nielsen, J. *et al.* A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. *Molecular and cellular biology* **19**, 1262-1270 (1999).

51 Degrauwe, N. *et al.* The RNA Binding Protein IMP2 Preserves Glioblastoma Stem Cells by Preventing let-7 Target Gene Silencing. *Cell reports* **15**, 1634-1647, doi:10.1016/j.celrep.2016.04.086 (2016).

52 Gong, C. *et al.* A long non-coding RNA, LncMyoD, regulates skeletal muscle differentiation by blocking IMP2-mediated mRNA translation. *Developmental cell* **34**, 181-191, doi:10.1016/j.devcel.2015.05.009 (2015).

53 Mineo, M. *et al.* The Long Non-coding RNA HIF1A-AS2 Facilitates the Maintenance of Mesenchymal Glioblastoma Stem-like Cells in Hypoxic Niches. *Cell reports* **15**, 2500-2509, doi:10.1016/j.celrep.2016.05.018 (2016).

54 Hammerle, M. *et al.* Posttranscriptional destabilization of the liver-specific long noncoding RNA HULC by the IGF2 mRNA-binding protein 1 (IGF2BP1). *Hepatology* **58**, 1703-1712, doi:10.1002/hep.26537 (2013).

55 Lei, M. The MCM complex: its role in DNA replication and implications for cancer therapy. *Current cancer drug targets* **5**, 365-380 (2005).

56 Rodriguez, P. *et al.* Functional characterization of human nucleosome assembly protein-2 (NAP1L4) suggests a role as a histone chaperone. *Genomics* **44**, 253-265, doi:10.1006/geno.1997.4868 (1997).

57 Rodriguez, P., Pelletier, J., Price, G. B. & Zannis-Hadjopoulos, M. NAP-2: Histone chaperone function and phosphorylation state through the cell cycle. *J Mol Biol* **298**, 225-238, doi:Doi 10.1006/Jmbi.2000.3674 (2000).

58 Medvedeva, Y. A. *et al.* EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database-Oxford*, doi:ARTN bav067 10.1093/database/bav067 (2015).

59 Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973, doi:10.1126/science.1237973 (2013).

4. DISCOVERY OF DIFFERENTIAL PROTEIN INTERACTOMES OF HIV-1 SPLICE VARIANTS.

This chapter is part of a future publication.

Rachel A. Knoener[*], Jordan T. Becker[‡], Mark Scalf[*], Nathan M. Sherer[‡], Lloyd M. Smith[*◊]

*Department of Chemistry, University of Wisconsin, Madison, Wisconsin, United States
‡McArdle Laboratory for Cancer Research and Institute for Molecular Virology, University of Wisconsin, Madison, Wisconsin, United States
◊Genome Center of Wisconsin, University of Wisconsin, Madison, Wisconsin 53706, United States

4.1 ABSTRACT:

Complex networks of transient RNA-protein interactions drive HIV-1 gene expression and replication. Though the interactions of various proteins with the viral genomic RNA have been individually characterized, a comprehensive view of the HIV splice variant-protein interactome is lacking. We present here a strategy for the isolation of the three classes of HIV splice variants from HIV infected Jurkat cells and identification of the protein interactomes for these full-length, partially spliced, and completely spliced HIV RNAs. Our analysis reveals differential interactomes for the distinct splice-forms and reveals potential candidates for further functional validation.

4.2 INTRODUCTION:

The HIV-1 genome is approximately 9.7 kilobases (kb) in length yet encodes nine proteins or polyproteins. This variety of protein products from a relatively modest-sized genome is made possible by the alternative splicing of one primary transcription product. As many as 109 different spliced HIV RNA transcripts have been identified by RNA sequencing, which

fall into 3 classes: unspliced or full-length (FL), partially spliced (PS), and completely spliced (CS) transcripts[1-3]. The full-length transcript, which retains both introns, codes for the polyproteins Gag and Gag-Pol and also serves as the genome for nascent virions. The partially spliced transcripts, approximately 4kb in length, have one excised intron and code for the Env protein and accessory proteins Vpu, Vpr, and Vif. The completely spliced transcripts, approximately 2kb in length, have both introns removed and code for viral regulatory proteins Nef, Rev and Tat (Figure 4.1A). As these splice variant classes each have unique characteristics and purpose, they utilize differing RNA-protein interactions to lead to the balanced expression of both the splice variants themselves and their protein products.

HIV RNA-protein interactions are pivotal regulators of HIV splice variant production. Figure 4.1B depicts a simplified visual of the complex processing steps of the various HIV RNA's, from transcription to protein production and virion formation. Core splicing signals including the 5'-splice sites (5'ss) and the 3'-splice sites (3'ss) can have strong or weak interactions with the splicing machinery thus influencing the efficiency of splicing at individual splice sites. In addition, *cis*-elements near the splice sites, including exonic splicing enhancers or silencers (ESE and ESS, respectively) and intronic splicing enhancers or silencers (ISE and ISS, respectively), interact with *trans*-elements such as regulatory proteins to facilitate or hinder splicing. For example, it has been shown that serine-arginine-rich (SR) proteins bind preferentially to ESEs to promote splicing and heterogeneous nuclear ribonucleoproteins (hnRNPs) can interact with ESSs to prevent splicing[4]. Interestingly, these same classes of proteins can have the respective converse effects on splicing when they interact with intronic *cis*-elements (ISE or ISS)[5]. These general splicing trends are also evident in HIV RNA splicing. A particularly well-studied splice site, 3'ssA7, is that which

contributes to the production of the Tat and Rev mRNAs. The splicing activity at this splice acceptor site has been shown to be inhibited by the interaction of hnRNPA1 with the *cis*-elements ESS3 and ISS[6,7]. Relatedly, another splicing event required for the formation of Tat mRNA is enhanced by the interaction of SR proteins SRSF2 and SRSF6 with multiple ESE elements located downstream from the 3'ssA3[8]. The complex interplay of such HIV RNA-protein interactions, many still not understood, is paramount to the balanced expression of the HIV splice variants over the course of HIV replication.

Following splicing, RNA-protein interactions continue to carry the different HIV splice variants through distinct pathways to achieve optimal expression of their protein products. Initially, the completely spliced transcript is the primary HIV RNA splice product and follows the canonical cellular NXF/NXT nuclear export pathway to translation in the cytoplasm. Protein product Nef is co-translationally N-myristoylated and translocated to the cellular membrane[9,10] while Rev and Tat are translocated to the nucleus where they serve as requisite regulatory factors for the efficient transcription, splicing, and nuclear export of the full length and partially spliced transcripts.

Typically, any RNA transcript in the nucleus that still contains introns will be spliced to completion or marked for degradation. The intron-containing HIV RNA splice variants, the full length and partially spliced transcripts, interact with viral and host proteins to avoid these fates. The Rev response element (RRE) located in the retained intron of the transcripts interacts with protein Rev leading to the expedited export of the transcripts via the Rev-mediated nuclear export pathway[11,12]. It has been determined that many cellular proteins including the RNA helicase MOV10[13] and RNA-binding protein STAU2[14] interact

with Rev and/or the HIV RNA to mediate the export of the intron-containing transcripts. This alternative export pathway helps these incompletely spliced transcripts to avoid the surveillance and degradation machinery associated with NXF/NXT nuclear export and to reach the cytoplasm for translation.

The cytoplasmic unspliced HIV RNA serves two purposes in HIV replication: it is the template for its protein products (Gag and Gag-Pol), it is the genome for nascent virions, or it could be both. Details regarding this complex interplay are elusive but it is evident that RNA-protein interactions are crucial to the cytoplasmic processing of the unspliced transcript. In particular, protein interactions with various regions within the 5'-UTR, a highly structured untranslated region of the unspliced HIV transcript, can influence translation or genome packaging[15]. RNA helicase A (RHA) has been shown to interact with the 5'-UTR of unspliced HIV RNA and promote the association of polysomes[16]. Conversely, HIV protein Gag binds to the packaging signal (psi) found in the 5'-UTR and promotes the packaging of the unspliced, genomic HIV RNA into virus particles [17-19]. These ribonucleoprotein complexes (RNPs) containing the genomic HIV RNA and Gag protein have been found to also contain host protein Staufen1 whose presence appears to influence particle formation[20]. This RNP has been further characterized and is estimated to contain over 200 different host proteins[21]. The cytoplasmic processing of full length HIV RNA is a highly regulated and specific process that requires many host proteins to efficiently produce Gag and Gag-Pol as well as infectious particles.

Clearly, HIV RNA-protein interactions are integral to HIV replication, from transcription of the nascent RNA to its eventual translation and packaging. The full length genomic HIV

RNA is a fundamental component of HIV replication and therefore, understandably, has been a focal point of study. The processing of the partially spliced and completely spliced transcripts, however, also influences the efficiency of HIV replication and therefore warrants study. For example, the partially spliced transcript must recruit the splicing machinery to excise the first intron while avoiding the removal of the RRE containing intron. RNA-protein interactions likely regulate this process. Furthermore, protein Tat, a product of completely spliced transcripts, has been shown to have a concentration dependent effect on apoptosis, requiring temporal regulation of its expression[22]. These examples demonstrate that RNA-protein interactions regulate the expression of all three HIV splice variant classes and their protein products in order to achieve the necessary balance required for efficient HIV replication. Identification of the host protein actors involved in this regulation is essential to gaining insight into such processes.

Appreciable progress has been made in the development of strategies to discover host factors that interact with HIV RNA. Studies have inserted MS2 binding sequences into HIV RNA constructs and, following isolation of the RNA using MS2 protein affinity purification, identified the proteins associated with the segments of HIV RNA. Variations of this strategy have lead to the discovery of MATR3[23] and hnRNPK[24] as pertinent host factors for HIV replication. However, these applications focused on only specific regions of the HIV RNA sequence and could be prone to artifactual interactions due to the inserted MS2 binding sites or altered secondary structures. Recently, we reported an alternative strategy, Hybridization Purification of RNA-protein-complexes followed by Mass Spectrometry, or HyPR-MS, for identification of full-length HIV RNA-protein interactors. This strategy utilized sequence-specific hybridization of biotinylated DNA oligonucleotides to the full

length HIV-RNA, isolation from a complex lysate using streptavidin coated magnetic beads, and identification of the associated proteins by mass spectrometry to discover nearly 200 proteins in the HIV-RNA *in vivo* interactome. This strategy can be applied to native HIV virus or virtually any modified HIV virus to study perturbations made to the system. However, only the interactome of the full-length variant of the HIV RNA could be detected in this initial study. In the present work we have expanded the HyPR-MS strategy described above to isolate not only the full length HIV RNA, but also, the partially spliced and completely spliced transcripts, in order to obtain a comprehensive view of the interactomes of each of these three important HIV RNA splice forms.

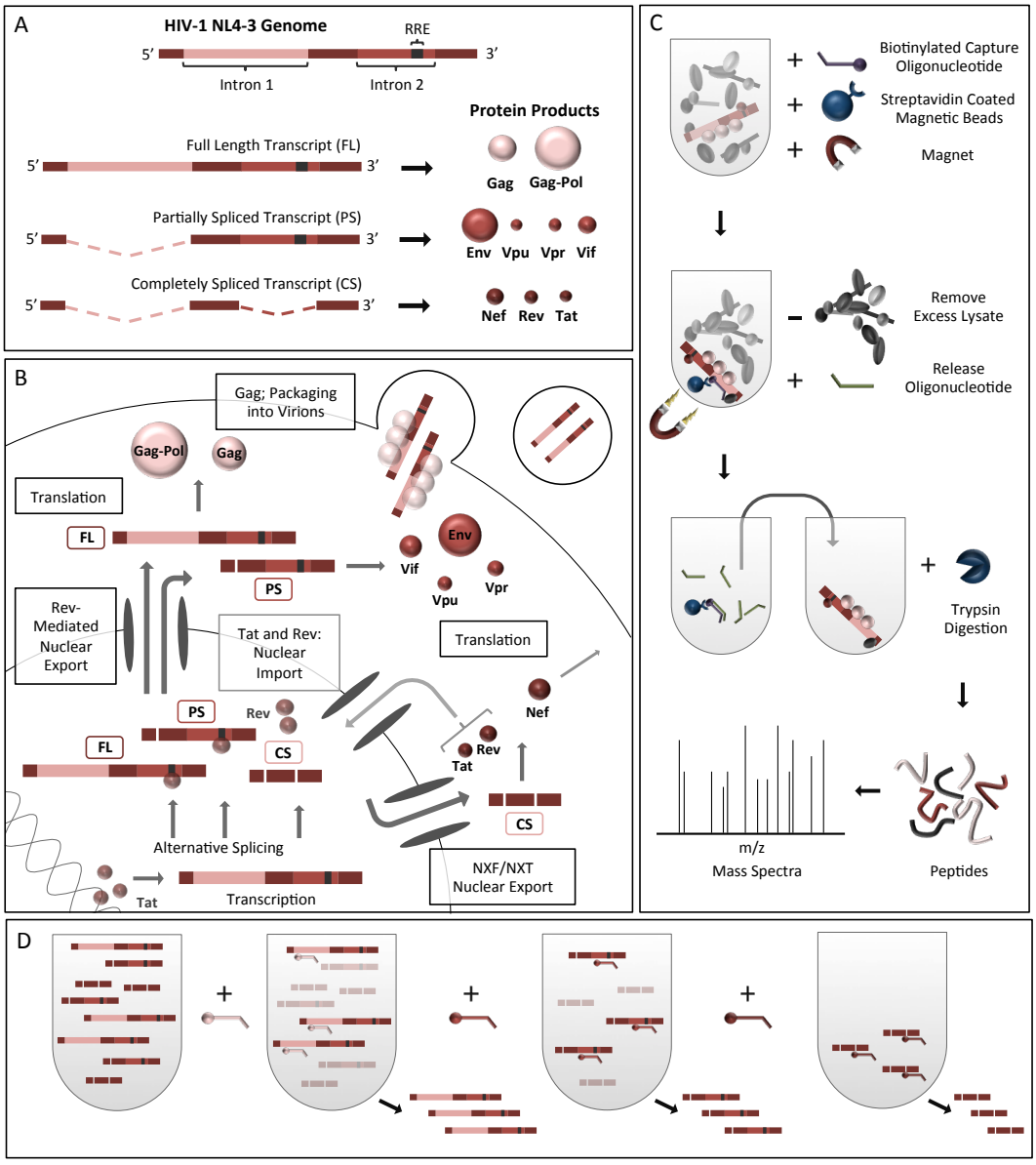


Figure 4.1: A. Mapping of the HIV-1 major classes of splice variants and their protein products. B. Diagram of late stages of HIV replication. C. Overview of HyPR-MS technology. D. Schematic of splice variant release from beads to achieve purification.

4.3 RESULTS

4.3.1 Workflow for the Purification of HIV-1 Splice Variants and Protein Interactomes.

To identify *in vivo* HIV-1 splice variant protein interactomes we adapted the HyPR-MS strategy (ref) to separately purify all three classes of HIV-1 splice variants: the full length unspliced (FL), the partially spliced (PS), and the completely spliced (CS) transcripts. This strategy utilizes the specific hybridization of a biotinylated DNA capture oligonucleotide to an RNA of interest followed by isolation of the RNA-protein complexes using streptavidin coated magnetic beads and identification of the proteins by mass spectrometry (Figure 4.1C).

To adapt this strategy to the isolation of the three HIV splice variant classes, the intricate nature of HIV RNA splicing must be considered. The PS splice variant class contains only nucleotide sequences that are also subsets of the FL splice variant; likewise, the CS variants contain only subsets of the PS and FL splice variants' sequences (Figure 4.1A). Thus, a capture oligonucleotide can only be unique to the PS or CS classes if it is complementary to their respective splice junctions. However, up to 109 different HIV splice variants have been detected in infected cells, all with varying splice junctions and all with temporally dynamic expression levels[1-3]. Therefore, to identify the comprehensive interactome of each splice variant class separately, the capture oligonucleotides for the PS and CS classes must be complementary to sequences that are present in all PS variants and CS variants, respectively. The obstacle of sequence similarity, as demonstrated in Figure 4.1A, is addressed by sequential isolation of the FL, PS, and CS splice variants, as is depicted in Figure 4.1D. For the study presented here, the FL transcript is first isolated using a capture oligonucleotide

that is complementary to a region of the retained intron that is not found in the PS or CS transcripts. Once the FL transcripts are removed from the lysate, the PS variants can be isolated using a capture oligonucleotide specific to its now unique intronic region. Finally, the CS transcripts, now the only HIV variant class remaining in the lysate, can be isolated by specific hybridization. The sequences of the designed capture oligonucleotides and their target locations on the HIV genomic sequence are indicated in Table S4.1.

4.3.2 Experimental Considerations. We sought to apply this capture strategy to identify proteins that interact with HIV RNAs, *in vivo*, and to elucidate proteins that may assist in functions unique to each splice variant class. To achieve this, three determinations for experimental design included: cell type, virus strain, and post-infection time-point for fixation. Jurkat cells, a commonly used T-cell line for HIV study, were selected as host cells not only because they largely recapitulate a cellular environment typical of HIV infection (CD4+ T-cells), they are also amenable to growth in large numbers. An HIV-1 molecular clone, NL4-3, bearing inactivating mutations in env and vpr, and expressing a cyan fluorescent protein (CFP) reporter from the nef reading frame (HIV-1 E-R-CFP), was chosen for infection of the Jurkat cells[25,26]. The inactivating mutations allowed for ease of handling of the infected Jurkat cell pellets while the CFP reporter permitted assessment of infection quality. Furthermore, knowledge of the NL4-3 clone sequence, gene expression and viral life cycle kinetics allowed for the maintenance of experimental reproducibility necessary for assurance the technology is effective. HIV-RNA protein interactions are temporal and transient with respect to the phases of HIV replication. Formaldehyde crosslinking covalently links direct RNA-protein and protein-protein interactions[27,28] thus preserving the *in vivo* interactions at the time of crosslinking while permitting the use of

harsh salt and detergent conditions to prevent artifactual interactions. We used formaldehyde treatment of the cells to stabilize the RNA-protein interactions at 48 hours post-infection, a time-point during which HIV genomic RNA is packaged into nascent virions. It may be noted that our strategy for isolation of HIV-1 splice variants by hybridization and identification of their protein interactomes is amenable to clinical or other studies investigating alternative HIV-1 strains or HIV replication time-points.

4.3.3 qPCR Results Demonstrate Efficient and Specific Isolation of HIV-1 Splice Variants. The yield and specificity of each HIV splice variant capture was measured using RT-qPCR. As was demonstrated above with regards to capture oligonucleotide design, the design and application of the qPCR assays must also consider sequence similarities among the different splice variants. The locations for qPCR amplification are shown in Figure 4.2A and specific sequences and locations are indicated in the SI. One qPCR assay (qPCR A) amplifies a region only found in the FL splice variant, another (qPCR B), a region found in both FL and PS variants, and a third (qPCR C), a region found in all three splice variant classes. To achieve relative quantification of each splice variant class in a given sample, the variability in reverse transcription and qPCR assay efficiencies was normalized using a full-length HIV RNA calibration curve (Details in SI).

The specificity of capture of each HIV splice variant was determined by analyzing an aliquot of purified RNA from each capture sample (FL capture, PS capture and CS capture) using RT-qPCR assays A, B, and C. Figure 4.2A shows which qPCR amplification regions are expected to be present as well as the attomoles of each region detected by the three RT-qPCR assays in each splice variant capture sample. Analysis of the RNA in the FL transcript

capture shows each region present at relatively the same quantity suggesting the primary HIV RNA variant captured is, indeed, the full-length transcript. Analysis of the PS and CS transcript capture samples shows the presence of at least 10-fold more of the regions amplified by qPCR assays B and C and qPCR assay C, respectively, than the regions amplified by the respective off-target qPCR assays. It should be noted that there is some variability in the quantification of the regions expected to be present in the FL and PS captures. For example, the PS capture samples show less amplification of region C than region B. This could be due to differences in secondary structure, and thus the reverse transcription efficiency, between the PS transcripts and the full-length transcripts used for the calibration curve. Similarly, these differences could be due to artifacts from protein crosslinking to the RNA. While this variability may affect the certainty of absolute quantitation of the splice variant present, the absence of the off-target splice variants is still evident. For example, if the full-length transcript were present in significant amounts in the PS capture, qPCR assay A would have significant amplification, which is not evident in the data presented here. The same logic holds for the CS capture data. Similarly, if there were large amounts of the PS or CS transcript variants present in the FL capture the quantification of qPCR assays B and C would be significantly higher than that of A. These results show that the primary HIV transcript in each of the splice variant capture samples is the intended target and is at least in 10-fold excess over the other splice variants.

We also sought to demonstrate the enrichment of the target HIV transcripts in the capture samples relative to off-target host cell RNA. To do this, a qPCR assay was used to quantify the highly expressed host transcript for GAPDH in the capture samples and the pre-capture lysate. The ratios of the HIV RNA target to the GAPDH transcript measured in the pre-

capture lysate and the capture samples indicate over a 200-fold enrichment of the respective target HIV RNA in the capture samples (Figures 4.2D-F).

The capture efficiency for each splice variant is defined here as the percent of the splice variant in the capture sample (capt) relative to the total amount of that transcript in the lysate prior to capture (pre-lys). Using the RT-qPCR assays described above the capture efficiencies for the FL, PS, and CS captures were 33, 76, and 2 percent, respectively (Figure S4.1). Additionally, the amount of each splice variant remaining in the lysate following each capture (post-lys) was measured and compared to the pre-lys measurements to determine the percentage lysate depletion of the transcripts. The FL, PS, and CS transcripts were 79, 87, and 78 percent depleted from the lysate following capture (Figure S4.1). Discussion and details of these measurements are in the SI.

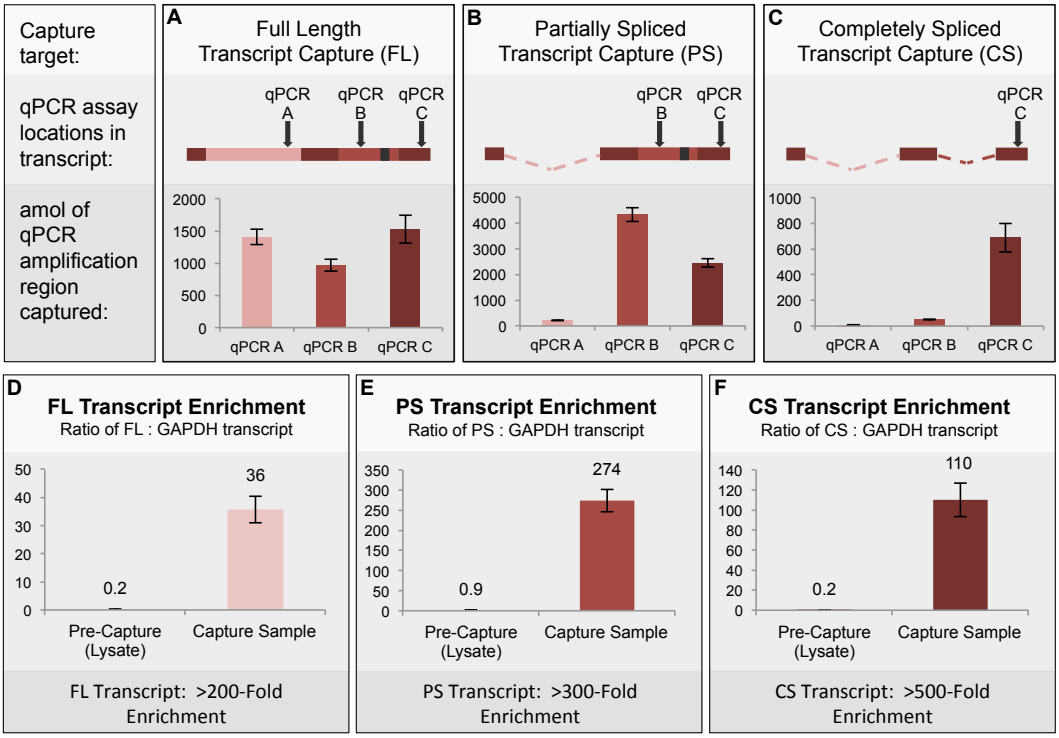


Figure 4.2: A-C: Specificity of capture of all three splice variant classes using qPCR assays. D-F: Enrichment of each splice variant in the capture samples. Fold enrichment is calculated by dividing the ratio of HIV:GAPDH in the capture sample by the same ratio in the lysate sample.

4.3.4 Spectral Analysis for Identification of Differential Splice Variant Interactomes

Following capture of the FL, PS and CS variants of HIV RNA, their associated proteins were purified and trypsin digested for mass spectrometric analysis. The resulting peptide spectra from the three biological replicates were searched and quantified using Maxquant search and label-free quantitation (LFQ) algorithms[29]. Statistical analysis was performed using the Perseus platform[30] to determine which proteins show significant differences in abundance among the three splice variant capture types. To do this, three pairwise comparisons were performed: FL vs PS, FL vs CS, and PS vs CS. Student's t-test p-values were calculated and a 5% permutation-based false discovery rate (FDR) was implemented for each pairwise comparison. In this way, two hundred twelve proteins were determined to have statistically significant differences in abundances in at least one pairwise comparison.

4.3.5 Hierarchical Clustering into Heatmap Shows Protein Enrichment Differences in lncRNA Captures.

A hierarchical clustering algorithm was employed to assist with interpretation of the protein interactomes of the three HIV-1 splice variant classes. A matrix of $\log_2$-transformed, mean-centered protein intensities for each splice variant capture biological replicate was compiled and includes all proteins enriched in at least one splice variant class relative to another splice variant class (212 proteins as stated above). By normalizing the data in this way we are able to remove experimentally introduced biases in protein quantitation such as varying peptide ionization efficiencies or protein solubilities that would hinder data interpretation. Next, the proteins were grouped, or clustered, based on the similarities between each protein's intensity profile across the three splice variants using a correlation-uncentered, centroid-linkage hierarchical clustering algorithm[31]. Finally, TreeView software[32] was used to create a heatmap to visualize these similarities and differences for all

212 proteins (Figure 4.3A). The protein intensities for three biological replicates for each of the three splice variants are represented in the nine columns of the heatmap. Each row contains the nine intensities for a single protein measured in each of the splice variant biological replicates. Protein intensity measurements with an intensity greater than the mean intensity for that protein across the nine columns (3 biological replicates of 3 different splice variant classes) are represented in yellow, with the brightness of the yellow reflecting the magnitude of the difference from the mean intensity. Analogously, proteins with an intensity measurement less than the mean intensity are represented in red. Black pixels represent protein intensities at or near the mean intensity for that protein. In Figure 4.3A it is clear that there are large clusters of proteins with higher abundance in the full length (FL) capture samples relative to the partially spliced (PS) and completely spliced (CS) capture samples (44 proteins), in the PS relative to the FL and CS capture samples (68 proteins), in the CS relative to the FL and PS clusters (46 proteins), and in both the FL and PS relative to the CS capture samples (29 proteins).

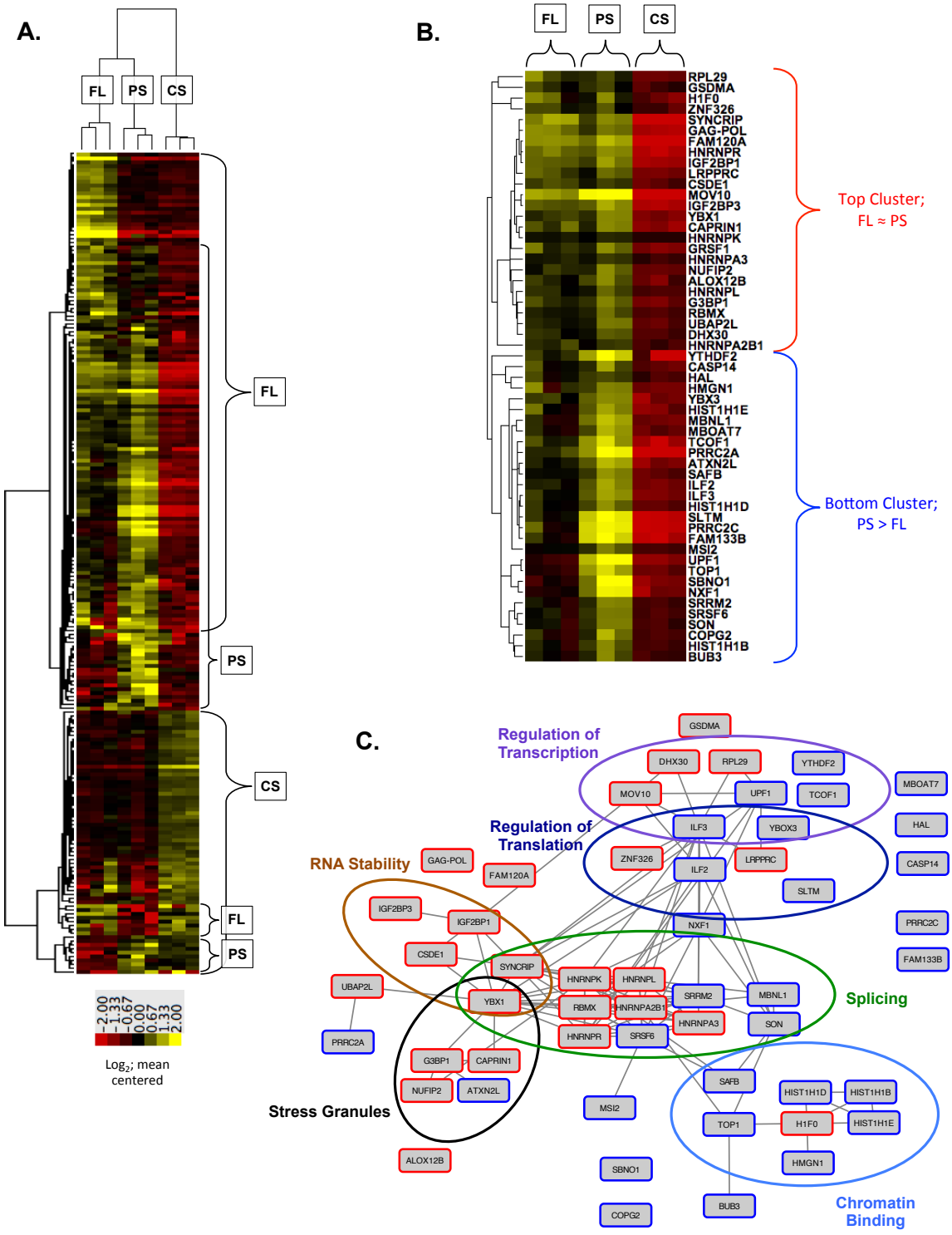


Figure 4.3: A. Heatmap displaying hierarchical clustering of HIV splice variant interactomes. B. Cluster of proteins enriched in full length and partially spliced variant interactomes. C. STRING analysis of cluster in B.

4.3.6 Analysis of Gene Ontology Term Enrichment. Gene ontology (GO) term enrichment analysis[33] was conducted to evaluate the significance of the proteins elevated in each splice variant's interactome. Table 4.1 shows the relevant GO terms that are over-represented in each splice variant interactome relative to what would be expected to occur randomly, including the p-value and the number of proteins annotated with that term. Satisfyingly, RNA-binding proteins are highly enriched in all three splice variant captures (p-values ranging from 3E-8 to 2E-35). Other notable enrichments include the enrichment of proteins involved in "translation", "translation initiation", and "cotranslational protein targeting to the membrane" in the CS capture (p-values 3E-12, 5E-10, and 1E-8 respectively) but enrichment of proteins involved in the "regulation of translation" in the FL and PS captures (p-values 8E-6 and 3E-6, respectively). Also interesting is the enrichment in the FL and PS captures of proteins that regulate RNA stability (p-values 5E-3 and 3E-2, respectively) and in FL that regulate RNA splicing (p-value 6E-3).

Category	Enriched GO Term	FL Capture		PS Capture		CS Capture	
		Enrichment p-value	# of prots	Enrichment p-value	# of prots	Enrichment p-value	# of prots
RNA-binding	RNA-binding	$2.06X10^{-35}$	60	$4.81X10^{-35}$	59	$3.10X10^{-8}$	25
	mRNA-binding	$7.82X10^{-9}$	12	$1.82X10^{-5}$	9	NA	2
	poly-A RNA binding	NA	56	NA	55	NA	24
RNA stability	regulation of mRNA stability	$4.82X10^{-2}$	7	$3.19X10^{-2}$	7	NA	2
	•CRD-mediated mRNA stabilization	$1.71X10^{-2}$	4	$1.42X10^{-2}$	4	NA	0
RNA splicing	RNA splicing	NA	10	$1.53X10^{-3}$	12	NA	4
	regulation of RNA splicing	$6.44X10^{-3}$	7	NA	6	NA	2
	•negative regulation of mRNA splicing	$3.45X10^{-2}$	3	NA	2	NA	0
Translation	translation	$2.82X10^{-3}$	7	NA	3	$2.91X10^{-12}$	9
	•translation initiation	NA	3	NA	1	$5.40X10^{-10}$	12
	•structural constituent of ribosome	$5.09X10^{-3}$	8	NA	1	$1.53X10^{-5}$	9
	regulation of translation	$7.99X10^{-6}$	12	$3.42X10^{-6}$	12	NA	5
	•negative regulation of translation	$9.94X10^{-6}$	9	$9.80X10^{-5}$	9	NA	2
DNA	histone H3-K27 trimethylation	NA	2	$2.45X10^{-2}$	3	NA	0
	DNA conformation change	NA	1	$1.77X10^{-2}$	2	NA	4
Protein localization	cotranslational protein targeting to membrane	NA	3	NA	1	$1.13X10^{-8}$	10
	protein localization to nucleus	NA	0	NA	1	$2.24X10^{-2}$	7
	•protein localization to nuclear envelope	NA	0	NA	0	$2.24X10^{-2}$	3
	nucleocytoplasmic transport	NA	4	NA	5	$3.59X10^{-2}$	8
Metabolic processes	acyl-CoA metabolic process	$4.08X10^{-2}$	7	NA	1	NA	1
	carboxylic acid metabolic process	$9.04X10^{-11}$	8	NA	1	NA	1
	dicarboxylic acid metabolic process	$9.46X10^{-6}$	11	NA	1	NA	2
	tricarboxylic acid metabolc process	$9.62X10^{-6}$	8	NA	0	NA	1
Miscellaneous	cytoplasmic stress granule	$4.77X10^{-8}$	7	$1.33X10^{-6}$	7	NA	1
	cell adhesion molecule binding	NA	5	NA	2	$4.53X10^{-3}$	10

Table 4.1: GO Term Enrichment Analysis for Splice Variant Interactomes.

4.3.7 Discussion of Splice Variant Interactomes. The GO term enrichment analysis, hierarchical clustering, and analysis of known protein-protein interactions using STRING[34] software together assist in evaluation of the splice variant interactomes. For example, a large section of the heatmap, shown in Figure 4.3B, contains proteins with abundances elevated in the full length (FL) and partially spliced (PS) transcripts relative to the completely spliced (CS) transcripts. This section is particularly interesting since the FL and PS transcripts, unlike the CS transcript, retain at least one intron and therefore require regulation of splicing and RNA stability for their nuclear export into the cytoplasm that the CS transcript does not require. Proteins with these functions are enriched in the FL and/or PS interactomes (Table 4.1) and are also represented in the cluster shown in Figure 4.3B. This cluster can also be subdivided into two smaller clusters, the top cluster (red) containing proteins with the

intensities approximately the same in the FL and PS captures and the bottom cluster (blue) containing proteins with intensities higher in the PS capture than the FL capture. Figure 4.3C shows the known protein-protein interactions, obtained from STRING, for the proteins contained within these two subdivisions of the cluster in Figure 4.3B. Proteins in the top cluster have a red border and proteins in the bottom cluster have a blue border. Within the top cluster are many protein involved in RNA stability, including CRD-mediated mRNA stabilization, as well as proteins involved with splicing, particularly the regulation of splicing. This is expected since the FL and PS transcripts must evade degradation and complete splicing prior to exiting the nucleus. Also present in the top cluster are proteins that are part of the cytoplasmic granule. Previous evidence shows that the FL transcript is often incorporated into stress granules or another cytoplasmic ribonucleoprotein complex (RNP) called the Staufen RNP [20,21,35]. Proteins with other known functions are also within this cluster though their role in HIV replication is unknown. For example, FAM120A, a known RNA binding protein implicated in the stress response[36], is clustered closely with proteins involved in RNA stability and splicing. Also, ALOX12B, a protein known to function in the regulation of gene expression[37], clusters with components of the stress granule. Though these two proteins are not known to function in this way with respect to HIV replication, they make prime candidates for further functional analysis during HIV replication. Within the bottom cluster are many proteins with known functions in chromatin binding, regulation of transcription and regulation of translation. Many have known roles in HIV replication, such as MOV10 and UPF1[38], while others will require further analysis to investigate their role in HIV replication.

4.4 FUTURE DIRECTIONS:

The HyPR-MS strategy is useful for discovery of proteins associated with the HIV RNAs and can differentiate proteins associated with the different splice variants. However, to understand the roles of these proteins in HIV replication requires further analysis. The clustering and heatmap analysis strategy provides a mechanism to interpret the data and develop informed hypotheses of the roles of such proteins. We are in the process of experimentally testing some of these hypotheses using siRNA knockdown techniques and a specially designed HIV viral strain to track the translational activity of the full length and completely spliced transcripts. As was done in Chapter 2, we will infect 293T cells with an HIV virus that incorporates CFP fluorescent protein with the Gag protein (translated from the full length transcript) and produces mCherry as a translational product from the completely spliced transcript. Each protein identified as a protein interactor with the HIV splice variant will be individually knocked-down using siRNAs and the impact on HIV replication will be evaluated using the signals from CFP and mCherry by fluorescence microscopy. We hypothesize that proteins that cluster together in the heatmap produced by HyPR-MS may work together to perform the same functions in HIV replication. Therefore, we predict that when those proteins are knocked-down, they will show the same fluorescent phenotype. Much information can be gleaned from these experiments regardless of confirmation of our hypothesis. This work will be conducted and the data combined with this chapter for publication.

4.5 SUPPLEMENTARY INFORMATION:

4.5.1 qPCR Assays Design and Strategy. To achieve relative quantitation of each splice variant class in a given sample, the variability in reverse transcription and qPCR assay efficiencies were normalized using a full length HIV RNA calibration curve. Briefly, full length HIV RNA was isolated from HIV virions using Trizol reagent and quantified using Ribogreen reagent and fluorescence detection. Since HIV virions primarily contain full length HIV RNA[1], the total RNA mass can be converted to moles of HIV RNA and each amplification region for the three qPCR assays is represented in an approximate 1:1:1 ratio. The resulting sample of HIV RNA was then diluted to make several calibration standards of known concentration, each of which was reverse transcribed and analyzed using all three qPCR assays. The three calibration curve equations from this analysis were then used for determining the concentration of their respective amplification regions in each splice variant capture sample.

4.5.2 Capture Efficiency and Lysate Depletion. Capture efficiency is defined here as the percent of the splice variant in the capture sample (capt) relative to the total amount of that transcript in the lysate prior to capture (pre-lys). Lysate depletion is defined here as the percent of the splice variant remaining in the lysate after capture (post-lys) relative to the amount prior to capture (pre-lys). These calculations are represented in the equations below and the results for the three splice variant classes using RT-qPCR assays A, B, and C are shown in Figure S4.1.

1) % Capture Efficiency: (capt)/(pre-lys) x 100
2) % Lysate Depletion: [(pre-lys) – (post-lys) / (pre-lys)] x 100

The accuracy of the capture efficiency measurement is limited by variability in reverse transcription efficiency caused by differences in sample characteristics. Specifically, the efficiency of the reverse transcription reaction is dependent on the complexity of the sample as well as the concentration of the transcript in the sample[(2)]. The capture samples have a high concentration of the target transcript but the complexity of the sample is relatively low. Conversely, the concentration of the target in the pre-lys sample is relatively low and the sample is much more complex. These differences likely cause different reverse transcription efficiencies for the target transcript in the capture and pre-lysate samples suggesting that the % capture efficiencies should be used only as guidance for assay functionality and not for absolute quantitation. The percent lysate depletion calculation, however, measures the target transcript in two samples with relatively similar complexities (pre-lys and post-lys) and thus likely is an accurate measurement. This measurement shows that the capture process removes approximately 80% of each of the splice variants from the lysate (Figure S1). This value shows that the vast majority of the FL and PS splice variants are removed from the lysate prior to the subsequent capture of the PS and CS variants, respectively, allowing for the specificity of capture among the splice variants as demonstrated in Figure 4.2. The lysate depletion value also represents a high estimation of % capture efficiency but does not account for potential losses due to non-specific binding to tubes and/or beads and losses due to washing of the beads prior to release into solution.

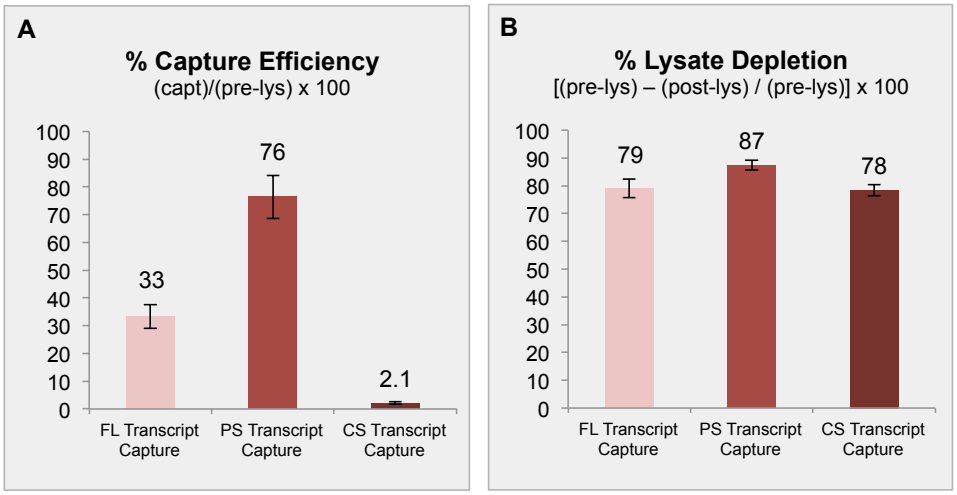


Figure S4.1: A. Capture efficiency calculations for each splice variant. B. Percent lysate depletion of each splice variant.

Table S1: Capture Oligonucleotide Sequences

HIV Splice Variant Target	Capture Oligonucleotide Sequence*	Release Oligonucleotide Sequence*	HIV NL4-3 Genome Complementarity; Nucleotide Positions
Full Length (FL)	5'-CAATTGACTTAC TATGCTTTCTGTGGC TATTTTTTGTA-3'	5'-TACAAAAAATAG CCACAGAAAGCATA GTAAGTCAATTG-3'	3210-3239
Partially Spliced (PS)	5'-CAATTGACTGTA CAATTAATTTCTACAG ATGTGTTCAG-3'	5'-CTGAACACAT CTGTAGAAATTAATTG TACAGTCAATTG-3'	6622-6651
Completely Spliced (CS)	5'-TCGTATCTCTTT TCTTTTAAAAAGTGG CTAAGATCTAC-3'	5'-GTAGATCTTAG CCACTTTTTAAAAGA AAAGAGATACGA-3'	8585-8614

*The nucleotides indicated in red letters are not complementary to the target transcript but are used as a "toe-hold" for the release mechanism as described in the text.

Table S2: qPCR Assays

Assay	Detected Splice Variant(s)	HIV NL4-3 Genome Amplification Region; Nucleotide Positions	Primer 1 (FWD) Sequence	Primer 2 (REV) Sequence	Probe Sequence
qPCR A	FL	3338-3480	GAGTTTGTCAA TACCCCTCCC	CCTCTGTCAGTT ACATATCCT GC	TTTCCCTATTGG CTGCCCCATCT
qPCR B	FL, PS	7069-7166	CCTCCCATC AGTGGACAAAT TA	CTGAAGATCTCG GACCCATTG	ACCACCATCTC TTGTTAATAGCA GCCC
qPCR C	FL, PS, CS	8794-8893	AGAGGCCAA TAAAGGAGAGA AC	GCTGTCAAACCT CCACTCTAA	TGTGAGCCTGC ATGGAATGGAT GA

qPCR primer-probes purchased from IDT; GAPDH qPCR assay is Hs.PT.39a.22214836 from IDT.

4.6 AUTHOR CONTRIBUTIONS:

Cell culture, virus production, and infection of cells for HyPR-MS was conducted by J.T.B and supported by NIH grant RO1AI110221 (PI: N.M.S). J.T.B received support from a National Science Foundation Graduate Research Fellow program (grant DGE-1256259), Research Competition Award from the UW—Madison Office of the Vice Chancellor of Research and Graduate Education, and a UW-Madison Graduate School Dissertation Completion Fellowship. HyPR-MS design and implementation, RT-qPCR, protein sample preparation, mass spectrometry data collection, siRNA knockdown with viral infection and fluorescence detection, data analysis and writing of manuscript were conducted by R.A.K. Western blot analysis was conducted by J.T.B and R.A.K. HPLC and mass spectrometer method development and maintenance were conducted by M.S. Work by R.A.K. and M.S. was funded by NIH grants 1P50HG004952 and R01CA193481 (PI: L.M.S). N.M.S. and L.M.S. provided advice and expertise throughout the project.

4.7 REFERENCES:

1 Ocwieja, K. E. *et al.* Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic acids research* **40**, 10345-10355, doi:10.1093/nar/gks753 (2012).

2 Purcell, D. F. & Martin, M. A. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *Journal of virology* **67**, 6365-6378 (1993).

3 Carrera, C., Pinilla, M., Perez-Alvarez, L. & Thomson, M. M. Identification of unusual and novel HIV type 1 spliced transcripts generated in vivo. *AIDS research and human retroviruses* **26**, 815-820, doi:10.1089/aid.2010.0011 (2010).

4 Stoltzfus, C. M. Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication. *Advances in virus research* **74**, 1-40, doi:10.1016/S0065-3527(09)74001-1 (2009).

5 Erkelenz, S. *et al.* Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *Rna* **19**, 96-102, doi:10.1261/rna.037044.112 (2013).

6 Marchand, V. *et al.* A Janus splicing regulatory element modulates HIV-1 tat and rev mRNA production by coordination of hnRNP A1 cooperative binding. *Journal of molecular biology* **323**, 629-652 (2002).

7 Damgaard, C. K., Tange, T. O. & Kjems, J. hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. *Rna* **8**, 1401-1415 (2002).

8 Erkelenz, S. *et al.* Balanced splicing at the Tat-specific HIV-1 3'ss A3 is critical for HIV-1 replication. *Retrovirology* **12**, 29, doi:10.1186/s12977-015-0154-8 (2015).

9 Kaminchik, J. *et al.* Genetic characterization of human immunodeficiency virus type 1 nef gene products translated in vitro and expressed in mammalian cells. *Journal of virology* **65**, 583-588 (1991).

10 Kaminchik, J. *et al.* Cellular distribution of HIV type 1 Nef protein: identification of domains in Nef required for association with membrane and detergent-insoluble cellular matrix. *AIDS research and human retroviruses* **10**, 1003-1010, doi:10.1089/aid.1994.10.1003 (1994).

11 Felber, B. K., Hadzopoulou-Cladaras, M., Cladaras, C., Copeland, T. & Pavlakis, G. N. rev protein of human immunodeficiency virus type 1 affects the stability and transport of the viral mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 1495-1499 (1989).

12 Zapp, M. L. & Green, M. R. Sequence-specific RNA binding by the HIV-1 Rev protein. *Nature* **342**, 714-716, doi:10.1038/342714a0 (1989).

13 Huang, F. *et al.* RNA helicase MOV10 functions as a co-factor of HIV-1 Rev to facilitate Rev/RRE-dependent nuclear export of viral mRNAs. *Virology* **486**, 15-26, doi:10.1016/j.virol.2015.08.026 (2015).

14 Banerjee, A., Benjamin, R., Balakrishnan, K., Ghosh, P. & Banerjee, S. Human protein Staufen-2 promotes HIV-1 proliferation by positively regulating RNA export activity of viral protein Rev. *Retrovirology* **11**, 18, doi:10.1186/1742-4690-11-18 (2014).

15 Butsch, M. & Boris-Lawrie, K. Destiny of unspliced retroviral RNA: ribosome and/or virion? *Journal of virology* **76**, 3089-3094 (2002).

16 Bolinger, C., Sharma, A., Singh, D., Yu, L. & Boris-Lawrie, K. RNA helicase A modulates translation of HIV-1 and infectivity of progeny virions. *Nucleic acids research* **38**, 1686-1696, doi:10.1093/nar/gkp1075 (2010).

17 Clever, J., Sassetti, C. & Parslow, T. G. RNA secondary structure and binding sites for gag gene products in the 5' packaging signal of human immunodeficiency virus type 1. *Journal of virology* **69**, 2101-2109 (1995).

18 Berkowitz, R. D., Luban, J. & Goff, S. P. Specific binding of human immunodeficiency virus type 1 gag polyprotein and nucleocapsid protein to viral RNAs detected by RNA mobility shift assays. *Journal of virology* **67**, 7190-7200 (1993).

19 Lever, A., Gottlinger, H., Haseltine, W. & Sodroski, J. Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions. *Journal of virology* **63**, 4085-4087 (1989).

20 Abrahamyan, L. G. *et al.* Novel Staufen1 ribonucleoproteins prevent formation of stress granules but favour encapsidation of HIV-1 genomic RNA. *Journal of cell science* **123**, 369-383, doi:10.1242/jcs.055897 (2010).

21 Milev, M. P., Ravichandran, M., Khan, M. F., Schriemer, D. C. & Mouland, A. J. Characterization of staufen1 ribonucleoproteins by mass spectrometry and biochemical analyses reveal the presence of diverse host proteins associated with human immunodeficiency virus type 1. *Frontiers in microbiology* **3**, 367, doi:10.3389/fmicb.2012.00367 (2012).

22 Timilsina, U. & Gaur, R. Modulation of apoptosis and viral latency - an axis to be well understood for successful cure of human immunodeficiency virus. *The Journal of general virology* **97**, 813-824, doi:10.1099/jgv.0.000402 (2016).

23 Kula, A. *et al.* Characterization of the HIV-1 RNA associated proteome identifies Matrin 3 as a nuclear cofactor of Rev function. *Retrovirology* **8**, 60, doi:10.1186/1742-4690-8-60 (2011).

24 Marchand, V. *et al.* Identification of protein partners of the human immunodeficiency virus 1 tat/rev exon 3 leads to the discovery of a new HIV-1 splicing regulator, protein hnRNP K. *RNA biology* **8**, 325-342 (2011).

25 Adachi, A. *et al.* Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *Journal of virology* **59**, 284-291 (1986).

26 Connor, R. I., Chen, B. K., Choe, S. & Landau, N. R. Vpr Is Required for Efficient Replication of Human-Immunodeficiency-Virus Type-1 in Mononuclear Phagocytes. *Virology* **206**, 935-944, doi:Doi 10.1006/Viro.1995.1016 (1995).

27 Moller, K., Rinke, J., Ross, A., Buddle, G. & Brimacombe, R. Use of Formaldehyde in Rna-Protein Cross-Linking Studies with Ribosomal-Subunits from Escherichia-Coli. *Eur J Biochem* **76**, 175-187, doi:Doi 10.1111/J.1432-1033.1977.Tb11583.X (1977).

28 Solomon, M. J. & Varshavsky, A. Formaldehyde-Mediated DNA Protein Crosslinking - a Probe for Invivo Chromatin Structures. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 6470-6474, doi:Doi 10.1073/Pnas.82.19.6470 (1985).

29 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

30 Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature methods* **13**, 731-740, doi:10.1038/nmeth.3901 (2016).

31 de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453-1454, doi:10.1093/bioinformatics/bth078 (2004).

32 Saldanha, A. J. Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**, 3246-3248, doi:10.1093/bioinformatics/bth349 (2004).

33 Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic acids research* **45**, D183-D189, doi:10.1093/nar/gkw1138 (2017).

34 Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research* **45**, D362-D368, doi:10.1093/nar/gkw937 (2017).

35 Chatel-Chaix, L., Boulay, K., Mouland, A. J. & Desgroseillers, L. The host protein Staufen1 interacts with the Pr55Gag zinc fingers and regulates HIV-1 assembly via its N-terminus. *Retrovirology* **5**, 41, doi:10.1186/1742-4690-5-41 (2008).

36 Tanaka, M. *et al.* A Novel RNA-Binding Protein, Ossa/C9orf10, Regulates Activity of Src Kinases To Protect Cells from Oxidative Stress-Induced Apoptosis. *Mol Cell Biol* **29**, 402-413, doi:10.1128/MCB.01035-08 (2009).

37 Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic acids research* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

38 Toro-Ascuy, D., Rojas-Araya, B., Valiente-Echeverria, F. & Soto-Rifo, R. Interactions between the HIV-1 Unspliced mRNA and Host mRNA Decay Machineries. *Viruses* **8**, doi:10.3390/v8110320 (2016).