UNIVERSITY OF CALIFORNIA SAN DIEGO

Revealing and understanding human protein interactome through high-throughput sequencing and meta-analysis

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioengineering

by

Zhijie Qi

Committee in charge:

Professor Sheng Zhong, Chair
Professor Jin Zhang, Co-Chair
Professor Ferhat Ay
Professor Bing Ren
Professor Kun Zhang

2022

The Dissertation of Zhijie Qi is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my grandmother, Qiaotang Liao, and my parents Qin Su and Zhangmin Qi, for their caring and support. To my cat, Loli, for her lovely paws and sweet purring. To all the metal bands and Japanese animations I listen and watch for their strong and encouraging songs and plot.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Sheng Zhong for his support as the chair of my committee. Through all these years, his guidance has proved to be invaluable along my journey toward the degree. I would also like to acknowledge Dr. Zhangming Yan and Dr. Weizhong Chen for their helpful discussion and suggestions. It is their support that helped me in an immeasurable way.

Chapter 1, in full, is an adapted reprint of materials that appears in Johnson, Kara L.; Qi, Zhijie; Yan, Zhangmin; Wen, Xingzhao; Nguyen, Tri C.; Zaleta-Rivera, Kathia; Chen, Chien-Ju; Fan, Xiaochen; Sriram, Kiran; Wan, Xueyi; Chen, Zhen Bouman; Zhong, Sheng. “Revealing protein-protein interactions at the transcriptome scale by sequencing”, Molecular Cell, Volume 81, Issue 19, 2021. The dissertation author was one of the primary investigators and the co-first author of this paper.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Qi. Zhijie; Zhong, Sheng. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Qi. Zhijie; Johnson, Kara L.; Chen, Junchen; Zhao, Wenxin; Wen, Xingzhao; Xue, Shuanghong; Zhong, Sheng. The dissertation author was the primary investigator and author of this paper.

VITA

2016 Bachelor of Science, Washington University in St. Louis

2022 Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Revealing and understanding human protein interactome
through high-throughput sequencing and meta-analysis

by

Zhijie Qi

Doctor of Philosophy in Bioengineering

University of California San Diego, 2022

Professor Sheng Zhong, Chair
Professor Jin Zhang, Co-Chair

Proteins play a central role in human cell activities through their interactions with proteins themselves and RNAs. Yet we lack technologies that can profile protein-protein interactions (PPIs) and RNA-protein interactions (RPIs) both effectively and efficiently. For the existing technologies, the search space of proteins is limited, and the up-scaling of the products is resource-intensive. Meanwhile, the variability of PPIs detected from different technologies also leads to the lack of consensus on the architecture of the profiled PPI networks. This dissertation work presents PROPER-seq (protein-protein interaction sequencing) and PRIM-seq (protein-RNA interaction mapping by sequencing), two time-effective technologies to map cell-wide protein-protein interactions (PPIs) and RNA-protein interactions (RPIs) respectively *in vitro*. This dissertation work also utilizes PPIs derived from PROPER-seq to induct human PPI network features. The technologies and analysis together provide rich resources to the community for studying protein-related interactions and understanding human proteome.

In Chapter 1, I describe PROPER-seq to map PPIs. I showed the PROPER-seq identified PPIs are of robust reproducibility, precision, and recall performance. I present PROPER v1.0, a human PPI network that consists of 8,635 proteins and 210,518 interactions. I delivered PROPERseqTools and PROPER v1.0 database, an open-source software, and an online database for people to process PROPER-seq libraries and to better access PROPER v1.0.

In Chapter 2, I describe my analysis of utilizing multiple human PPI datasets to systematically examine the architectural characteristics of human PPI networks. I found consistent evidence to support that a comprehensive human PPI network should be a scale-free network filled with many completed or close-to-completed cliques. The hub proteins with similar molecular functions are often highly inter-connected in cliques and serve as building blocks in small network motifs.

In Chapter 3, I describe PRIM-seq to systematically map RNA-protein interactions (RPIs) *in vitro*. I present PRIM v.1.0, a human RPI network that consists of 117,516 RPIs from 8,440 RNAs and 7,691 proteins. I showed the enrichment of previously characterized RNA-binding proteins (RBPs) in PRIM v.1.0 and found evidence to support PHGDH as an RBP.

CHAPTER 1 High throughput mapping of protein-protein interactions

1.1 Abstract

Protein-protein interactions (PPIs) are one of the most fundamental units involved in numerous cellular processes. However, a technique for efficiently mapping large-scale protein-protein interactions is still missing. Here, we describe PROPER-seq (protein-protein interaction sequencing) to map protein-protein interactions (PPIs) at the transcriptome scale *in vitro*. PROPER-seq first converts transcriptomes of input cells into RNA-barcoded protein libraries, in which all interacting protein pairs are captured through nucleotide barcode ligation, recorded as chimeric DNA sequences, and decoded at once by sequencing and mapping. We applied PROPER-seq to human embryonic kidney cells, T lymphocytes, and endothelial cells. We showed the PROPER-seq identified PPIs are of robust reproducibility, precision, and recall performance. We present PROPER v1.0, a human PPI network that consists of 8,635 proteins and 210,518 interactions. Within PROPER v1.0, 8,100 PPIs are supported by previous data, 17,638 PPIs are predicted by the prePPI algorithm without previous experimental validation, and 100 PPIs overlap with human synthetic lethal gene pairs. In addition, four previously uncharacterized interaction partners with poly(ADP-ribose) polymerase 1 (PARP1) (a critical protein in DNA repair), including XPO1, MATR3, IPO5, and LEO1 are validated in vivo. We delivered PROPERseqTools and PROPER v1.0 database, an open-source software, and an online database for people to process PROPER-seq libraries and to better access PROPER v1.0. PROPER-seq presents a time-effective technology to map PPIs at the transcriptome scale, and PROPER v.1.0 provides a rich resource for studying PPIs.

1.2 Introduction

Our ability to interpret the human genome function is greatly improved by our understanding of the interaction networks formed by the genome products. Recent technological breakthroughs enabled genome-wide mapping of DNA-DNA (Dekker et al. 2017), protein-DNA (Consortium 2004, 2012), RNA-DNA (Sridhar et al. 2017; Yan et al. 2019; Li et al. 2017), and RNA-RNA (Lu, Gong, and Zhang 2018; Sharma et al. 2016; Aw et al. 2017; Nguyen et al. 2016) interactions. However, genome-wide mapping of human protein-protein interactions (PPI) remains a resource-intensive task.

Large-scale PPI mapping methods can be grouped into 3 classes, that are “parallelized one-to-one”, “one-to-many”, and “many-to-many” approaches. The “parallelized one-to-one” methods leverage automation and parallelization to enhance the throughput of yeast two-hybrid (Y2H) assays(Rual et al. 2005; Luck et al. 2020; Rolland et al. 2014). These include High-Throughput Y2H (Walhout and Vidal 2001), MAPPIT (Lievens et al. 2009), and QIS-seq (Lewis et al. 2012), which massively parallelized the binary interactions, and RLL-Y2H (Yang et al. 2018), Stitch-Seq (Kawalia et al. 2015), CrY2H-seq (Trigg et al. 2017), BFG-Y2H (Yachie et al. 2016), in which gene sequences of interacting PPI pairs were fused and sequenced. The “one-to-many” methods start with purifying or tagging a target (or “bait”) protein to identify the co-purified proteins in spatial proximity using affinity purification(Vermeulen, Hubner, and Mann 2008), proximity biotinylation (BioID) (Touchette et al. 2017), GFP fusion (Zhang et al. 2017), or protein microarray (Kukar et al. 2002). The “many-to-many” approach, aiming to read out all the pairwise PPIs from a single experiment, has been applied to resolve ligand-target pairs (McGregor, Jain, and Liu 2014) and antibody-antigen pairs (Gu et al. 2014).

The aforementioned methods can also be grouped into protein interaction assays and spatial proximity assays, depending upon the property of the protein pair unraveled. The protein interaction assays can be further divided into binary and non-binary assays (Yu et al. 2008). Whereas binary assays such as Y2H yield direct pairwise protein interactions, non-binary assays such as AP-MS and co-IP yield physical associations, where each protein identified in a pair may not directly interact with each other, as in a multi-protein complex. Finally, spatial proximity assays including BioID (Touchette et al. 2017) reveal proteins that may not form physical interactions or associations, other than being spatially proximal.

In this work, we introduce protein-protein interaction sequencing (PROPER-seq), a resource-efficient many-to-many non-binary assay for PPI mapping. The central idea of PROPER-seq is to convert each PPI into a unique sequence of DNA and then to leverage extremely high throughput DNA sequencing to decode these PPIs. To implement this idea, we developed a technique called SMART-display to attach a unique RNA barcode to every protein (Figure 1.1A) and a method called incubation, ligation, and sequencing (INLISE) to sequence the pair of DNA barcodes that are attached to two interacting proteins (Figure 1.1B). We named the overall technology combining SMART-display and INLISE as PROPER-seq (Figure 1.1A). The input to PROPER-seq is a group of cells, and PROPER-seq's output is a list of identified PPIs and their associated read counts and test statistics. We demonstrated that PROPER-seq is capable of scanning on the order of 10,000 × 10,000 protein pairs in one experiment and of identifying both binary and multiway protein interactions. Applying PROPER-seq on human embryonic kidney cells, T lymphocytes, and endothelial cells, we constructed a map of human PPIs (PROPER v.1.0) that includes 210,518 PPIs involving 8,635 proteins.

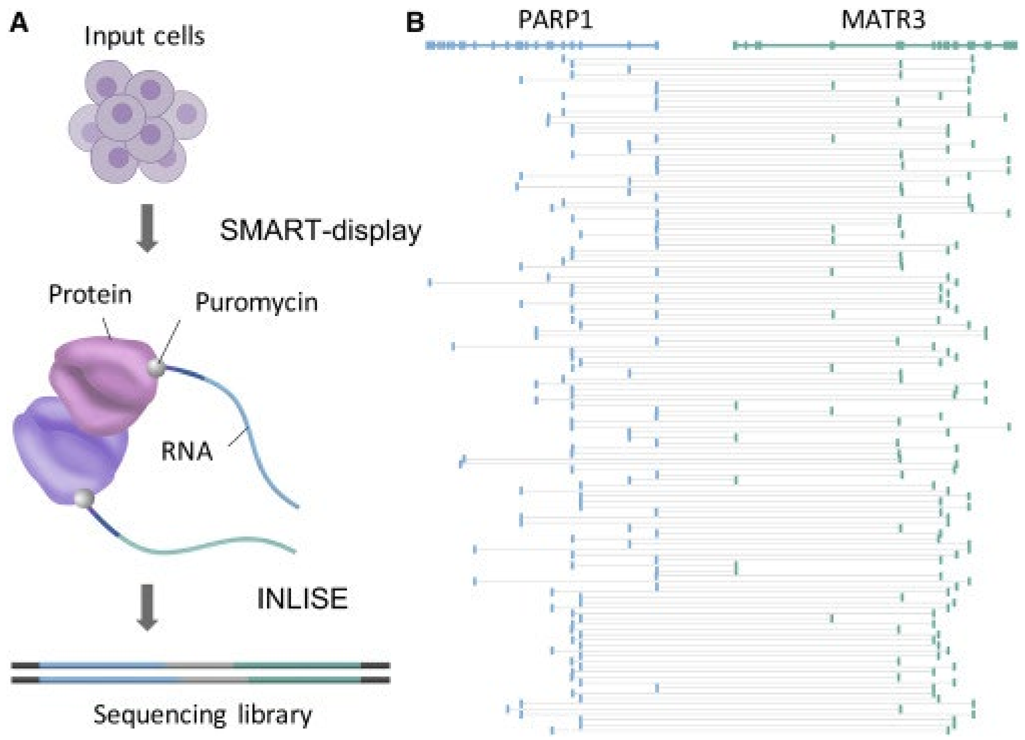


Figure 1.1: Overview of PROPER-seq experimental pipeline
(A) PROPER-seq starts with SMART-display that transforms the input cells into a library of RNA-barcoded proteins (the first arrow), followed by INLISE that transforms the barcoded proteins into a sequencing library, such that the barcodes of interacting protein pairs form a chimeric sequence (the second arrow).
(B) Alignment of the barcodes to reveal the identities of the two genes (top track) between which the chimeric sequences (rows) were formed.

1.3 Design

1.3.1 SMART-display

We developed a modified mRNA-display method, called SMART-display, to efficiently generate a protein library in which the proteins are conjugated with their mRNA (Figures 1.1A and 1.2). Thus, the mRNA serves as the unique nucleic acid barcode for each protein. Similar to mRNA display (Roberts and Szostak 1997; Barendt et al. 2013), SMART-display is designed to create mRNA-protein fusions, specifically by adding an amino acid analog puromycin (“P” in the purple circle, Figure S1.1A) near the 3′ end of the mRNA. The translated protein from this mRNA is then covalently linked with its mRNA when puromycin enters the A site of the ribosome and is joined to the amino acid chain. This generates an mRNA-protein fusion, which is then released from the ribosome (Figure S1.1).

In SMART-display, we replaced the gene-by-gene cloning (or gene-by-gene PCR) step in mRNA display with reactions that can be carried out with a mixture of genes (or mRNAs) without the need for independent purification of each gene. This was achieved by replacing the gene-specific primers in mRNA display with template switching oligos (TSOs) (Petalidis et al. 2003; Zhu et al. 2001) that are universal for all genes. The input to SMART-display is a user-selected cell population. An important intermediate product of SMART-display is a gene library suitable for mRNA display, in which the sequences for transcription initiation, translation initiation, and puromycin attachment have been incorporated in the appropriate places for every gene (Figure 1.2A). The output of SMART-display is a library of display complexes in the form of mRNA-linker-protein (Figure 1.2H; Figure S1.1D).

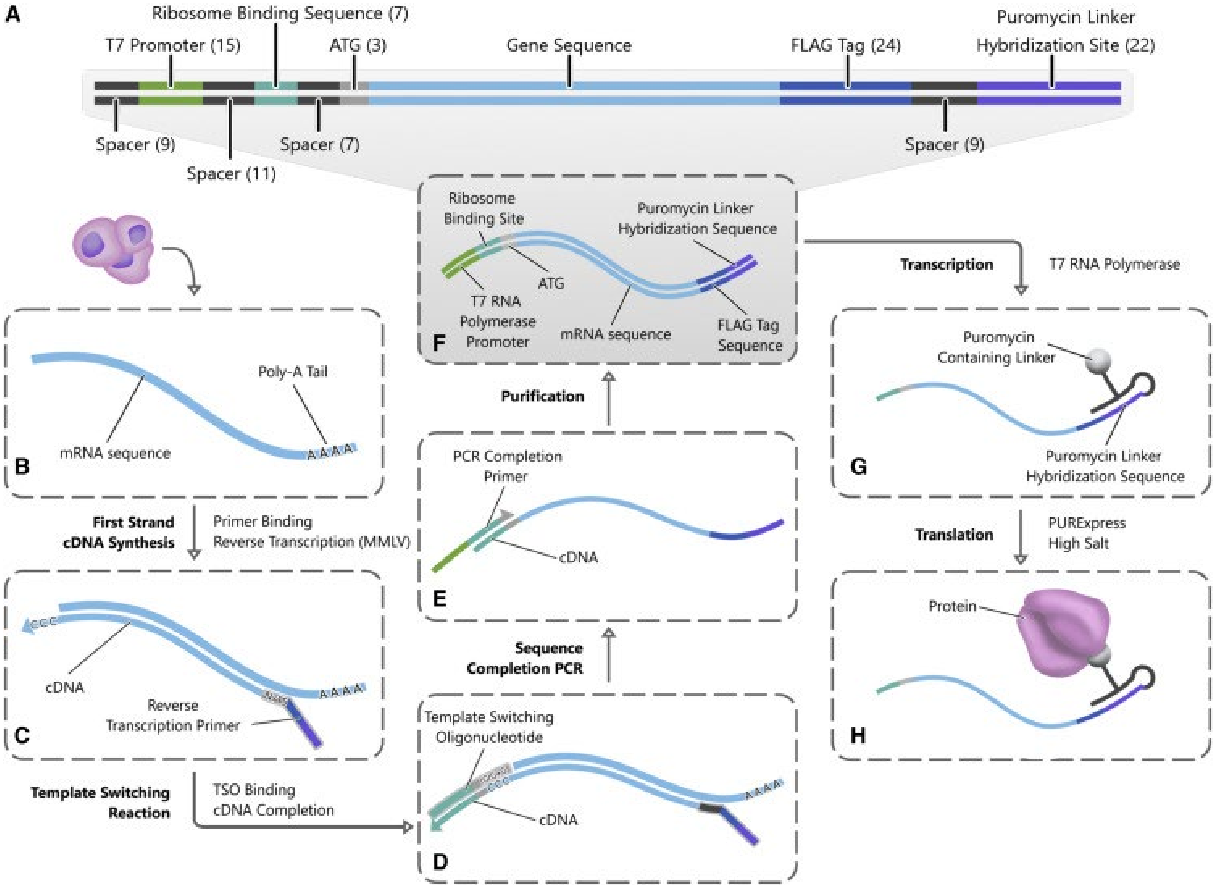

A
Ribosome Binding Sequence (7)
T7 Promoter (15)
ATG (3)
Gene Sequence
FLAG Tag (24)
Puromycin Linker Hybridization Site (22)
Spacer (9)
Spacer (11)
Spacer (7)
Spacer (9)
Ribosome Binding Site
Puromycin Linker Hybridization Sequence
ATG
T7 RNA Polymerase Promoter
mRNA sequence
FLAG Tag Sequence
F
Transcription
T7 RNA Polymerase
Poly-A Tail
mRNA sequence
B
Purification
PCR Completion Primer
cDNA
E
Puromycin Containing Linker
Puromycin Linker Hybridization Sequence
G
First Strand cDNA Synthesis
Primer Binding Reverse Transcription (MMLV)
Translation
PURExpress High Salt
Protein
cDNA
Reverse Transcription Primer
C
Sequence Completion PCR
H
Template Switching Oligonucleotide
cDNA
D
Template Switching Reaction
TSO Binding cDNA Completion

Figure 1.2: SMART-display
(A) Structure of gene templates produced by SMART-display (the product of F).
(B) Poly(A)-selected and rRNA-depleted mRNA is collected from the input cells.
I Reverse transcription primer containing a random sixteen base pair region followed by the sequences for a FLAG tag and a GC-rich puromycin linker hybridization site is annealed to the mRNA.
(D) Reverse transcription and incorporation of the template switching oligo (TSO).
(E) PCR is performed with a primer that partially overlaps the TSO sequences to introduce the T7 promoter and complete the ribosome binding site.
(F) Double-stranded DNA is purified.
(G and H) Transcribed RNA is ligated to a puromycin-containing linker sequence (G) and subsequently translated to form mRNA-protein fusion products (H).

1.3.2 INLISE

As the second key step of PROPER-seq, INLISE converts PPIs into chimeric sequences with the structure cDNA1-linker-cDNA2 (Figure 1.3). The inputs of INLISE are two display libraries generated by SMART-display. Each display library contains approximately 15,000 mRNA-protein fusions. One library, called the bait library, is immobilized on streptavidin beads through the biotin on the puromycin linker sequence ("B" in the blue circle, Figure S1.1A). The other library, called the prey library, is not immobilized, as the biotin is cleaved from the puromycin linker and is mixed with the bait library to allow interactions. After the removal of spurious interactions, the mRNA barcodes of interacting proteins are ligated to create a chimeric sequence in the form of cDNA1-linker-cDNA2, where cDNA1 and cDNA2 represent the two interacting proteins. These chimeric sequences are subsequently selected for and subjected to paired-end sequencing.

1.3.3 Identification of PPIs by statistical tests

Our overarching goal is to examine as many protein pairs as possible and assign a binary indicator (interacting or not) to every protein pair. Toward this goal, we subjected the mapped read pairs on each gene pair to an association test. The null hypothesis is that the mapping of a read pair to one gene is independent of the mapping of this read pair to the other gene (Figure 1.4A). We used Bonferroni-Hochberg (BH) correction to account for multiple hypothesis tests (Benjamini and Hochberg 1995). A pair of proteins was identified as interacting (i.e., a PPI) by two criteria. First, the BH-corrected p-value derived from the association (Chi-square) test is smaller than 0.05 (Figure 1.4A). Second, the number of the chimeric read pairs mapped to this gene pair is no less than 4 times the average number of chimeric reads mapped to any gene pair (4 × the number of all mapped chimeric read pairs / the number of all mapped gene pairs). Hereafter, we call these the default threshold, denoted as BH-corrected $p < 0.05$ and number of read pairs > 4X, where X is the expected number of read pairs mapped on a randomly chosen gene pair. Unless otherwise specified, all PPIs presented in the rest of this chapter were identified based on this default threshold. To facilitate reproducible analysis, we have implemented all data processing and statistical test steps into an open-source software package called PROPERseqTools (Figure 1.4B) (https://doi.org/10.5281/zenodo.5009171).

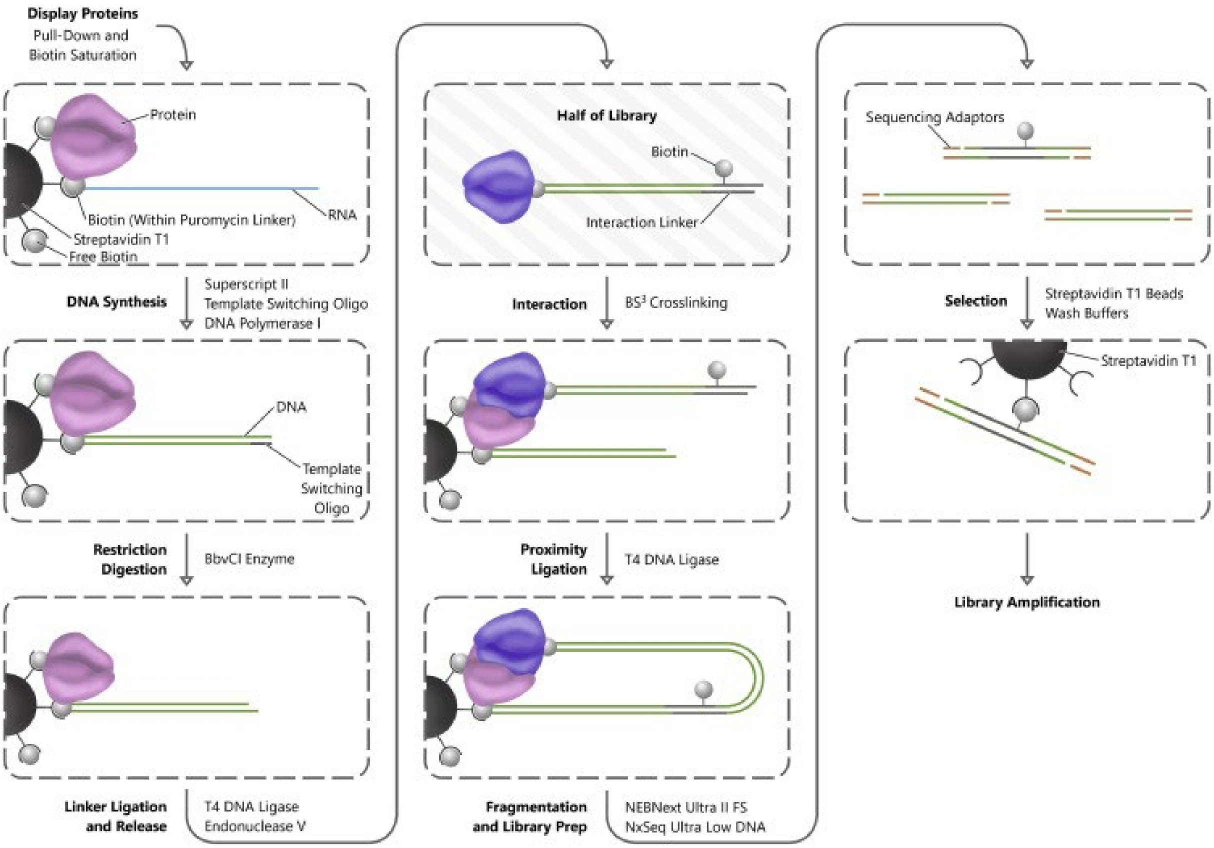
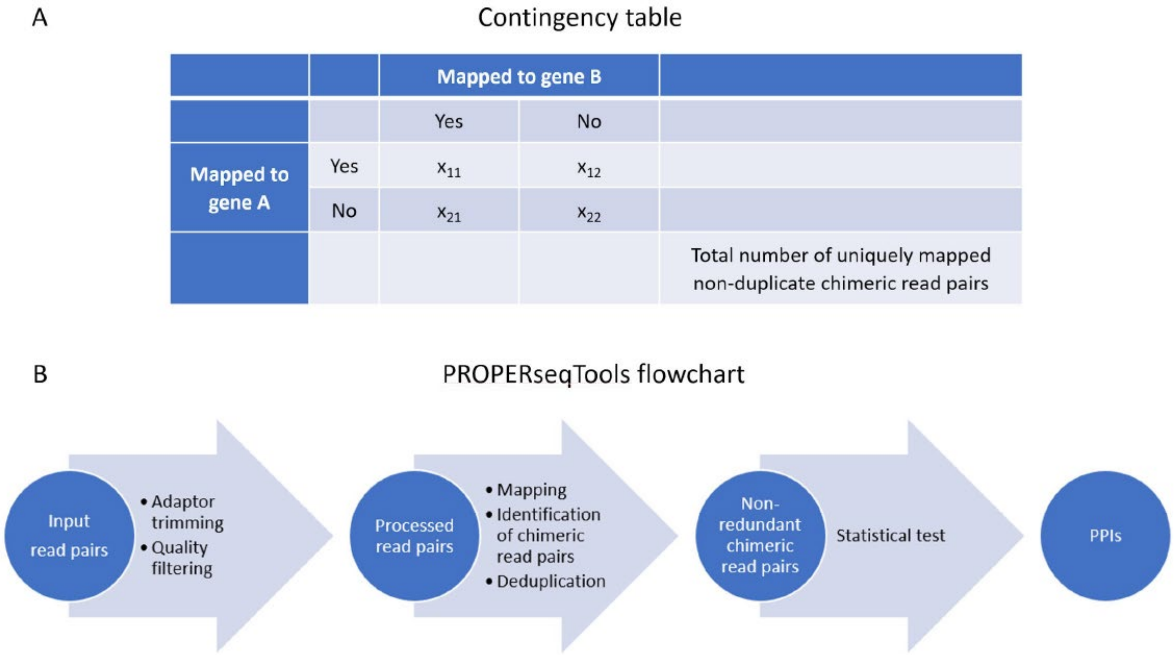


Figure 1.3: INLISE
Steps are indicated in bold font to the left of each process arrow, and primary enzymes or reagents used to accomplish each step are indicated to the right of the process arrow. The process begins with the stabilization of the display complexes on streptavidin magnetic beads. Subsequently, the RNA component of each display complex is converted to double-stranded DNA and digested with a non-palindromic restriction enzyme. The library of display proteins is then split into two populations. One half of the display protein complex is ligated to the biotinylated interaction linker and then digested to remove the complexes from the streptavidin beads. The free half of the display protein library is combined with the half still on the beads to perform the interaction step, and the interacting proteins are crosslinked. The beads are washed to remove non-specific interactions, and then proximity ligation between the display nucleic acids is performed. The DNA is then fragmented, and adaptor ligation for sequencing is performed before final streptavidin selection for the biotin-containing interaction linker and library amplification.



		Mapped to gene B		
		Yes	No	
Mapped to gene A	Yes	x_{11}	x_{12}	
	No	x_{21}	x_{22}	
				Total number of uniquely mapped non-duplicate chimeric read pairs

Figure 1.4.: PROPERseqTools
(A) A contingency table for the read pairs mapped to gene A (rows) and gene B (columns). Every mapped read pair is assigned to one and only one cell in this contingency table. The null hypothesis is that the mapping of a read pair to one gene (gene A) is independent of the mapping of this read pair to the other gene (gene B), where a read pair is considered mapped to a gene when either end of this read pair is mapped to that gene.
(B) Flowchart of PROPERseqTools for processing PROPER-seq data. Linker sequence and adaptor sequences were trimmed (Adaptor trimming). Low-quality reads and reads that were too short were removed (Quality filtering). The resulting read pairs were mapped to RefSeq genes (Mapping), and those with the two ends mapped to two different genes were obtained (Identification of chimeric read pairs). Non-redundant chimeric read pairs were used as the input to test of association (Statistical test).

1.4 Results

1.4.1 Assessments of SMART-display and INLISE

We assessed SMART-display in two aspects. First, we asked whether the display products exhibit specificity in antibody-protein interactions. To test whether a specific PPI can be detected by using the mRNA barcode on the display protein, we used the GFP antibody and GFP protein as the tested PPI. We constructed a small SMART-display library as follows. We started from four full-length mRNAs: GFP, creatine kinase, mitochondrial 2 (CKMT2), MAPK activated protein kinase 2 (MAPKAPK2), and dihydrofolate reductase (DHFR). After the display process (Figure S1.2A and S1.2B), we mixed the resulting mRNA-protein fusions equimolarly to create a small SMART-display library. We used qPCR to quantify each mRNA in this mixture (pre-selection value), used a GFP antibody for pull-down on magnetic beads, and applied stringent washes to remove non-specific RNA-bead attachments. qPCR was then used to quantify each mRNA in the mixture (post-selection value). A greater ratio of post- to pre-selection values suggests a higher anti-GFP antibody interaction with the protein. As expected, the ratios of the other three mRNAs (CKMT2, MAPKAPK2, and DHFR) were lower than that of the GFP mRNA (Figure S1.2C). This test suggests that the display protein can be specifically recognized by its antibody and that the mRNA barcodes could provide a quantitative readout of the PPIs.

Second, we evaluated the proportion of mRNAs from the original sample that were converted to display complexes by SMART-display. To this end, we split a population of HEK293T cells equally into two, one for RNA sequencing (RNA-seq) and the other for SMART-display, and we purified the display complexes by their protein moiety and sequenced the co-purified RNA moiety. Although the RNA-seq reads were mapped to 15,191 protein-coding genes (Transcripts per million [TPM] > 0.1), the sequencing reads from SMART-display were mapped to 14,805 protein-coding genes (TPM > 0.1) (displayed genes), 14,658 of which overlapped those revealed by RNA-seq (Figure S1.2D). This level of overlap in the detected mRNAs is comparable to that between two RNA-seq experiments carried out with the same cell type (Li et al. 2014; Su et al. 2014). Thus, the SMART-display-generated product library largely recapitulated the diversity of mRNAs from input cells. We subjected two HEK293T samples to the SMART-display. The samples yielded 14,805 and 14,104 displayed genes (Figure S1.2E), with 13,835 genes overlapping (odds ratio = 274.8, p < 10−32, Chi-square test), suggesting limited variation between two SMART-display repeats.

Several experimental steps in INLISE were designed to promote the formation of chimeric sequences. To test whether this design goal was achieved, we carried out the INLISE procedure with two variations, one with the interaction linker excluded (no-linker column, Figure S1.4A) and the other with the bait library pre-incubated with proteinase (proteinase column, Figure S1.3A). Compared with the standard INLISE procedure, both variations yielded less DNA in the second-to-last step (streptavidin T1 selection) (Figure S1.3A and S1.3B) and final sequencing libraries with lower concentrations (Figure S1.3C and S1.3D). These results suggest that INLISE's experimental steps improved the efficiency of forming chimeric sequences, in line with our design goal.

1.4.2 Evaluations of PROPER-seq identified PPIs

We evaluated PROPER-seq identified PPIs based on their reproducibility, precision, and recall. To test these properties, we generated six PROPER-seq libraries from HEK293T cells, Jurkat cells, and human umbilical vein cells (HUVECs). Two biological replicates from each cell type were used to generate two libraries of that cell type. These libraries are named HEK1, HEK2, JKT1, JKT2, HUVEC1, and HUVEC2 (Table S1.1). Sequencing of these libraries yielded approximately 350 million read pairs per library. Among these, approximately 8 million are non-duplicate chimeric read pairs, each mapped to two different coding genes (Table S1.1). By using PROPERseqTools with default threshold, we identified

62,637 PPIs from HEK1, 51,611 PPIs from HEK2, 41,516 PPIs from JKT1, 40,879 PPIs from JKT2, 21,494 PPIs from HUVEC1 and 33,716 PPIs from HUVEC2.

Reproducibility between biological replicates

We tested the reproducibility of PROPER-seq identified PPIs between the two biological replicates of each cell line. A total of 34,244 PPIs were shared between HEK1 and HEK2 (odds ratio = 14,242, $p < 2.2 \times 10{-}16$, Chi-square test) (Figure S1.4A), suggesting significant overlap between experimental repeats. We also tested how sensitive the reproducibility is to the threshold applied for PPI calling. We started from the default threshold and then varied the threshold (BH-corrected $p < 0.05$, number of read pairs > nX) by changing n from 4 (default) to 40 (Figure S1.4B and S1.4C). As the criterion (n) increased, the number of identified PPIs decreased as expected. However, the relative size of the overlap exhibited a monotonic increase (Figure S1.4C). These data suggest that the reproducibility of PROPER-seq increases as the threshold increases. We repeated these analyses with the two Jurkat libraries and the two HUVEC libraries and detected a similar increase in reproducibility, evident by the monotonic increase of the proportions of the overlaps as the threshold increases (Figure S1.4D–S1.4I). These results indicate that among the statistically significant PPIs, the more read pairs supporting a PPI, the more likely this PPI is to be reproducible by another repeat experiment.

Precision and recall of PROPER-seq identified PPIs

Next, we evaluated the precision and recall (Saito and Rehmsmeier 2015) of the PROPER-seq-identified PPIs with reference to known PPIs. We obtained reference datasets from the Agile Protein Interactomes DataServer (APID) (Alonso-Lopez et al. 2019; Alonso-Lopez et al. 2016), which has integrated experimentally reported PPIs from more than 6,689 curated articles and BIND (Bader, Betel, and Hogue 2003), BioGRID (Stark et al. 2006), DIP (Xenarios et al. 2000), HPRD (Peri et al. 2003), IntAct (Hermjakob et al. 2004), and MINT (Licata et al. 2012) databases. Based on this most up-to-date archive of PPIs (Alonso-Lopez et al. 2019), three types of non-binary assays yielded more than 10,000 PPIs per experimental type. These are AP-MS, coIP, and liquid chromatography-mass spectrometry (LC-MS), which have reported 131,224, 50,290, and 33,195 human PPIs, respectively (Table S1.2). These were compared with 109,539 PPIs identified in two merged PROPER-seq libraries from HEK. We plotted the precision and recall using the collection of all human coding genes as the search space (Venkatesan et al. 2009) and generated a dataset by permutating the assignment of chimeric read pairs to gene pairs. The precision-recall curve of this permutated dataset (gray dots, Figure S1.5A-S1.5C) is far beneath that of the actual data (black dots, Figure S1.5A-S1.5C), confirming that PROPER-seq's read pairs were distinguished from the background of randomly sampled gene pairs. We repeated these analyses with PROPER-seq data from Jurkat and HUVEC, using the merged data of two replicates (Figure S5) or each replicate separately (Figure S1.6). In all analyses, increases in thresholds resulted in larger precisions and smaller recalls (Figures S1.5 and S1.6). Furthermore, PROPER-seq-identified PPIs exhibited better precisions and recalls than the permutation data (Figures S1.5 and S1.6). Altogether, PROPER-seq-identified PPIs are supported by the PPIs identified by previous literature.

Identification of PPIs from PROPER-seq with control libraries and subsampling

For each of the HEK1 and HEK2 libraries, we carried out two control experiments to evaluate the noise of chimeric read pairs falsely generated by ligation errors and mapping errors. The first control, named as no-linker, has the linker sequence excluded to prevent the formation and selection of chimeric read pairs. The second control, named as proteinase, has the protein library immobilized on streptavidin beads digested. We identified 4,699,752 chimeric read pairs from the merged library of HEK1-noLinker and HEK1-proteinase, and 3,918,507 chimeric read pairs from the merged library of HEK2-noLinker and HEK2-proteinase (Table S1.1). Then we incorporated these chimeric read pairs from the control libraries into the process of identifying PPIs from the positive libraries by using an alternative association test. In detail, we subjected the number of read pairs on each gene pair that appeared in the positive library together with the number of read pairs on the corresponding gene pair in the merged control library to the alternative association test. The null hypothesis is that the mapping of a read pair in the positive library is independent of the mapping of this read pair in the control library (Figure S1.7A). While keeping the other steps and thresholds the same as in PROPERseqTools, we identified 38,977 PPIs for HEK1 with control libraries (named HEK1-wControl) and 37,948 PPIs for HEK2 with control libraries (named HEK2-wControl) under default thresholds (BH-corrected $p < 0.05$, number of read pairs > 4X).

18,752 PPIs are shared between HEK1-wControl and HEK2-wControl (odds ratio=9896.8, P-value<1e-20, fisher's exact test, Figure S1.7B). After varying the thresholds of identifying PPIs (BH-corrected $p < 0.05$, number of read pairs > nX) by changing n from 4 to 40, we found the relative size of the overlapped PPIs between HEK1- wControl and HEK2-wControl exhibited monotonic increase. Such an increase in reproducibility, quantified by odds ratio, is comparable to that between HEK1 and HEK2 without control libraries (Figure S1.7C). These data suggest a similar reproducibility of PROPER-seq identified PPIs regardless of control libraries.

The HEK1 and HEK2 positive and control libraries were then merged to infer the protein interactome in HEK cells. We identified 86,338 PPIs from the merged positive and control libraries under default thresholds, referred as HEK-wControl (BH-corrected $p < 0.05$, number of read pairs > nX). We found HEK- wControl and HEK shared an increased level of overlap of PPIs as we increased the positive read count threshold (Figure S1.7D). We computed precisions and recalls for HEK-wControl using AP-MS, coIP, and LC-MS as the reference dataset. We found HEK-wControl exhibits similar precisions and recalls as those of HEK without control libraries (Figure S1.7E-S1.7G). This suggests that HEK-wControl captures a consistent level of known PPIs as HEK without control libraries. In general, PROPER-seq identified PPIs exhibit robust reproducibility, precision, and sensitivity regardless of control libraries.

We also subsampled 75%, 50% and 25% of the input reads from HEK1 and HEK2 libraries to ask whether varying the sequencing depth of PROPER-seq libraries will affect the scale, reproducibility, precision, and recall of the identified PPIs. We found the number of identified PPIs decreases as the subsampling rate decreases (Figure S1.8A). Meanwhile, the reproducibility between HEK libraries at different subsampling rates all exhibits a monotonic increase with an increased threshold of the number of read pairs from 4X to 40X, a trend similar to that of the original HEK libraries (Figure S1.8B). We then computed precision and recall values of the merged HEK libraries at different subsampling rates using AP-MS, coIP, and LC-MS as the reference dataset. We found that the resulting precision-recall curves overlap with each other (Figure S1.8C-S1.8E). We repeated the same subsampling analyses with the two Jurkat libraries and the two HUVEC libraries. For all these libraries at different subsampling rates, they exhibit similar reproducibility and have overlapped precision-recall curves (Figure S1.8F-S1.8O). These results suggest that the sequencing depth of PROPER-seq libraries only affects the scale of identified PPIs but does not affect their reproducibility, precision, and recall. In other words, increasing the sequencing depth of PROPER-seq libraries will reveal more protein-protein interactions of equal validity.

1.4.3 PROPER v.1.0: An extensive human PPI network

To generate a comprehensive human PPI network, we combined all six PROPER-seq libraries (HEK1, HEK2, JKT1, JKT2, HUVEC1, and HUVEC2) into one dataset, composed of approximately 1.4 billion read pairs. This combined dataset revealed 210,518 pairwise PPIs involving 8,635 proteins, which are collectively termed the PROPER v.1.0 network (Figure 1.5A). We have developed a web interface to download, search, and visualize PROPER v.1.0 (https://genemo.ucsd.edu/proper).

To evaluate the topology of the network, we examined the degree distribution of PROPER v.1.0 (Barabasi 2009; Barabasi and Bonabeau 2003; Navlakha et al. 2014). The proportion of proteins (nodes) is inversely correlated with the number of interactions (edges) (Figure 1.5B), suggesting that PROPER v.1.0 is a scale-free network (Barabasi 2009; Barabasi and Bonabeau 2003). A major characteristic of scale-free networks is that they contain a small proportion of highly connected nodes, called hubs (Barabasi 2009; Barabasi and Bonabeau 2003). For example, poly(ADP-ribose) polymerase 1 (PARP1), a key regulator of various biological processes, emerged as a hub of PROPER v.1.0 by participating in 605 PPIs (edges) (Figure 1.5B and 1.7A). PROPER v.1.0's clustering coefficient (C(k)) exhibits a reverse correlation to the degree (k) (Figure S1.9I), which is in line with hierarchical networks' C(k) distributions (Barabasi and Oltvai 2004). PROPER's C(k) approaches 1 when k becomes small, suggesting that the nodes with small degrees are embedded in highly connected neighborhoods.

We asked whether functional groups are enriched in PROPER v.1.0. We plotted the enrichment level of every biological process Gene Ontology (GO) term in PROPER v.1.0 against the total number of human genes of that GO term (Figure 1.5C). To avoid generic GO terms that involve too many genes, we focused our analysis on GO terms that contained no more than 300 genes (green dots, Figure 1.5 C). The most enriched GO terms were translation (Bonferroni-corrected $p < 9.4 \times 10-51$) and RNA splicing (Bonferroni-corrected $p < 8.9 \times 10-41$) (Figure 1.5C). By intersecting PROPER v.1.0 with each GO term, we obtained a subnetwork associated with each GO term, including a translation subnetwork and an RNA splicing subnetwork. Considering the successes of previous research in elucidating the central dogma, we expected large fractions of the PPIs in the translation and the RNA splicing subnetworks to be known PPIs. Indeed, the translation subnetwork included 2,520 PPIs, where 1,185 PPIs (47%) overlapped APID-documented PPIs (Figure 1.5D). The RNA splicing subnetwork included 2,081 PPIs, where 468 PPIs (23%) overlapped APID-documented PPIs (Figure 1.5E).

Following Yu et al. (2008) and Venkatesan et al. (2009) (see also Cusick et al., 2009) (Yu et al. 2008; Venkatesan et al. 2009; Cusick et al. 2009), we calculated the screening completeness, sampling sensitivity, assay sensitivity, overall sensitivity, and precision of PROPER v.1.0 (Venkatesan et al. 2009) (Table S1.3). PROPER v.1.0’s sequencing reads covered 16,305 human protein coding genes, of which 8,635 protein coding genes were involved in PROPER v.1.0’s PPIs (Table S1.2). We further tested whether PROPER v.1.0 is enriched with either binary or non-binary PPIs by comparing with three pairs of binary and non-binary PPIs, namely, APID binary versus APID non-binary, Lit-BM-13 versus Lit-NB-13 (Kovacs et al. 2019), and L3-BM versus L3-NB (Kovacs et al. 2019) (Table S1.2). Association tests suggested enrichments of non-binary PPIs in PROPER v.1.0 ($p < 2.2 \times 10{-}16$, $p = 0.081$, and $p = 9.8 \times 10{-}9$, Chi-square tests with the three pairs of binary and non-binary datasets). These results are consistent with our expectation that PROPER v.1.0 includes both binary and non-binary PPIs, because both binary and multiway interactions are allowed when the two display libraries are incubated at the INLISE step. Altogether, PROPER v.1.0 expands the profile of human protein interactome with more than 200,000 previously uncharacterized PPIs.

Support of 17,638 computationally predicted PPIs by PROPER v.1.0

A genome-wide structure-based prediction of human PPIs was accomplished based on the prePPI (predicting protein-protein interactions) algorithm (Zhang et al. 2012; Zhang et al. 2013). Among the 1,273,679 computationally predicted and previously uncharacterized human PPIs (previously uncharacterized prePPIs) that currently do not have experimental support (not recorded in the APID database), 17,638 previously uncharacterized prePPIs appeared in PROPER v.1.0 (1.38% of the previously uncharacterized prePPIs, 8.38% of PROPER v.1.0, odds ratio = 14.83, $p < 2.2 \times 10{-}16$, Chi-square test). We also examined whether the PROPER-seq-supported prePPIs were enriched with predicted domain-domain or domain-peptide interactions (Zhang et al. 2012; Chen et al. 2015; Garzon et al. 2016). As expected, PROPER-seq-supported PrePPIs exhibited smaller structure scores that reflect a direct interaction between two protein domains (Zhang et al. 2012; Chen et al. 2015; Garzon et al. 2016) compared with the entire prePPI (Figure S1.10D). This is because the prePPI algorithm used the structure score as an important component to predict what protein pairs can interact (Zhang et al. 2012). However, the PROPER-seq-supported prePPIs exhibited a distribution of domain-peptide scores (Zhang et al. 2012; Chen et al. 2015; Garzon et al. 2016) similar to that of the entire prePPI (Figure S1.10H), suggesting little difference in domain-peptide interactions between computationally derived and PROPER-seq-supported PPIs.

Correlation between human synthetic lethal (SL) gene pairs and human PPIs

We asked whether human genetic interactions exhibit a correlation with physical interactions. To this end, we compared human SL gene pairs identified by DAISY (data mining synthetic lethality identification pipeline) (Jerby-Arnon et al. 2014; Lee et al. 2018) with three sets of human PPIs, namely, PROPER v.1.0, APID, and HuRI (Luck et al. 2020). DAISY included 2,816 SL pairs (Jerby-Arnon et al. 2014), whereas PROPER v.1.0, APID, and HuRI contained 210,518, 322,260, and 52,544 human PPIs, respectively. DAISY and PROPER v.1.0 shared 100 gene pairs (odds ratio = 27.6, $p < 2.2 \times 10-16$, hypergeometric test) (Figure S1.11A), DAISY and APID shared 74 gene pairs (odds ratio = 13.2, $p < 2.2 \times 10-16$, hypergeometric test), and DAISY and HuRI shared 4 gene pairs (odds ratio = 4.2, p = 0.015, hypergeometric test). Although the association between DAISY and HuRI was weaker than DAISY's associations with PROPER v.1.0 and APID, all three comparisons revealed positive associations. These data suggest a positive correlation between human SL gene pairs and human PPIs.

Next, we tested whether the hubs (proteins with many interactions) and the other nodes of PROPER v.1.0 are equally likely to participate in synthetic lethality. To this end, we identified the 121 nodes in PROPER v.1.0 that are involved in the human SL pairs (SL nodes) (Figure S1.11A). The SL nodes exhibited an average degree of 538 in PROPER v.1.0, far above the average degree of the entire PROPER v.1.0 ($p < 2.2 \times 10-16$, Kolmogorov-

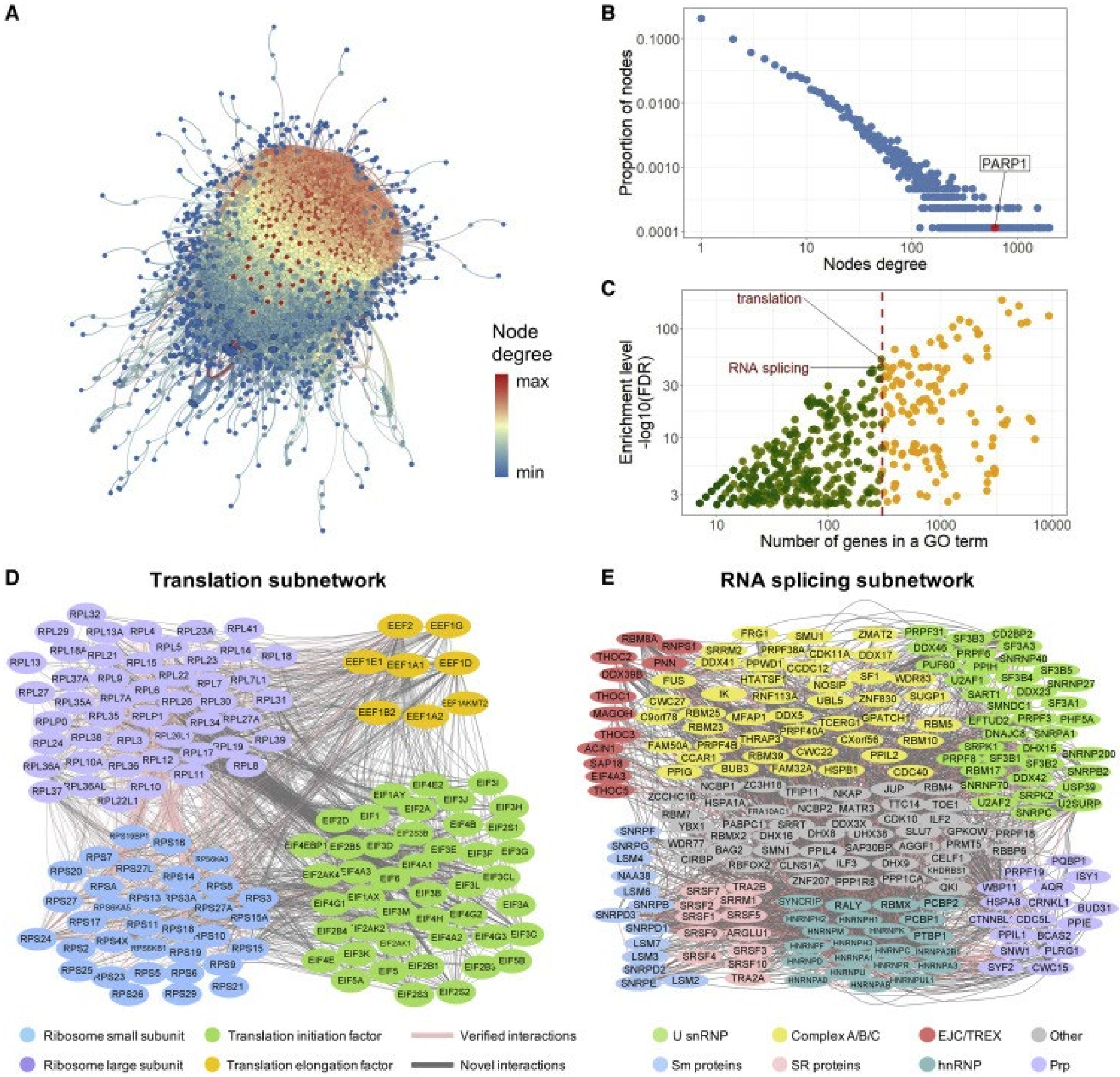
Smirnov test) (Figure S1.11B). These data suggest that the human genes involved in SL tend to be the hubs of the human PPI network, in line with the notion that the hubs of a scale-free network are more important than the other nodes for maintaining the integrity of the network (Buldyrev et al. 2010).

Cell-type-associated subnetworks

When we designed PROPER-seq, we did not anticipate it to be sensitive enough to reveal cell-type differences. After evaluating PROPER v.1.0 (the integrated result from three input cell lines), we tested whether the cell-type-specific gene expression could lead to the differential contribution of PROPER-seq data from each cell type to the identified PPIs in PROPER v.1.0. We tested this possibility at two levels, namely, for every PPI and for every subnetwork (as defined by GO terms). At the level of individual PPIs, approximately 33% of PROPER v.1.0's PPIs were identified primarily due to the read pairs from a specific cell type, including approximately 14,000 (6.8%), 25,000 (12%), and 29,000 (14.1%) PPIs attributable to HEK, Jurkat, and HUVEC data, respectively (Figure 1.6A).

At the subnetwork level, we obtained 431 subnetworks by extracting the nodes in PROPER v.1.0 associated with each GO term and the edges connecting the extracted nodes. We quantified the association of each subnetwork to each cell type by the proportions of PPIs (edges) attributable to that cell type. Most subnetworks (402 of 431) did not preferentially associate with any one of the three cell types (clustered at the center, Figure 1.6B), consistent with the idea that most biological processes as defined by GO terms are shared across these cell types. Specifically, no subnetwork exhibited preferential association with HEK (top corner, Figure 1.6B). The T cell activation and positive regulation of T cell proliferation subnetworks emerged as the top 2 subnetworks with the strongest associations with Jurkat cells, consistent with the T lymphocyte origin of Jurkat cells (lower-left corner, Figures 1.6B and 1.6C). Several subnetworks were associated with vascular endothelial cells, including regulation of the extracellular matrix, cell mobility, cell-matrix, and cell-substrate adhesion, and the integrin-mediated signaling pathway (lower-right corner, Figures 1.6B and 1.6D), reflecting the crucial functional properties of endothelial cells (Deanfield, Halcox, and Rabelink 2007). These data suggested a strong potential of applying PROPER-seq to reveal cell-type-specific PPIs.

Figure 1.5: PROPER v.1.0
(A) Entire PROPER v.1.0 network with proteins as nodes and PPIs as edges. The degree of nodes is color coded from high (red) to low (blue).
(B) PROPER's degree distribution, with the degree (number of connections of a node) (x axis) plotted against the proportion of nodes in that degree (y axis). Arrow, PARP1 node. The fitted probability density function of the degree distribution is proportional to k−1.076, where k is the degree.
(C) Number of genes (x axis) of each GO term (dot) versus the enrichment level of this GO term in PROPER v.1.0 (y axis). The colors of the dots show GO terms with less (green) and more (yellow) than 300 genes.
(D) Translation subnetwork.
(E) RNA splicing subnetwork, including the core components of human spliceosomes (U small nuclear ribonucleoprotein particle [snRNP]), components of the pre-spliceosome complex, the pre-catalytic spliceosome, and catalytic step 1 spliceosome (complex A/B/C), the exon junction complex (EJC), and the transcription and export complex (TREX), as well as SR proteins, Sm proteins, heterogeneous nuclear ribonucleoproteins (hnRNP), and pre-mRNA processing factors (Prp). Pink edges, known PPIs (as documented in the APID database); gray edges, previously uncharacterized PPIs.


A
Node degree
max
min
B
Proportion of nodes
0.1000
0.0100
0.0010
0.0001
1
10
100
1000
Nodes degree
PARP1
C
Enrichment level -log10(FDR)
100
30
10
3
10
100
1000
10000
Number of genes in a GO term
translation
RNA splicing
D
Translation subnetwork
RPL32 RPL29 RPL13A RPL4 RPL23A RPL41 RPL18A RPL21 RPL5 RPL14 RPL18 RPL13 RPL15 RPL23 RPL37A RPL9 RPL22 RPL7 RPL7L1 RPL27 RPL35A RPL7A RPL6 RPL26 RPL30 RPL31 RPLP0 RPL35 RPLP1 RPL34 RPL27A RPL24 RPL38 RPL3 RPL26L1 RPL19 RPL39 RPL36A RPL10A RPL36 RPL12 RPL17 RPL8 RPL11 RPL37 RPL36AL RPL10 RPL22L1
EEF2 EEF1G EEF1E1 EEF1A1 EEF1D EEF1B2 EEF1A2 EEF1AKMT2
RPS19BP1 RPS16 RPS7 RPS6KA3 RPS20 RPS27L RPS14 RPS27 RPSA RPS8 RPS13 RPS3A RPS3 RPS6KA5 RPS27A RPS15A RPS17 RPS11 RPS18 RPS10 RPS24 RPS2 RPS4X RPS6KB1 RPS19 RPS15 RPS25 RPS23 RPS5 RPS6 RPS9 RPS26 RPS29 RPS21
EIF4E2 EIF3I EIF1AY EIF2A EIF3J EIF3H EIF2D EIF1 EIF4B EIF2S1 EIF4EBP1 EIF2B5 EIF3D EIF2B3 EIF3E EIF3F EIF3G EIF2AK4 EIF4A3 EIF4A1 EIF6 EIF3CL EIF4G1 EIF1AX EIF3B EIF3L EIF3A EIF3M EIF4H EIF4G2 EIF2B4 EIF2AK2 EIF2AK1 EIF4A2 EIF4G3 EIF3C EIF4E EIF3K EIF2B1 EIF2B2 EIF5B EIF5A EIF5 EIF2S3 EIF2S2
Ribosome small subunit
Ribosome large subunit
Translation initiation factor
Translation elongation factor
Verified interactions
Novel interactions
E
RNA splicing subnetwork
RBM8A RNPS1 THOC2 PNN CDX38B THOC1 MAGOH THOC3 ACIN1 SAP18 EIF4A3 THOC5
FRG1 SMU1 ZMAT2 SRRM2 PRPF38A CDK11A DDX17 DDX41 PPWD1 CCDC12 FUS HTATSF1 NOSIP SF1 WDR83 IK RNF113A UBL5 ZNF830 SUGP1 CWC27 C9orf78 RBM25 MFAP1 DDX5 TCERG1 GPATCH1 RBM5 RBM23 PRPF40A FAM50A PRPF4B THRAP3 CXorf56 RBM10 CCAR1 RBM39 CWC22 PPIL2 PPIG BUB3 FAM32A HSPB1 CDC40
PRPF31 SF3B3 CD2BP2 DDX46 PRPF6 SF3A3 PUF60 PPIH SNRNP40 U2AF1 SF3B4 SF3B5 SNRNP27 SART1 DDX23 SMNDC1 SF3A1 LEFTUD2 PRPF3 PHF5A DNAJC8 SNRPA1 SRPK1 DHX15 SNRNP200 PRPF8 SF3B1 RBM17 SF3B2 SNRPB2 SNRNP70 DDX42 USP39 U2AF2 SRPK2 U2SURP SNRPC
NCBP1 ZC3H18 TFIP11 NKAP JUP RBM4 ZCCHC10 HSPA1A FRA10AC1 NCBP2 MATR3 TTC14 TOE1 RBM7 YBX1 PABPC1 SRRT DDX3X CDK10 ILF2 WDR77 RBMX2 DHX16 DHX8 DHX38 SLU7 GPKOW PRPF18 BAG2 SMN1 PPIL4 SAP30BP AGGF1 PRMT5 RBBP6 CIRBP RBFOX2 CLNS1A ILF3 DHX9 CELF1 ZNF207 PPP1R8 PPP1CA KHDRBS1 QKI
SNRPF SNRPG LSM4 NAA38 LSM6 SNRPB SNRPD3 SNRPD1 LSM7 LSM3 SNRPD2 SNRPE LSM2
SRSF7 TRA2B SRRM1 SRSF2 SRSF1 SRSF5 SRSF9 ARGLU1 SRSF4 SRSF3 SRSF10 TRA2A
SYNCRIP RALY RBMX PCBP2 HNRNPH2 HNRNPH1 PCBP1 HNRNPM HNRNPK PTBP1 HNRNPF HNRNPH3 HNRNPD HNRNPA1 HNRNPC HNRNPA2B1 HNRNPR HNRNPA3 HNRNPA0 HNRNPU HNRNPUL1 HNRNPAB
PQBP1 PRPF19 ISY1 WBP11 AQR HSPA8 CRNKL1 BUD31 CTNNBL1 CDC5L PPIE PPIL1 BCAS2 SNW1 PLRG1 SYF2 CWC15
U snRNP
Sm proteins
Complex A/B/C
SR proteins
EJC/TREX
hnRNP
Other
Prp

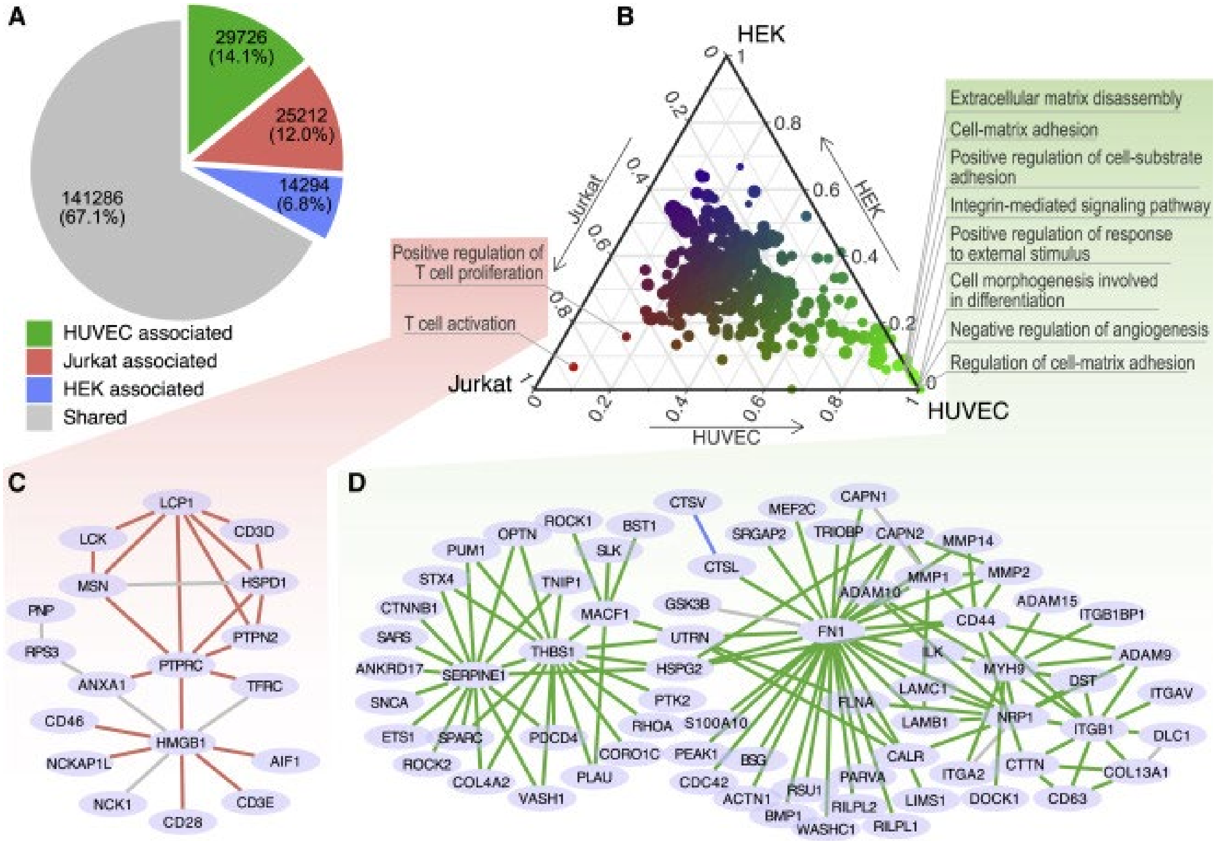


Figure 1.6: Cell-type-associated subnetworks
(A) Numbers of PPIs associated with HEK, Jurkat, and HUVEC and those that did not associate with any cell type (shared).
(B) Associations of subnetworks and cell types. The proportion of PPIs that are associated with each cell type (each axis on the edge of the triangle) in every GO-term-defined subnetwork (dot). The relative associations to the three cell types are also represented in a color gradient from red (Jurkat), to green (HUVEC), to blue (HEK). Dot size, number of genes in a GO term.
(C and D) Expanded view of the combined subnetwork of the subnetworks associated with Jurkat (C) and those associated with HUVEC (D). Edge colors denote shared PPIs (gray), as well as the PPIs associated with Jurkat (red) or HUVEC (green).

1.4.4 Experimental validation of previously uncharacterized PPIs

We subjected select previously uncharacterized PPIs to experimental validation. We first investigated whether previously uncharacterized PPIs in PROPER v.1.0 exhibit spatial proximity *in situ* by PLA (Gullberg et al. 2004; Soderberg et al. 2006), which enables direct observation of protein interactions by generating fluorescence signals specifically from interacting protein pairs in unmodified cells (Gullberg et al. 2004; Soderberg et al. 2006). We decided to choose a hub in PROPER v.1.0 and selectively test a few previously uncharacterized PPIs involving this hub. We elected several previously uncharacterized PARP1-participating PPIs, i.e., PARP1-exportin 1 (XPO1), PARP1-matrin 3 (MATR3), and PARP1-importin 5 (IPO5), to PLA tests. XPO1 and IPO5 (importin 5) regulate export and import through nuclear pores (Fornerod et al. 1997; Jäkel and Görlich 1998). MATR3 is a nuclear matrix protein.

As a positive control, we assayed for PARP1-small ubiquitin-like modifier 1 (SUMO1), a known PPI (Messner et al. 2009). The HEK293 cells co-incubated with PARP1 and SUMO1 antibodies exhibited 3 to 12 PLA foci per cell, compared with 0 to 2 foci per cell in the control cells (p = 1.1 × 10−4 for PARP1+none control, p = 1.2 × 10−6 for none+SUMO1 control, Wilcoxon test) (Figures 1.7B, 1.7C, 1.7H, and 1.7I). In parallel, cells co-incubated with PARP1 and XPO1 antibodies exhibited 13 to 34 PLA foci per cell, compared with 0 to 6 foci per cell in the cells incubated with PARP1 or XPO1 antibody alone (p = 7.4 × 10−5, for PARP1+none control, p = 7.7 × 10−5 for none+XPO1 control, Wilcoxon test) (Figures 1.7B, 1.7D, 1.7H, and 1.7J). Similarly, tests for PARP1-IPO5 and PARP1-MATR3 yielded more PLA foci per cell than their respective controls (the largest p = 1 × 10−4, Wilcoxon test) (Figures 1.7B, 1.7E, 1.7F, 1.7H, 1.7K, and 1.7L). Furthermore, all additional controls, including co-incubation of PARP1 and GFP antibodies, GFP antibody alone, and a no-antibody control, yielded fewer foci compared with the experimental groups (the largest p = 4.5 × 10−4, Wilcoxon test) (Figures 1.7B, 7G, 7M, and 7N).

We selected another previously uncharacterized PPI, PARP1-LEO1, for a coIP test. LEO1 is a component of the PAF1 complex that associates with the RNA polymerase

II (RNA Pol II) (Yu et al. 2015). In HEK293, immunoprecipitation (IP) with LEO1 antibody (Figure S1.12) resulted in coIP of PARP1 (IP/LEO1 lane and input lane, Figure 1.7O), whereas the lysates immunoprecipitated with immunoglobulin G (IgG) antibody did not exhibit any signal when immunoblotted with PARP1 antibody (IP/IgG lane, Figure 1.7O). Altogether, 4 of the 4 previously uncharacterized PPIs have been confirmed by PLA or coIP.

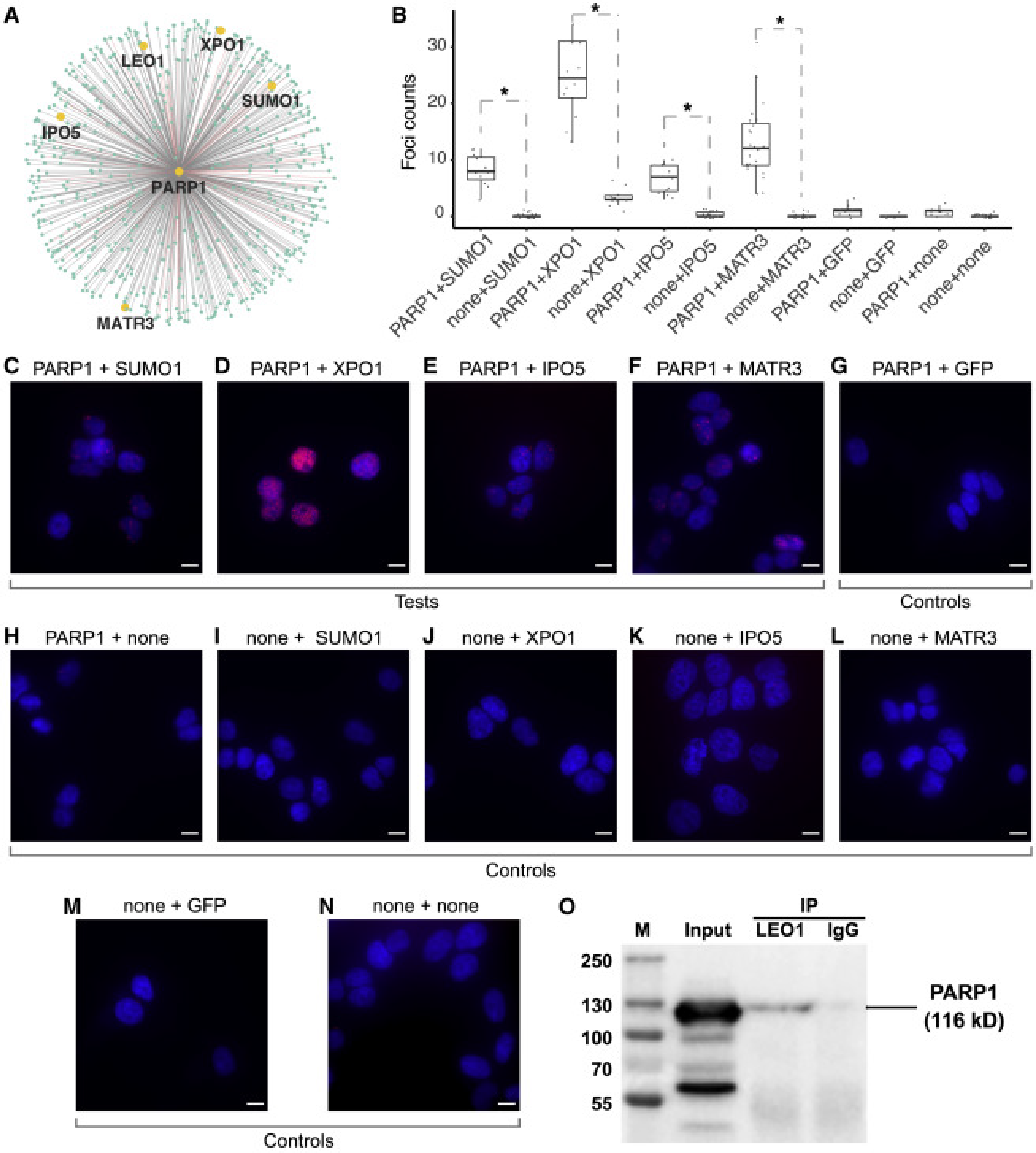
Figure 1.7: Experimental validations of previously uncharacterized PPIs

(A) 605 PPIs involving PARP1. Pink edges, known PPIs; gray edges, previously uncharacterized PPIs. The 5 PPIs tested are labeled.

(B) Boxplots of the number of PLA foci. Columns, experimental conditions, including 4 test conditions (PARP1+SUMO1, PARP1+XPO1, PARP1+IPO5, and PARP1+MATR3) and 8 control conditions (the other columns). $*p < 0.05$, Wilcoxon test.

(C–N) Representative microscopic images in each experimental condition corresponding to columns C–N in (B), with DAPI staining (blue) and PLA signals (red). Scale bar: 10 μm.

(O) CoIP analysis of PARP1 and LEO1. PARP1 immunoblots in LEO1 antibody (IP/LEO1) and IgG antibody-immunoprecipitated materials (IP/IgG). M, marker lane from a pre-stained protein ladder; input, 5% of precleared cell lysates.


A
LEO1
XPO1
SUMO1
IPO5
PARP1
MATR3
B
Foci counts
0
10
20
30
*
PARP1+SUMO1
none+SUMO1
PARP1+XPO1
none+XPO1
PARP1+IPO5
none+IPO5
PARP1+MATR3
none+MATR3
PARP1+GFP
none+GFP
PARP1+none
none+none
C PARP1 + SUMO1
D PARP1 + XPO1
E PARP1 + IPO5
F PARP1 + MATR3
G PARP1 + GFP
Tests
Controls
H PARP1 + none
I none + SUMO1
J none + XPO1
K none + IPO5
L none + MATR3
Controls
M none + GFP
N none + none
Controls
O
M
Input
IP
LEO1
IgG
250
130
100
70
55
PARP1
(116 kD)

1. 5 Discussion

PROPER-seq provides a time-effective approach to map PPIs at the transcriptome scale in a single experiment. It does not require specialized resources or reagents such as antibodies and can be applied to various input cells. Thus, PROPER-seq may be a useful profiling tool to assist users in a broad scientific community to discover PPIs relevant to many cells or tissues of interest.

The PROPER v.1.0 database expands the human protein interactome by contributing approximately 200,000 previously uncharacterized PPIs. For example, PROPER v.1.0 adds several hundred interaction partners to PARP1. Markedly, PROPER v.1.0 lends experimental support to more than 17,000 computationally predicted PPIs that have not been experimentally validated, suggesting the strong predictive ability of structure-based computational models. Furthermore, the hub proteins of PROPER v.1.0 are more likely to overlap the genes in SL gene pairs than the non-hub proteins, suggesting a connection between the human protein interactome's connectivity and the human genes' synthetic lethality.

This study has several limitations. First, PROPER-seq is an *in vitro* assay, and it may miss PPIs that rely on posttranslational modifications or *in vivo* protein localizations. Second, we have only validated a small number of previously uncharacterized PPIs, and future studies are warranted to interrogate many other previously uncharacterized PPIs. Third, we have not tested whether the DNA tags of proteins can interfere with protein-protein interactions. Fourth, we cannot rule out all possible false-positive interactions, e.g., those resulting from high-abundance proteins (Mellacheruvu et al. 2013) and protein-DNA interactions. To control for high-abundance proteins, we accounted for unligated reads belonging to each protein in the

Chi-square test; we also marked 13 PPIs in PROPER v.1.0 as potential background contaminations, because they include proteins that appear at high frequencies in negative control AP-MS experiments (Mellacheruvu et al. 2013). To minimize protein-DNA binding, PROPER-seq uses a protein-specific crosslinker, BS3, which only crosslinks amines to other amines. After crosslinking by BS3, we included multiple rounds of washes in PROPER-seq to minimize spurious binding.

This study is not designed to identify cell-type-specific interactions with statistical rigor. To identify cell-type-specific PPIs, we anticipate that future work is required to characterize the within-cell-type variation and dissect the with-cell-type variation into biological variation (among different cell sources, batches, culture conditions, and cell-cycle phases) and technical variation (among sufficient replicate experiments on the cells for which biological variation has been controlled). With within-cell-type variation fully characterized and accounted for, we anticipate that a comparison among different cell types can identify cell-type-specific PPIs.

1. 6 Supplementary information

Supplementary figures

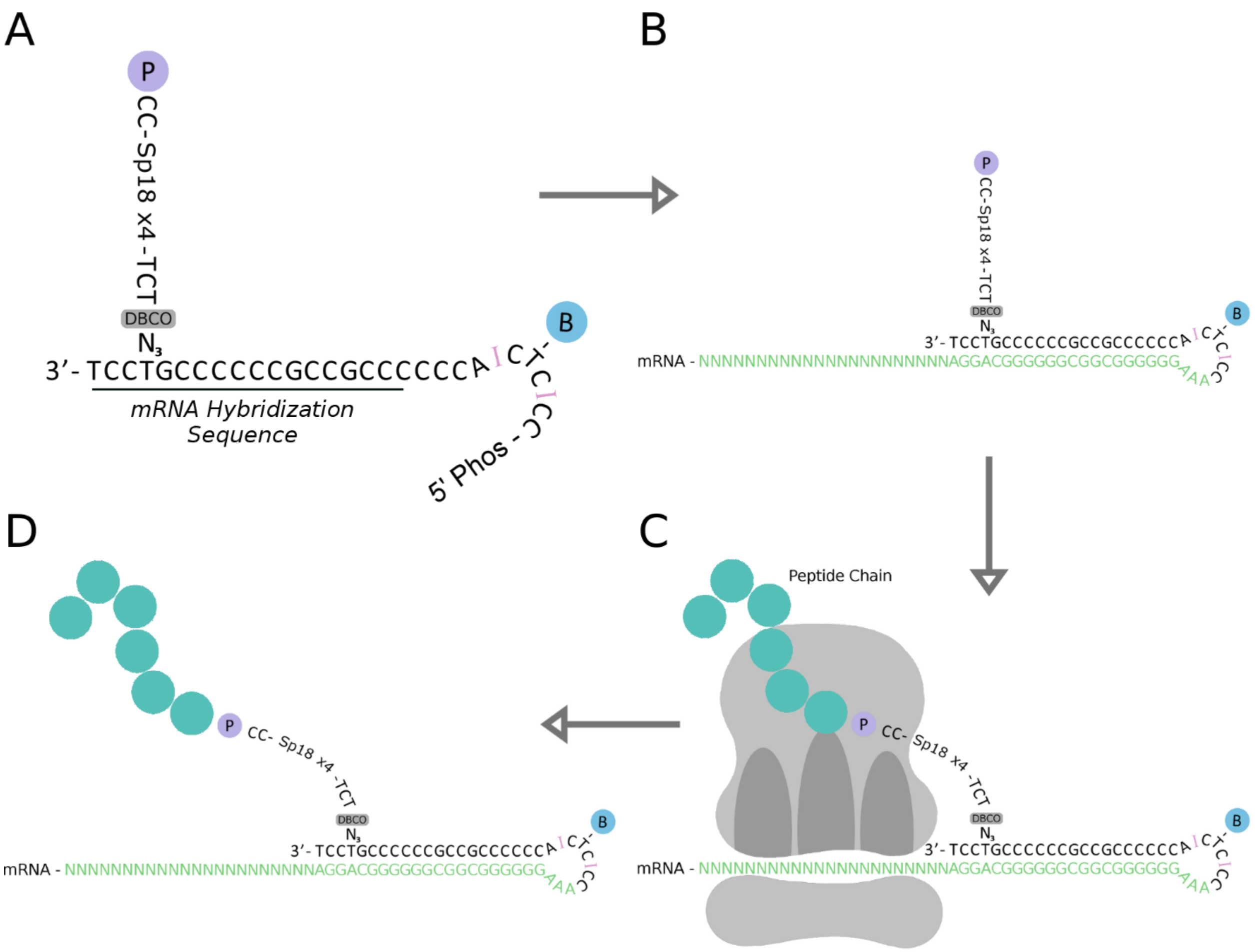
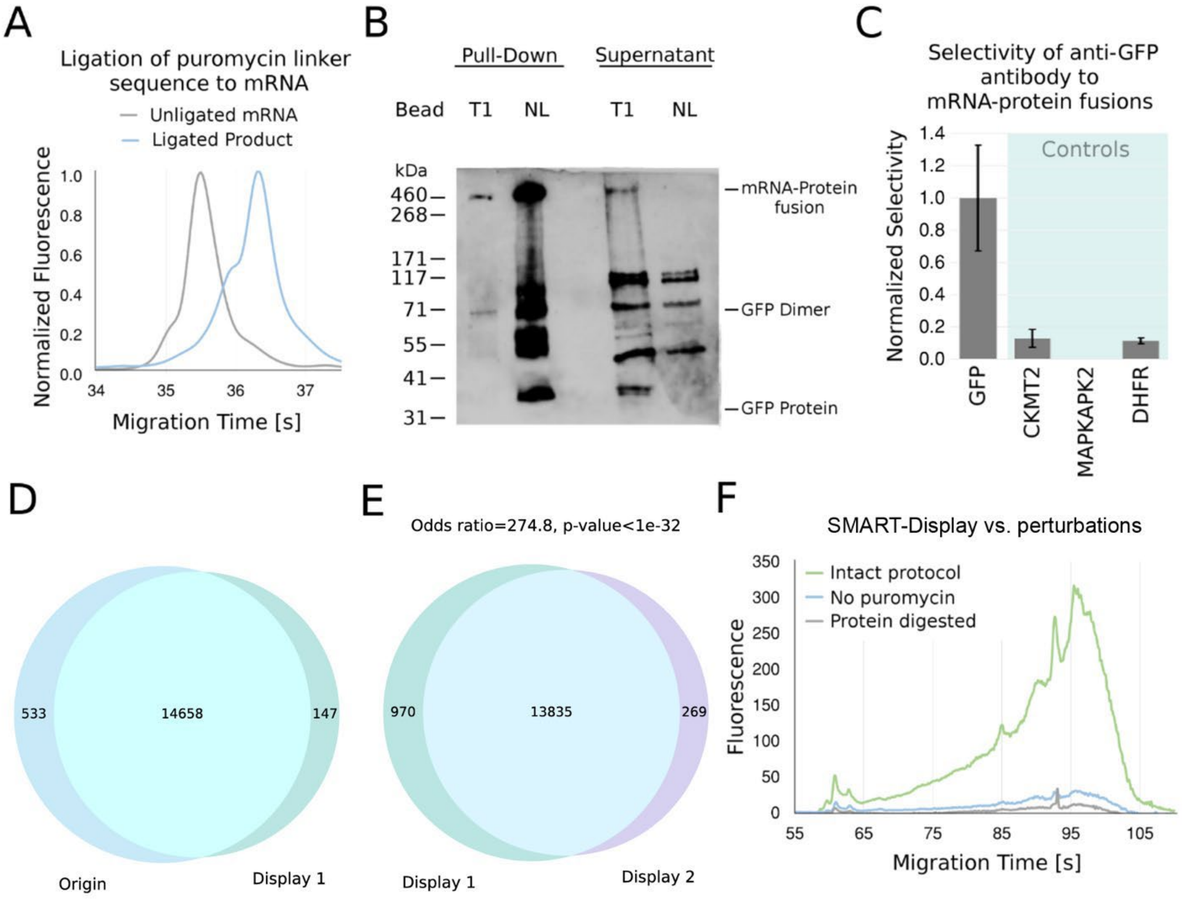
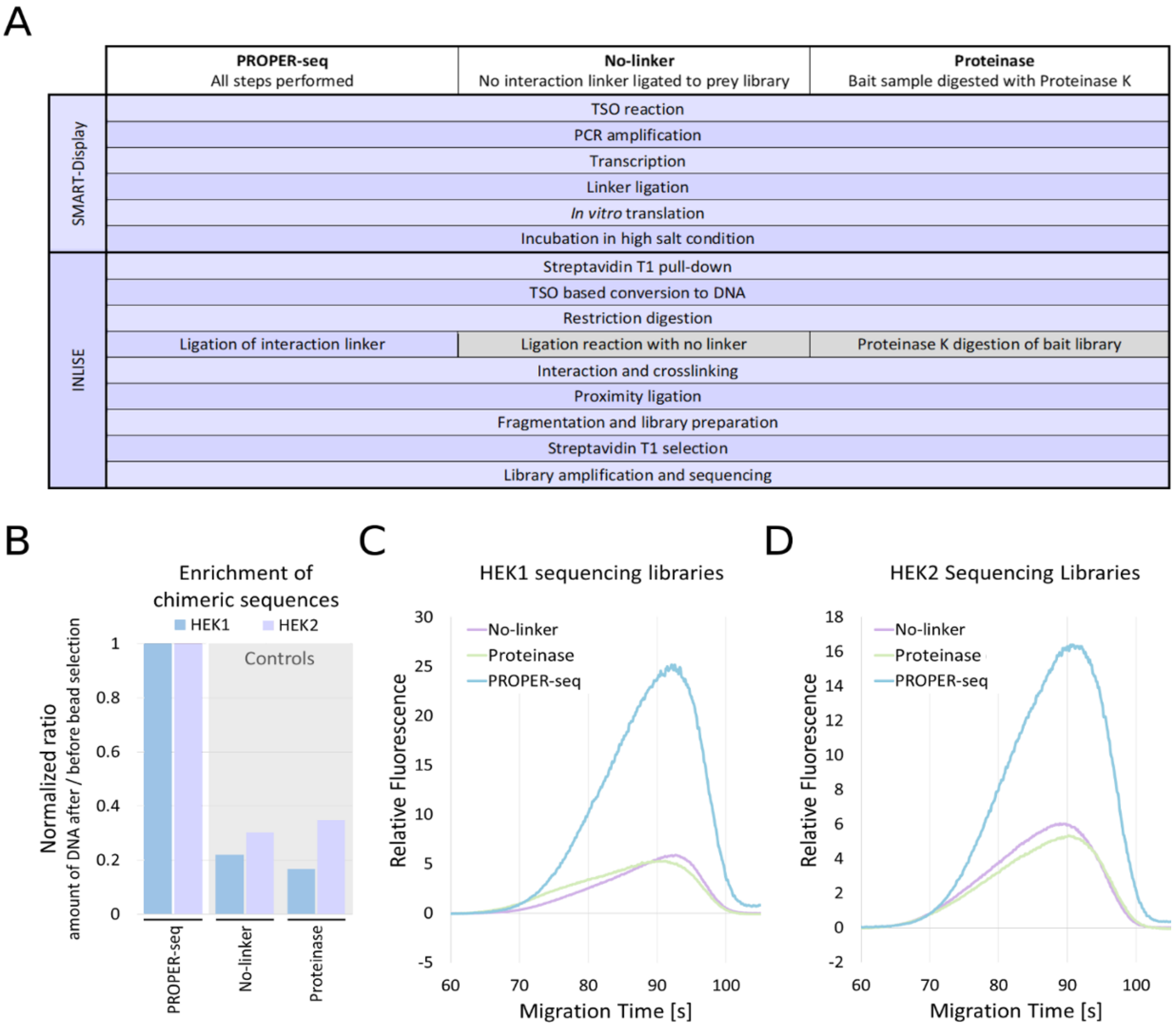

A
P
CC-Sp18 x4-TCT
DBCO
N_3
3'- TCCTGCCCCCCGCCGCCCCCCA I CTC I CC - 5' Phos
B
mRNA Hybridization Sequence
B
P
CC-Sp18 x4-TCT
DBCO
N_3
B
3'- TCCTGCCCCCCGCCGCCCCCCA I CTC I CC
mRNA - NNNNNNNNNNNNNNNNNNNNNNAGGACGGGGGGCGGCGGGGGGAAA
C
Peptide Chain
P
CC- Sp18 x4 -TCT
DBCO
N_3
B
3'- TCCTGCCCCCCGCCGCCCCCCA I CTC I CC
mRNA - NNNNNNNNNNNNNNNNNNNNNNAGGACGGGGGGCGGCGGGGGGAAA
D
P
CC- Sp18 x4 -TCT
DBCO
N_3
B
3'- TCCTGCCCCCCGCCGCCCCCCA I CTC I CC
mRNA - NNNNNNNNNNNNNNNNNNNNNNAGGACGGGGGGCGGCGGGGGGAAA

Figure S1.1: Overview of the mRNA-display process

(A) The structure of the universal puromycin-containing linker oligo, which contains puromycin ("P" in the purple circle), biotin ("B" in the blue circle), and two inosine bases ("I" in pink). The underscored sequence in this linker can hybridize with a "linker hybridization sequence" at the 3' end of a SMART-display generated mRNA that lacks a stop codon. (B) This hybridization facilitates the ligation of the 3' end of the SMART-display generated mRNA with the 5' end (5'Phos-) of this universal puromycin-containing linker sequence. (C) At the end of the translation process, puromycin enters the A-site of the ribosome and forms a covalent link with the translated peptide. (D) The fusion product is released from the ribosome.

Figure S1.2: Testing antibody specificity to displayed fusion products
(A) Size difference between unligated mRNA and puromycin-containing linker ligated mRNA. Bioanalyzer RNA Pico traces for the mRNA transcribed from a FLAG tag containing the GFP gene before (grey) and after ligation to the puromycin linker sequence (blue). Migration time (x axis) reflects fragment size. The increase in fragment size between the unligated and the ligated sequences, based on the difference in migration time, is about 100 bases. (B) Western blot of the display products. The translation outputs of the puromycin-containing linker ligated mRNA were purified with either MyOne Streptavidin T1 beads (T1 column) or with NanoLink streptavidin beads (NL column) that reacted with the biotin on the puromycin-containing linker. The released materials from the beads were blotted with an anti-FLAG antibody (T1, NL columns). The supernatants of the bead selections were blotted as controls (T1 supernatant, NL supernatant). The Streptavidin T1 beads were used in the PROPER-seq protocol. The expected size of GFP protein with a FLAG tag is approximately 27 kDa. The expected size of the display complex (GFP protein, puromycin-containing linker, and mRNA) is approximately 350 kDa. (C) Specificity of antigen-antibody interaction. The selectivity of the anti-GFP antibody was measured by the ratio of qPCR quantifications of each mRNA (column) in mixed bead purified mRNA-protein fusion products after vs. before pulling down with the anti-GFP antibody (y axis). The ratio for MAPKAPK2 was 0 because MAPKAPK2 was not detected post-selection. Error bar: standard error. (D) Venn diagram of the RNAs generated by the SMART-display process (Display 1) (Step G, Figure 2) and the original RNAs (Origin) (Step B, Figure 2). (E) Overlap of displayed genes between two repeated experiments (Display 1, Display 2). (F) Bioanalyzer traces of cDNA libraries generated from SMART-display generated fusion products (Intact protocol, green curve) and two control display libraries (blue and grey curves). One control library was generated by the same SMART-display process without ligating the puromycin-containing linker to the RNA (No-puromycin). The other control library was generated by digesting the SMART-display output library with proteinase K and removing all released contents (Protein digestion).


A
Ligation of puromycin linker sequence to mRNA
Unligated mRNA
Ligated Product
Normalized Fluorescence
1.0
0.8
0.6
0.4
0.2
0.0
34
35
36
37
Migration Time [s]
B
Pull-Down
Supernatant
Bead
T1
NL
T1
NL
kDa
460
268
171
117
71
55
41
31
mRNA-Protein fusion
GFP Dimer
GFP Protein
C
Selectivity of anti-GFP antibody to mRNA-protein fusions
Controls
Normalized Selectivity
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
GFP
CKMT2
MAPKAPK2
DHFR
D
533
14658
147
Origin
Display 1
E
Odds ratio=274.8, p-value<1e-32
970
13835
269
Display 1
Display 2
F
SMART-Display vs. perturbations
Intact protocol
No puromycin
Protein digested
Fluorescence
350
300
250
200
150
100
50
0
55
65
75
85
95
105
Migration Time [s]



	PROPER-seq All steps performed	**No-linker** No interaction linker ligated to prey library	**Proteinase** Bait sample digested with Proteinase K
SMART-Display	TSO reaction		
	PCR amplification		
	Transcription		
	Linker ligation		
	In vitro translation		
	Incubation in high salt condition		
INLISE	Streptavidin T1 pull-down		
	TSO based conversion to DNA		
	Restriction digestion		
	Ligation of interaction linker	Ligation reaction with no linker	Proteinase K digestion of bait library
	Interaction and crosslinking		
	Proximity ligation		
	Fragmentation and library preparation		
	Streptavidin T1 selection		
	Library amplification and sequencing		

Figure S1.3: Comparison of the standard INLISE procedure with two variations
(A) Flowchart of the standard protocol (PROPER-seq column) and the two variations (No-linker column and Proteinase column). (B) The ratio (y axis) of the quantities of DNA after vs. before the second last step (Streptavidin T1 selection) in the standard INLISE procedure (first column) as well as in the two variations (2nd and 3rd columns). All the ratios of a biological replicate (HEK1 or HEK2) were normalized to the ratio of the standard INLISE procedure of the same biological replicate. (C-D) Bioanalyzer traces of the sequencing library generated by the standard INLISE procedure (blue curve) and the two variations (green curves) in HEK1 (C) and HEK2 (D). The fluorescence signals are made comparable by normalizing to the concentration of the input sample (relative fluorescence, y axis).

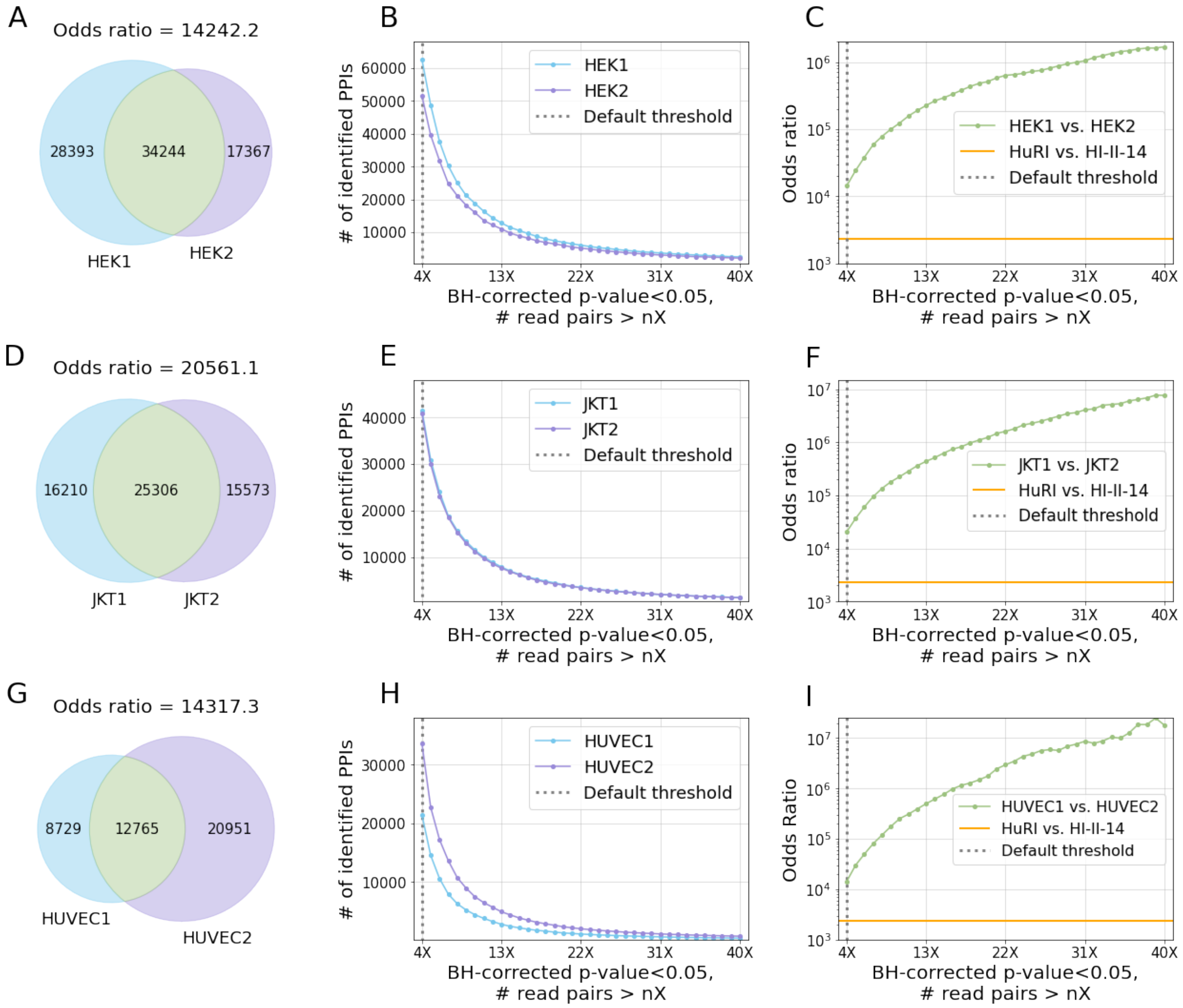


Figure S1.4: Reproducibility between biological replicates
(A) A Venn diagram of the identified PPIs from each of the two HEK293T replicates (HEK1, HEK2). (B) The number of identified PPIs (y axis) from each biological replicate (HEK1, HEK2) with respect to the criteria of calling PPIs. The criteria were BH-corrected p-value < 0.05 and # read pairs > nX, where n was changed from 4 (default, dotted vertical line) to 40 (x axis). (C) The odds ratio of the two sets of PPIs identified from the two replicates (y axis) with respect to nX. For reference, the odds ratio between HuRI and HI-II-14 is marked as a horizontal line. (D-F) The same plots as (A-C) for the two Jurkat replicates (JKT1, JKT2). (G-I) The same plots as (A-C) for the two HUVEC replicates (HUVEC1, HUVEC2).

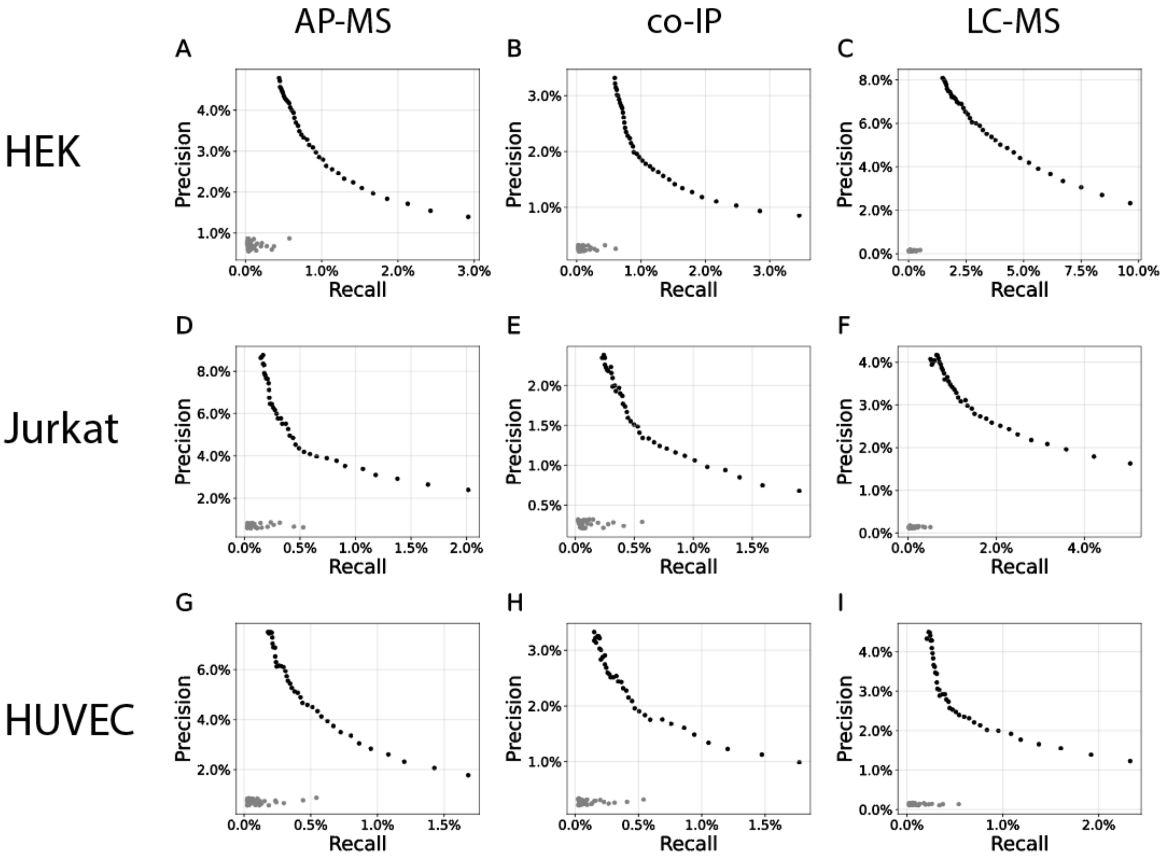


Figure S1.5: Precisions and recalls
PROPER-seq derived PPIs from HEK (A-C), Jurkat (D-F), and HUVEC (G-I) were compared to three types of known PPIs that were retrieved from APID, including all the PPIs that were identified by affinity purification mass spectrometry (AP-MS), co-immunoprecipitation (co-IP), and liquid chromatography-mass spectrometry (LC-MS) derived PPIs (columns). The precisions of recalls of the PPIs identified from PROPER-seq's permutation dataset are marked in grey dots. The permutations were based on only the genes involved in PROPER-seq detected PPIs.

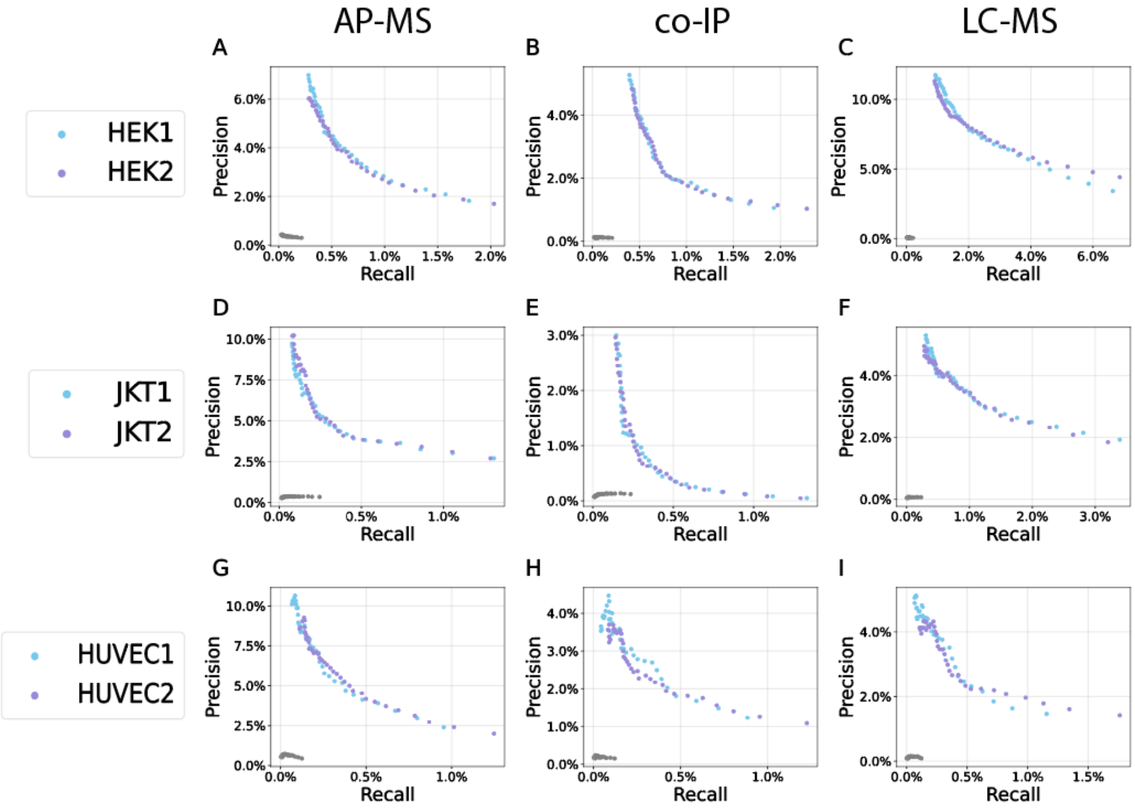


Figure S1.6: Precisions and recalls of each replicate

Precision-recall curves of PROPER-seq derived PPIs from two biological replicates of HEK293T (blue and purple dots, A-C), Jurkat (blue and purple dots, D-F), and HUVEC (blue and purple dots, G-I) compared to three types of PPIs that are derived from other experimental methods, including all the APID PPIs that are detected by affinity purification-mass spectrometry (AP-MS), co-immunoprecipitation (co-IP), and liquid chromatography-mass spectrometry (LC-MS) derived PPIs (columns). The precisions and recalls calculated from permutation data (grey dots) are included for reference. The permutations were based on only the genes involved in PROPER-seq detected PPIs.

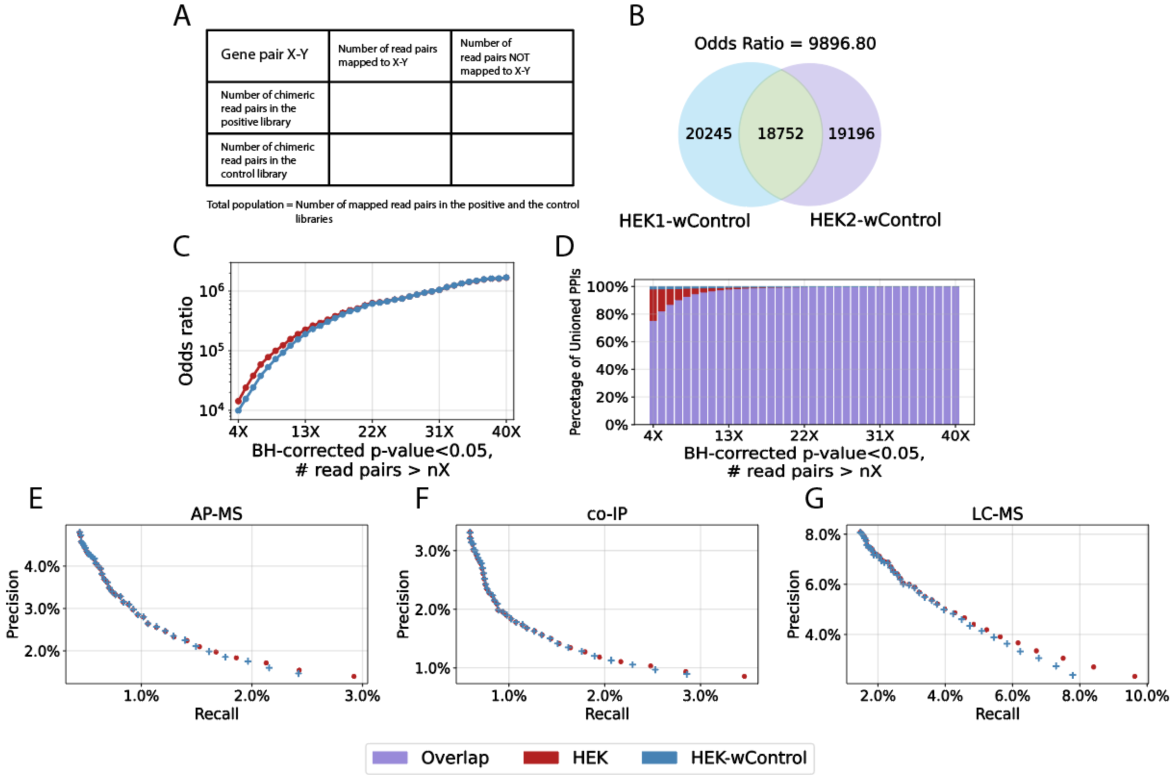


Figure S1.7: PROPER-seq identified PPIs with control libraries
(A) A continency table for the read pairs mapped to gene pair X-Y. The null hypothesis is that the mapping of a read pair to gene pair X-Y in the positive library is independent of the mapping of this read pair to the gene pair in the merged control library. (B) A Venn diagram of the identified PPIs from each of the two HEK293T replicates with control libraries (HEK1-wControl, HEK2-wControl). (C) The odds ratio of the two sets of PPIs identified from the two replicates (y axis) with respect to nX with (blue, HEK-wControl) and without (red, HEK) control libraries. (D) Percentage of overlapped PPIs (y axis, purple) with respect to nX between PPIs identified from merged HEK293T library with (blue, HEK-wControl) and without (red, HEK) control libraries. (E-G) Precision-recall curves of PPIs identified from HEK293T with (blue, HEK-wControl) and without (red, HEK) control libraries, compared to three types of PPIs that are derived from other experimental methods, including all the APID PPIs that are detected by affinity purification-mass spectrometry (AP-MS), co-immunoprecipitation (co-IP), and liquid chromatography-mass spectrometry (LC-MS) derived PPIs.

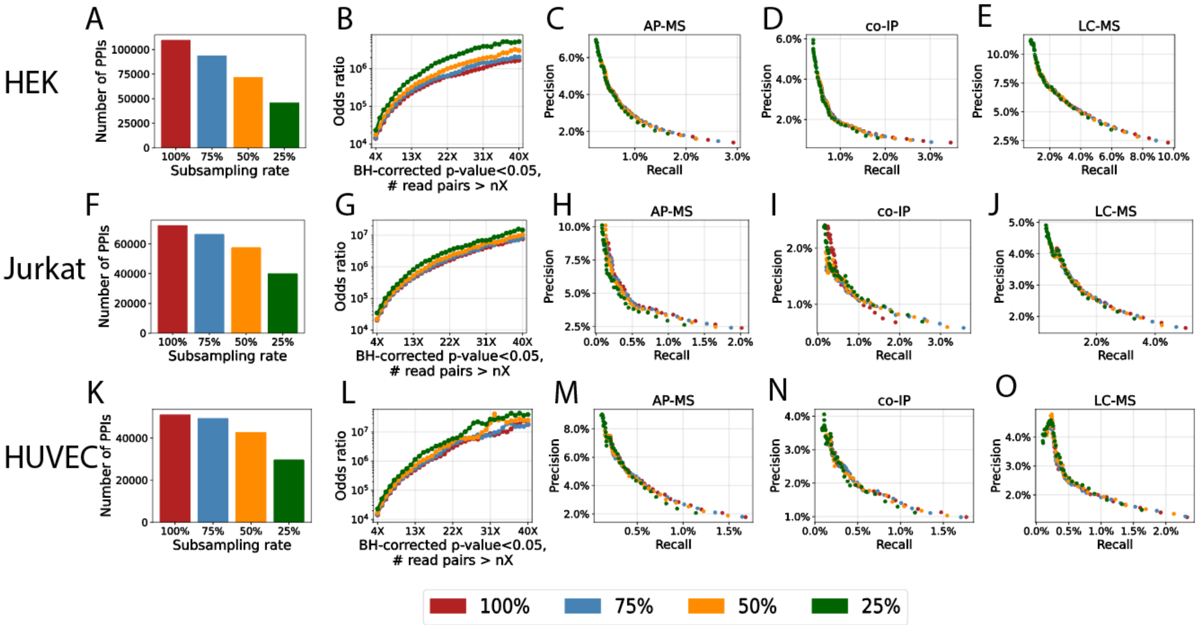


Figure S1.8: PROPER-seq identified PPIs with subsampling

Number of PPIs identified from PROPER-seq libraries of HEK (A), JKT (F), and HUVEC (K) at a subsampling rate of 100% (red), 75% (blue), 50% (orange), and 25% (green).

The odds ratio of the two sets of PPIs identified from the two replicates of HEK (B), JKT (G), and HUVEC (L) with respect to nX at a subsampling rate of 100% (red), 75% (blue), 50% (orange) and 25% (green). Precision-recall curves of PPIs identified from HEK (C-E), JKT (H-J), and HUVEC (M-O) at a subsampling rate of 100% (red), 75% (blue), 50% (orange), and 25% (green), compared to three types of PPIs that are derived from other experimental methods, including all the APID PPIs that are detected by affinity purification-mass spectrometry (AP-MS), co-immunoprecipitation (co-IP), and liquid chromatography-mass spectrometry (LC-MS) derived PPIs.

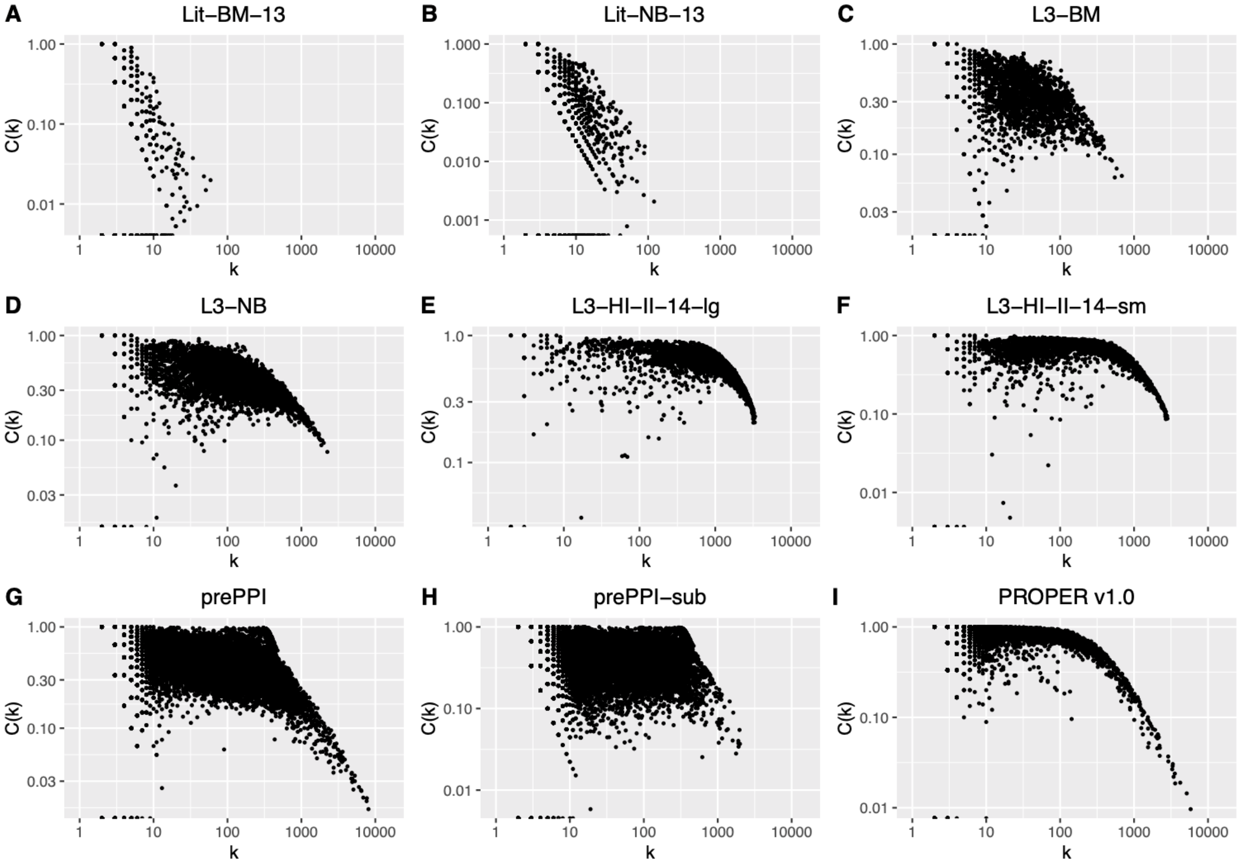


Figure S1.9: Log-log plots of clustering coefficient vs. degree
Scatterplots (log-log plots) of clustering coefficient C(k) vs. degree (k), based on (A) Binary PPIs curated by Kovacs et al. (DOI: 10.1038/s41467-019-09177-y) (Lit-BM-13), (B) Non-binary PPIs curated by Kovacs et al. (Lit-NB-13), (C) The subset of predicted binary PPIs using Lit-BM-13 as the input data by the L3 algorithm with L3 scores > 50% quantile (L3-BM), (D) The subset of predicted non-binary PPIs using Lit-NB-13 as the input data by the L3 algorithm with L3 scores > 50% quantile (L3-NB), (E) The subset of PPIs predicted using HI-II-14 as the input data by the L3 algorithm with L3 scores > 25% quantile (L3-HI-II-14-lg), (F) The subset of PPIs predicted using HI-II-14 as the input data by the L3 algorithm with L3 scores > 75% quantile (L3-HI-II-14-sm), (G) the entire prePPI, (H) the subset of prePPI with structure score > 1, and (I) PROPER v1.0.

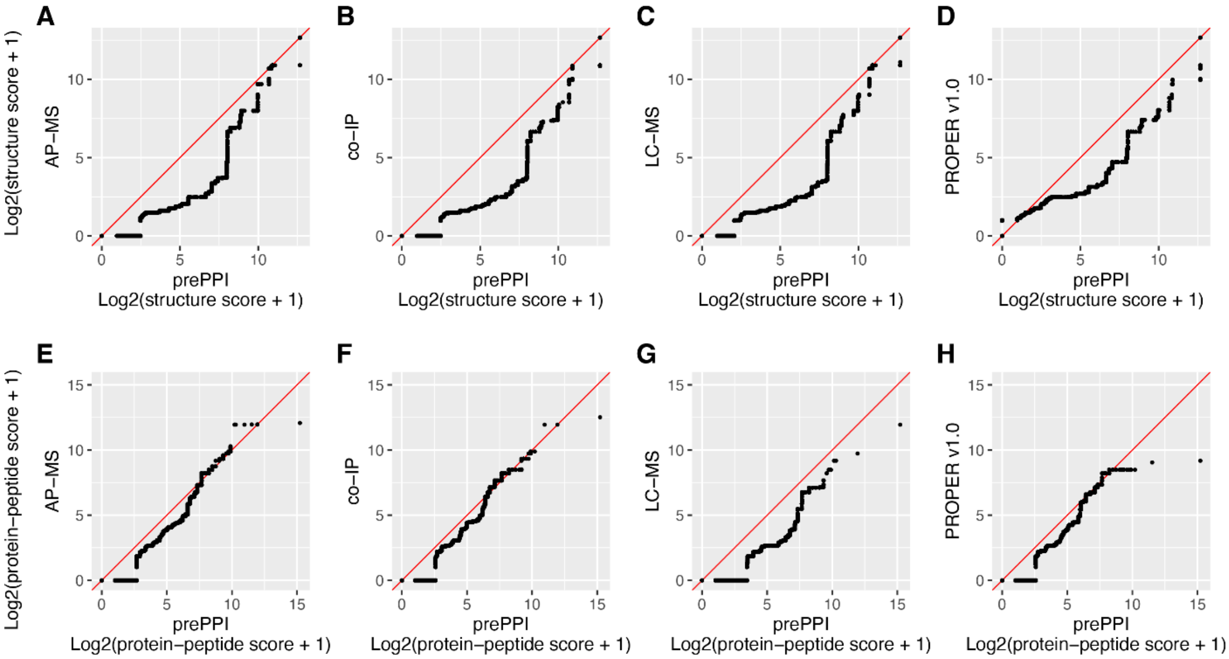


Figure S1.10: Q-Q plots

Q-Q plots of AP-MS, Co-IP, LC-MS, and PROPER v1.0 confirmed prePPIs (y axis) vs. the entire prePPI (x axis), based on structure score that reflects domain-domain interactions (A-D) and protein-peptide score that reflects domain-peptide interactions (E-H). See Table S1.1 for the descriptions of AP-MS, Co-IP, LC-MS datasets.

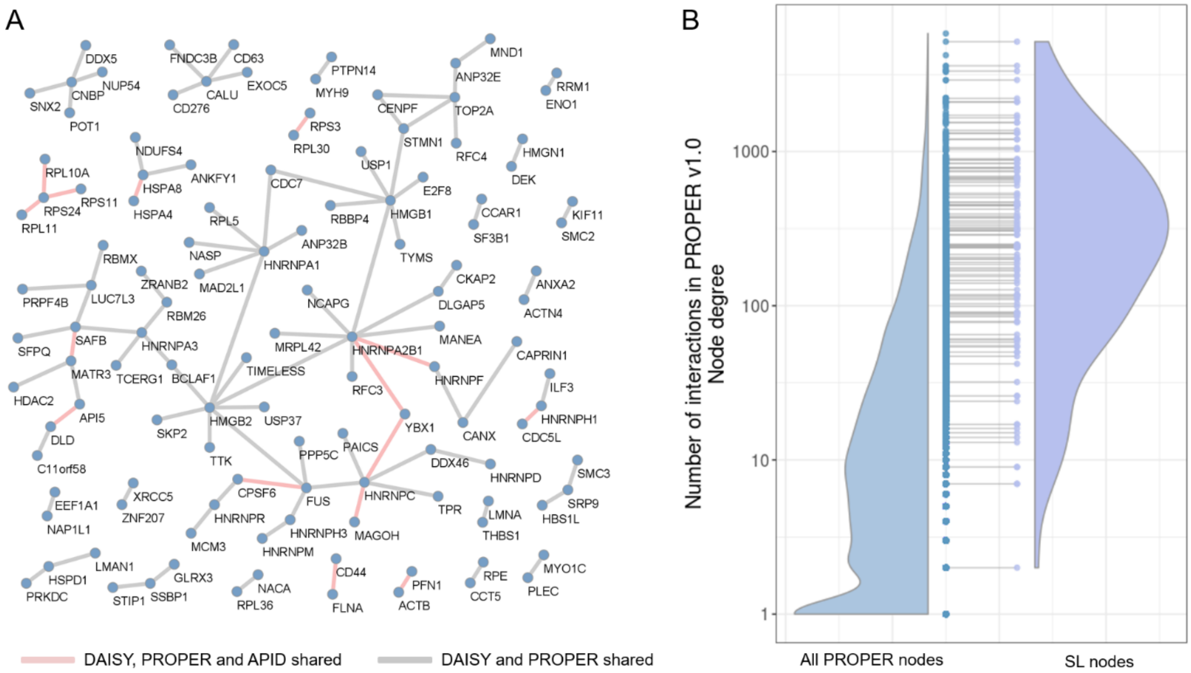


Figure S1.11: The overlap between DAISY SL gene pairs and PROPER v1.0
(A) The overlap between DAISY SL gene pairs and PROPER PPIs. Grey edge: DAISY SL gene pair that is also a PROPER PPI. Pink edge: DAISY SL gene pair that is a PROPER PPI and an APID documented PPI. (B) The degree distribution (half violin plot in blue) of all the PROPER v1.0 nodes (blue dots) vs. the degree distribution (half violin plot in purple) of all the SL nodes (purple dots). All degrees are based on the PROPER v1.0 network. The nodes with the same degrees are indicated by horizontal lines.

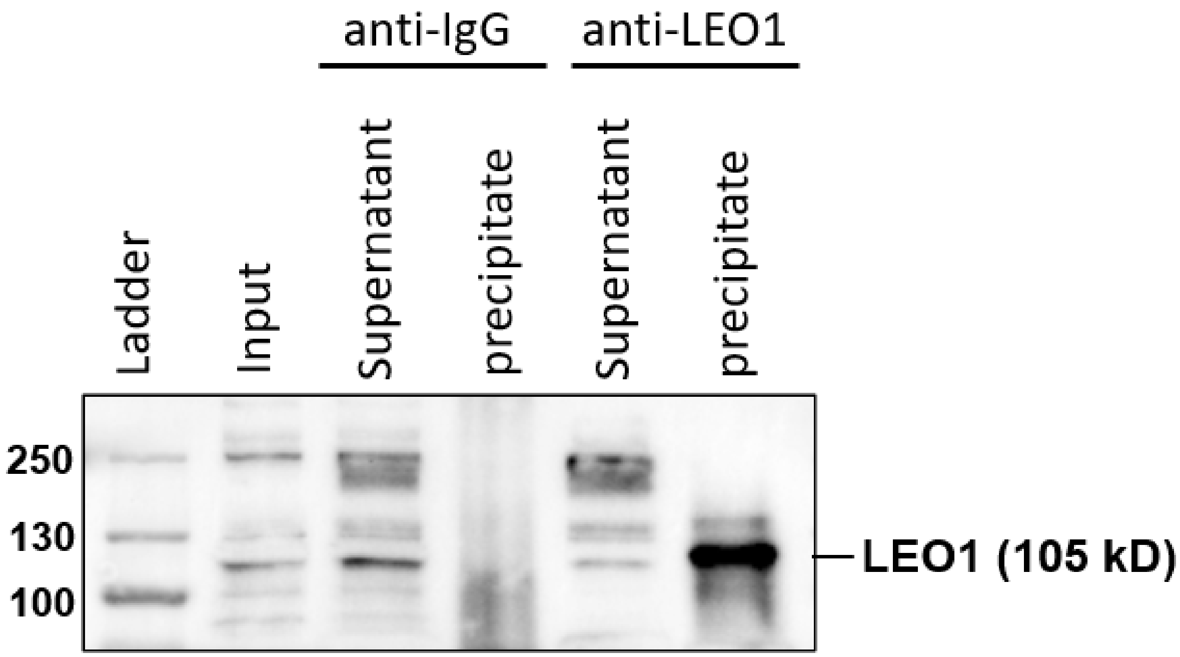


Figure S1.12: Immunoprecipitation of LEO1
HEK293T lysates were immunoprecipitated with rabbit anti-human LEO1 antibody (anti-LEO1) or anti-rabbit IgG as an isotype control (anti-IgG). Both the precipitate and the supernatant were blotted with LEO1 antibody. Ladder: pre-stained protein ladder. Input: 5% of precleared cell lysates. The precipitates were used as input in PARP1 blots (Figure 1.7O).

Supplementary tables

Table S1.1: Summary of PROPER-seq and perturbation libraries
The libraries generated at the same time were given the same experiment ID (Exp ID). The total number of read pairs, the number of non-duplicate read pairs mapped to protein coding genes and the number of non-duplicate chimeric read pairs that were mapped to two different protein coding genes were listed in the last three columns.

Library ID	Expt ID	Cell line	Number of read pairs	# of non-duplicate read pairs mapped to coding genes	Non-duplicate uniquely mapped chimeric read pairs
HEK1	1	HEK293T	343,861,373	205,881,483	12,581,208
HEK2	2	HEK293T	248,657,713	173,300,648	7,747,982
JKT1	3	Jurkat	444,413,111	262,211,890	9,988,056
JKT2	4	Jurkat	390,643,931	236,283,970	9,385,745
HUVEC1	5	HUVEC	359,807,741	194,690,153	6,404,274
HUVEC2	6	HUVEC	483,597,124	283,434,465	9,705,398
HEK1-noLinker	7	HEK293T	97,353,678	64,671,472	2,462,180
HEK1-proteinase	8	HEK293T	64,497,521	46,428,119	2,237,572
HEK2-noLinker	9	HEK293T	69,732,554	42,197,977	2,152,084
HEK2-proteinase	10	HEK293T	87,444,917	41,828,629	1,766,423

Table S1.2: The datasets used
The datasets used in this work, including APID and APID's subsets, collections of literature reported binary and non-binary PPIs, computationally predicted PPIs, PROPER-seq derived PPIs, and synthetic lethal gene pairs.

Name	**Description**	**# PPIs**	**# proteins**
APID	All the experimentally-derived human PPIs in APID, downloaded from http://cicblade.dep.usal.es:8080/APID/init.action	322,260	16,965
AP-MS	Affinity purification-mass spec detected PPIs that are included in APID	131,224	13,650
Co-IP	Co-IP detected PPIs that are included in APID	50,290	9,088
LC-MS	Liquid chromatography–mass spec detected PPIs that are included in APID	33,195	4,548
APID-binary	The experimentally derived binary PPIs curated into the APID database (level 2)	63,954	12,572
APID-non-binary	The PPIs derived from non-binary methods in the APID database	258,306	15,847
Lit-BM-13	Binary PPIs curated by Kovacs *et al. Nat Commun* 10, 1240 (2019).	4,386	3,249
Lit-NB-13	Non-binary PPIs curated by Kovacs *et al. Nat Commun* 10, 1240 (2019).	10,152	5,382
prePPI	Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 490, 556-560, doi:10.1038/nature11503 (2012)	1,279,381	16,903
prePPI-sub	Subset of prePPI with structure scores > 10.	619,619	13,222
L3-BM	The subset of predicted binary PPIs using Lit-BM-13 as the input data by the L3 algorithm (Kovacs *et al. Nat Commun* 10, 1240) with L3 scores > 50% quantile.	56,890	2,726
L3-NB	The subset of predicted non-binary PPIs using Lit-NB-13 as the input data by the L3 algorithm with L3 scores > 50% quantile.	387,971	4,694
PROPER v1.0	The PPIs derived from the merged PROPER-seq libraries of HEK1, HEK2, JKT1, JKT2, HUVEC1 and HUVEC2	210,518	8,635
HEK	The PPIs derived from merged PROPER-seq libraries of HEK1 and HEK2	109,539	7,292
Jurkat	The PPIs derived from merged PROPER-seq libraries of JKT1 and JKT2	72,409	5,136
HUVEC	The PPIs derived from merged PROPER-seq libraries of HUVEC1 and HUVEC2	51,125	4,266
		# gene pairs	**# genes**
DAISY	Jerby-Arnon, L. *et al.* Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* 158, 1199-1209 (2014).	2,802	2,077

Table S1.3: PROPER v.1.0 framework
The estimated screening completeness, sampling sensitivity, assay sensitivity, overall sensitivity, precision, and protein interactome size for PROPER v1.0.

	PROPER v1.0
Screening completeness	64.7%
Sampling sensitivity	35.4%
Assay sensitivity	43.38%
Overall sensitivity	15.36%
Precision	5.77%
Protein interactome size	8.5x10^5

Table S1.4: Contingency table to compute odds ratio to quantify reproducibility

	Within set II	Outside set II
Within set I	A	B
Outside set I	C	D

Table S1.5: Contingency table to test cell type association of a PROPER-seq derived PPI

		The read pair is generated from this cell type	
		Yes	No
Mapped to this gene pair	Yes	A	B
	No	C	D

1. 7 Materials and methods

SMART-display

mRNA Purification

Total RNA was isolated from HEK with TRIzol™ Reagent (Invitrogen, 15596026) according to the manufacturer's recommendations. Subsequently, poly-A RNAs were enriched with the Dynabeads™ mRNA Purification Kit (Invitrogen, 61006). The reduction of rRNA was evaluated against the total RNA using Agilent's Bioanalyzer RNA 6000 Pico Kit (Agilent Technologies, 5067-1513). The remaining rRNA was depleted with the Ribo-Zero H/M/R Kit (Illumina, MRZH116) or the RiboMinus Transcriptome Isolation Kit (Invitrogen, K155002) adjusting the input amount based on the estimated rRNA removed by the oligo-dT selection (For example, if rRNA was 50% depleted, input was twice as much RNA as recommended). The final quality of the RNA was assessed with Agilent's Bioanalyzer RNA 6000 Pico Kit.

Generation of DNA Library

To hybridize the Right/Random primer (5' TTT CCC CGC CGC CCC CCG TCC TGC TGC CGC CCT TGT CGT CAT CGT CTT TGT AGT C(Nx15) 3'), 0.5 pmols of mRNA, 2.33 uM primer, and 2.33 mM dNTPs were mixed in a total volume of 10.75 uLs. This reaction was brought to 72 °C for 3 minutes and then cooled to 25 °C for 10 minutes. The template switching reaction was performed by adding 250 U SuperScript II Reverse Transcriptase (Thermo Scientific, 18064014), SuperScript II First Strand Buffer (to 1x), 5 mM DTT, 20 U SUPERase•In™ RNase Inhibitor (Thermo Scientific, AM2694), 1 M Betaine (Sigma-Aldrich, 61962), 6 mM MgCl2 (Invitrogen, AM9530G), and 1 uM Library TSO (5' /5Biosg/GGC TCA CGA GTA AGG AGG ATC CAA CAT rGrGrG 3') to a total volume of 25 uLs. The reaction was incubated at 25 °C for 2 minutes, 42 °C for 50 minutes, 10 cycles of 50 °C for 2 minutes and 42 °C for 2 minutes, and 70 °C for 15 minutes. Purification was performed with 1.8x Agencourt RNAClean XP Beads (Beckman Coulter, A63987) and the product was quantified with the Qubit™ dsDNA BR Assay Kit (Invitrogen, Q32853).

Amplification of 1 ng of cDNA/RNA product was performed per 25 uL NEBNext

High-Fidelity 2X PCR Master Mix (NEB, M0541L) reaction, containing 0.5 uM Left PCR primer (5' GCG AAT TAA TAC GAC TCA CTA TAG GGC TCA CGA GTA AGG AGG 3') and 0.3 uM Right PCR primer (5' TTT CCC CGC CGC CCC CCG TC 3'). Reactions were cycled twice with a 65 °C annealing step and a 3-minute 72 °C extension step, and 13 cycles with a single 3-minute 72 °C combined annealing and extension step. Approximately 24 reactions were performed simultaneously to generate enough material for in vitro transcription; the products were co-purified with 1.8x Agencourt AMPure XP Beads (Beckman Coulter, A63881) and quantified with the Qubit™ dsDNA BR Assay Kit.

Synthesis of Puromycin containing linker

All oligo components of the puromycin containing linker were reconstituted to 1 mM with 1x PBS pH 7.2 (Thermo Scientific, 20012027). To generate the dI containing puromycin containing linker, the Biotin Arm (w/dI) (5’ /5Phos/CC/ideoxyI/ C/iBiodT/C /ideoxyI/AC CCC CCG CCC CCC CCG /iAzideN/CCT 3’) was mixed in a 1:1 ratio with the Puromycin Arm (5’ /5DBCON/TCT /iSp18/iSp18/iSp18/iSp18/CC/3Puro/ 3’). To generate puromycin containing linker without dI bases, the Biotin Arm (w/o dI) (5’ /5Phos/CCG C/iBiodT/C GAC CCC CCG CCC CCC CCG /iAzideN/CCT 3’) was mixed in a 1:1 ratio with the Puromycin Arm (5’ /5DBCON/TCT /iSp18/iSp18/iSp18/iSp18/CC/3Puro/ 3’). The mixtures were incubated at 40 °C overnight with agitation.

The mixtures were run on a 15% TBE-UREA Gel (Invitrogen, EC6885BOX) prepared in a 1:1 ratio with Formamide Running Buffer (1 part 10x TBE Buffer Running Buffer (Invitrogen, LC6675), 9 parts Deionized Formamide (EMD Millipore, 4610-100ML)) at 200V for 1 hour. The gel was removed from the cassette and was exposed to UV while on a TLC Silica gel 60 F_{254} Plate (EMD Millipore, 1.05715.0001) to visualize the DNA bands. Two bright bands appeared, the largest was removed with a clean scalpel and transferred to a clean 2 mL tube. The gel fragment was crushed with the plunger from a 1 mL syringe and suspended in 500 uLs Elution Buffer (0.5M Ammonium Acetate (Invitrogen, AM9070G), 10 mM Magnesium Acetate (Sigma-Aldrich, 63052-100ML)). The gel fragment was incubated at room temperature with rotation overnight. The gel and buffer mixture were transferred to a 0.45 uM Nanosep® MF spin filter (Pall Corporation, ODM45C33), and the liquid collected by spinning at 5,000 xg for 10 minutes. The flow through was precipitated with 0.5x volume LiCl Precipitation Solution (Invitrogen, AM9480), 6 uLs Co-Precipitant Pink (Bioline, BIO-37075), and 3x volume of 100% Ethyl Alcohol (Sigma-Aldrich, 493546) and incubated overnight at -80 °C. The linker was then pelleted by centrifugation at 22,000 xg for 20 minutes, washed with 70% Ethyl Alcohol, and air dried. The pelleted linker was suspended in nuclease-free water (Thermo Scientific, 10977023).

Generation of Puromycin Ligated RNA Library

RNA libraries were generated with 500 ngs of DNA Library using the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB, E2040S). After synthesis, DNA was removed with TURBO™ DNase (Invitrogen, AM2238). The RNA was precipitated with 2.5 M LiCl Precipitation Solution, quantified with the Qubit™ dsDNA BR Assay Kit (Invitrogen, Q32853), and the distribution checked with the Agilent RNA 6000 Pico Kit. RNA libraries were annealed to the appropriate puromycin containing linker in a 1:1.25 molar ratio in Annealing Buffer (10x: 100 mM Tris-HCI Buffer, pH 7.5 (Invitrogen, 15567027), 500 mM NaCl (Thermo Fisher Scientific, AM9759), 10 mM EDTA (Research Products International, E14100-50.0)), incubating at 75 °C for 5 minutes and cooling slowly to 25 °C. Ligation was performed with 0.4 U/uL of T4 RNA Ligase 1 (NEB, M0204S), 1 mM ATP, and 1.6 U/uL of

SUPERase• In™ RNase Inhibitor for 30 minutes at 25 °C. NEBuffer 4 was added to 1x, and unligated linker was digested with 0.2 U/uL of T5 Exonuclease (NEB, M0363S) at 37 °C for 30 minutes. The ligated RNA was purified with an RNeasy Mini Column (Qiagen, 74104).

Translation and Display

Protein products were generated using 25 pmols of ligated RNA product per 25 uL reaction of the PURExpress® In Vitro Protein Synthesis Kit (NEB, E6800S). Translation reactions were performed in an air incubator for 90 minutes at 37 °C. After translation, KCl (Invitrogen, AM9640G) and MgCl2 (Invitrogen, AM9530G) were added to a final concentration of 800 mM and 80 mM respectively. The reaction was incubated at room temperature for 30 minutes and then stored at -20 °C for a minimum of 12 hours.

VALIDATION by anti-GFP Selection

Preparation of SMART-display Library

Templates for the target genes were ordered from IDT with all display sequences already incorporated on the 5' and 3' ends of the template. From these templates, RNA was generated and SMART-display proceeded as described above.

Pull-Down with anti-GFP antibody

The products of the SMART-display process for each of the target genes were mixed in a 1:1 ratio. The mixture was precleared with 50 µL of Streptavidin T1 magnetic beads. The mixture was incubated at 4°C with gentle rotation for 1 hour. The Streptavidin T1 beads were separated with a magnetic rack for 1 minute and the supernatant was transferred into a new microcentrifuge tube placed on ice. To the precleared solution, Normal Goat Serum (NGS) (Thermo Fisher Scientific, 31873) in PBS was added to 5% for blocking. Primary anti-GFP antibody (Thermo Fisher Scientific, A10259) diluted in PBS was added to a final concentration of 0.2 µg/mL. The sample was incubated at 4°C overnight with gentle rotation. 50 uLs Streptavidin T1 magnetic beads were added to the samples and incubated at room temperature for 1 hour with gentle rotation. The tubes were placed on a magnetic rack for 1 minute and the supernatant discarded. The beads were suspended in wash buffer (5% NGS in PBS, 1% Triton® X-100, 3% BSA (NEB, B9000S) by pipetting gently up and down. The tubes were rotated gently for 10 minutes. The wash process was repeated two more times.

cDNA Synthesis

A reverse transcription reaction solution was prepared for the selected sample (immobilized on the Streptavidin T1 beads) and for the pre-selection samples. The 100 uL reactions contained 800 U SuperScript II, 1x First Strand buffer, 10 mM DTT, and 0.5 mM dNTPs. The same volume of pre-selection sample was used for each of the genes; the entire bead volume was used in the post-selection reactions. The reactions were incubated at 42°C for 90 min with agitation.

Protein Removal

1.6 units of Proteinase K was added to each sample and incubated for 15 minutes at 65°C. Samples were purified with 1.2x Ampure beads and eluted in 30 uL of water.

Gene Identification using qPCR

Three 25 uL qPCR reactions containing 1x Power SYBR® Green PCR Master Mix (Thermo Fisher Scientific, 4367659) and 10 mM of each of the gene specific primers was prepared for each sample and for the no template controls. Three 25 uL reactions were also prepared for each sample without primers as a no primer control. 1 uL of sample was used in each reaction. The qPCR assay was run on a QuantStudio 3 Real-Time PCR System with an initial denaturation of 95 °C for 2 minutes, 30 cycles of 95 °C for 30 seconds, 55 °C for 15 seconds, and 72 °C for 30 seconds, and a final extension of 72 °C for 5 minutes. A melt curve was run to assess the purity of the qPCR products.

Comparison of SMART-display product library and control libraries

SMART-display libraries were prepared as described above up to the puromycin containing linker ligation.

Generation of Puromycin Ligated RNA Library

RNA libraries were annealed to a puromycin containing linker with no biotin (5' /5Phos/CC/ideoxyI/CTC/ideoxyI/ACCCCCCGCCGCCCCCCGTCCT/iSp18/iSp18/iSp18/iSp 18/CC/3Puro/ 3') in a 1:1.25 molar ratio in Annealing Buffer (10x: 100 mM Tris-HCI Buffer, pH 7.5, 500 mM NaCl, 10 mM EDTA). The "no puromycin" control was subject to the same reaction with the omission of the puromycin containing linker. The reactions were incubated at 75 °C for 5 minutes and cooled slowly to 25 °C. Ligation was performed with 0.4 U/uL of T4 RNA Ligase 1, 1 mM ATP, and 1.6 U/uL of SUPERase•In™ RNase Inhibitor for 30 minutes at 25 °C. NEBuffer 4 was added to 1x, and unligated linker was digested with 0.2 U/uL of T5 Exonuclease at 37 °C for 30 minutes. The ligated RNA was purified with an RNeasy Mini Column.

Translation and Display

Protein products were generated using 25 pmols of RNA product and 2 uLs Transcend™ tRNA (Promega, L5061) per 25 uL reaction of the NEB PURExpress IVT kit. 2 uLs of Proteinase K was added to the "protein digested control". Translation reactions were performed in an air incubator for 90 minutes at 37 °C. After translation, KCl and MgCl2 were added to a final concentration of 800 mM and 80 mM respectively. The reaction was incubated at room temperature for 30 minutes and then stored at -20 °C for a minimum of 12 hours.

Protein Selection and Pull-Down

75 uLs of Dynabeads MyOne Streptavidin T1 Beads were prepared per IVT reaction according to the manufacturer’s directions. The IVT reaction was added to the suspended beads and incubated for 1 hour with rotation at room temperature. The beads were washed 3 times with 8M Urea wash buffer (8M Urea, 50 mM Tris, 5 mM EDTA, 0.1% NP40, 500 mM LiCl, 2% SDS), and 3 times with 1x B&W buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1M NaCl).

Library Preparation and Sequencing

The beads were subject to a SuperScript III One-Step RT-PCR (Invitrogen, 12574018) reaction at 5x the original volume of streptavidin beads, with 0.5 uM of each a universal forward primer (5’ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGAGTAAGGAGGATCCAACATG 3’) and an indexed reverse primer (5’ CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTGTCGTCATCGTCTTTGTAGTC 3’, where X represents the index bases). The cycle number was optimized for each sample, using the minimum number of cycles to generate a library. Samples were mixed 3:2 with PhiX and sequenced 150 base pairs from each end on an Illumina MiniSeq.

INLISE

Purification and Immobilization of Display Products

75 uLs of Dynabeads™ MyOne™ Streptavidin T1 (Thermo Fisher Scientific, 65601) were prepared by washing twice in an equivalent volume of 1x PBS pH 7.4 (Thermo Fisher

Scientific, 70011044). The IVT reaction was added to the suspended beads in 1.8 mLs of 1x PBS pH 7.4 (Thermo Fisher Scientific, 70011044) with 0.1% Triton™ X-100 (Sigma-Aldrich, T8787-50ML) and incubated for 1 hour with rotation at room temperature. D-Biotin (Ivitrogen, B20656) was added to 2.25 uM and incubated at room temperature for 10 minutes with rotation. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100 (Sigma-Aldrich, T8787-50ML).

DNA Synthesis

50 uLs of first strand reaction was mixed per sample containing 500 U of SuperScript II Reverse Transcriptase (Thermo Scientific, 18064014), 1x SuperScript II FS Buffer, 5 mM DTT, 1 uM dNTP mix (NEB, N0447S), 1 M Betaine (Sigma-Aldrich, 61962), 6 mM MgCl2, 500 pmol of End Capture TSO (5' /5dSp/AGT AAA GGA GAC CTC AGC TTC ACT GGA rGrGrG 3'), and 40 U of SUPERase• In™ RNase Inhibitor. The mix was added to the beads and incubated at 42°C for 50 minutes with agitation, and then cycled 10 times at 50°C for 2 minutes followed by 42°C for 2 minutes. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100. 100 uLs of first strand reaction was mixed per sample containing 20 U DNA Polymerase I (NEB, M0209S), 1x NEBuffer 2, 2.4 mM DTT, and 0.25 mM dNTP mix. The mix was added to the beads and incubated at 37°C for 30 minutes with agitation. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Restriction Digestion and Control Digestion

All samples were digested with 10 U of BbvCI (NEB, R0601S) in 1x CutSmart Buffer at 500 uLs. The digestion was incubated at 37°C for 1 hour with agitation. After the restriction enzyme digestion, but without washing the beads, the bait population used in the

Proteinase control was generated by the addition of 5 uLs of Proteinase K (NEB, P8107S) to the sample. The sample was incubated for an additional 30 minutes at 37°C with agitation. All samples were then washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Synthesis of Interaction Linker

The top and bottom strands of the interaction linker were reconstituted to 200 uM with Annealing Buffer. The two strands were mixed in a 1:1 molar ratio, incubated at 75 °C for 5 minutes, and cooled slowly to 25 °C.

Interaction Linker Ligation and Release of Prey

Samples with a dI containing puromycin containing linker were ligated to the Interaction Linker and subsequently released from the Dynabeads™ MyOne™ Streptavidin T1 beads to generate the prey population. Ligation was performed at 37°C with agitation for 30 minutes, with 200 pmol Interaction Linker, 4000 U T4 DNA Ligase (NEB, M0202M), and 1x T4 DNA Ligase Buffer in 500 uLs. The interaction linker was omitted in the prey reaction used in the No-linker control. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100. The release of the complexes from the beads was performed at 37°C with agitation for 30 minutes, with 40 U of Endonuclease V (NEB, M0305S) in 50 uLs of 1x NEBuffer™ 3 (NEB, B7003S).

Interaction

The samples without deoxyinosine (dI) bases in the puromycin containing linker were retained on the Dynabeads™ MyOne™ Streptavidin T1 beads to become the bait libraries. These samples were suspended in 150 uLs Binding Buffer (10 mM HEPES (Fisher Scientific, BP299100), 50 mM KCl, 4 mM MgCl2, 2mM DTT, 0.2 mM EDTA, 0.1% Tween® 20

(Sigma-Aldrich, P9416-100ML)). The 50 uL of supernatant from the Endonuclease V digestion (the prey library), was added to the bait samples with the following conventions. PROPER-seq reaction: bait and prey libraries with the full PROPER-seq protocol; No-linker control: bait library with the full PROPER-seq protocol, prey library created without the interaction linker ligated; and Proteinase control: bait library treated with Proteinase K and the prey library created with the full PROPER-seq protocol. The mixtures were incubated at room temperature with rotation for 1 hour. 800 uLs of Binding Buffer was added to each reaction to bring the volume to 1 mL, and they were rotated an additional 10 minutes at room temperature.

Crosslinking and Proximity Ligation

Crosslinking was performed at room temperature for 30 minutes with 0.5 mM BS3 (Thermo Fisher Scientific, A39266). The reaction was quenched with 50 mM Tris-HCI Buffer, pH 7.5 with rotation for 15 minutes. The beads were washed 3 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Proximity ligation was performed with 20,000 U of T4 DNA Ligase in 1 mL of 1x T4 DNA Ligase Buffer. The reaction was incubated with constant rotation for 30 minutes at room temperature. The enzyme was inactivated before the beads were gathered by heating to 65°C for 10 minutes. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Sequencing Library Generation and Sequencing

The DNA was released from the beads with the NEBNext® Ultra™ II FS DNA Module (NEB, E7810S) using twice the reaction volume and a fragmentation time of 5 minutes. The end repair step was not performed. Libraries were then generated with the

NxSeq® UltraLow DNA Library Kit (Lucigen, 15012-1) up to the final AMPure XP Bead purification before amplification. Each sample was eluted in 50 uLs Nuclease-free water and added to 10 uLs of Dynabeads™ MyOne™ Streptavidin T1beads suspended in 50 uLs 1x PBS pH 7.4 with 0.1% Triton X-100. The selection was performed at room temperature for 1 hour. Beads were washed 2 times with 500 uLs Low Salt buffer [0.1% SDS (Invitrogen, AM9820), 0.1% Triton™ X-100, 2 mM EDTA, 20 mM Tris-HCI buffer, pH 8 (Invitrogen, 15568025), 150 mM NaCl], 2 times with 500 uLs 1x B&W buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1M NaCl), and 2 times with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100. Library amplification was then performed with the NxSeq® UltraLow DNA Library Kit as directed.

Each library was paired end sequenced for 100 cycles on each end on an Illumina HiSeq 4000 or NovaSeq 6000.

Validation by proximity ligation assay (PLA)

Cell Culture

HEK 293T cells were cultured in Dulbecco's modified Eagle medium (DMEM; GIBCO, 11960044) supplemented with 10% FBS (Gemini, 100-500), 2 mM Glutamax (GIBCO, 35050061), and 5,000 U/ml penicillin/streptomycin (GIBCO, 15070063), at 37°C with 5 % CO2.

Fixation and Permeabilization

Approximately 0.5 million HEK cells per well were fixed with 4% formaldehyde (Thermo Fisher Scientific, 28906) in PBS pH 7.2 (Life Technologies, 20012027) at room temperature for 30 minutes on a Lab-Tek 8-well Chamber Slide (Thermo Fisher Scientific, 154534). Cells were washed once with PBS pH 7.2, then permeabilized with 200 uLs of 0.1%

Triton X-100 (Sigma-Aldrich, T8787-50ML) in PBS for 15 minutes at room temperature with rocking.

Blocking

Cells were blocked by adding 40 uLs Duolink Blocking Solution (Sigma-Aldrich, DUO92101-1KT) and incubating in a humidity chamber for 1 hour at 37°C.

Staining with Primary Antibody

Primary antibodies were added to the cells at the dilutions listed below in a total of 40 uLs. The slides were incubating in a humidity chamber for 1 hour at 37°C.

Staining with PLA Probes, Ligation, and Amplification

Slides were washed 2x with 70 mL of wash buffer A and stained with PLA probes according to the Duolink Assay instructions. Slides were wash 2x with 70 mL of wash buffer A, and ligation performed according to the Duolink Assay instructions. Slides were wash 2x with 70 mL of wash buffer A, and amplification performed according to the Duolink Assay instructions. Slides were then washed 2x with wash buffer B and 1x with 1:100 wash buffer B.

Imaging

Coverslips were mounted with 12 uLs Duolink PLA mounting medium with DAPI per well and sealed with clear nail polish. Images were acquired on Olympus Inverted Microscope using a 60X/1.518 oil objective (GE Healthcare Life Sciences) (pixel size = 0.1075 μm). A series of z-stack images across the cells were acquired with 0.3 μm sample thickness (3 sections).

Validation by co-IP

Five million HEK293T cells were lysed in RIPA buffer [150 mM NaCl, 5 mM EDTA, 50 mM Tris pH 7.5, 1% NP-40, 0.5% sodium deoxycholate (Sigma-Aldrich, 30970-25G), 0.1% SDS, and a protease inhibitor cocktail (Sigma Aldrich, P8340)] for 30 minutes on ice and subsequently centrifuged at 10,000 xg for 10 minutes. The supernatants were precleared by incubation with Protein-G Dynabeads (Thermo Fisher Scientific, 10003D) for 30 minutes at 4˚C. Antibody-coated beads were prepared by incubating rabbit anti-human Leo1 antibody (5 µg per sample, Bethyl Laboratories, A300-175A) or control rabbit IgG (5 µg per sample; Abcam, AB37415) with pre-washed Protein-G Dynabeads for 2-3 hours at room temperature. 5% of the precleared lysate (input) was saved for later analysis, and the remaining lysate was split equally among the Leo1- or IgG-coated beads for immunoprecipitation (IP). IP was carried out overnight at 4˚C. 10% of the flow through (FT) was retained for analysis. The Dynabeads were washed 3 times for 5 mins each with RIPA buffer. The washed beads were eluted in reducing sample buffer (Thermo Fisher Scientific, 39000) before resolving on an 8% SDS-PAGE and immunoblotting (IB) with indicated antibodies.

Quantification and Statistical Analysis

Processing proper-seq read pairs

The following data processing steps are implemented in the PROPERseqTools pipeline: https://github.com/Zhong-Lab-UCSD/PROPERseqTools. The sequencing reads were subjected to Cutadapt 2.5(Martin 2011) to remove the 3' linker sequence and the 5' adapter sequence. The remaining read pairs were subsequently subjected to Fastp 0.20.0(Huang et al. 2018) to remove low-quality reads (average quality per base < Q20) and short reads (<20 bp). The remaining read pairs were subsequently mapped to RefSeq transcripts (O'Leary et al. 2016) (based on GRCh38.p13, NCBI Homo sapiens Annotation Release 109.20190607) using BWA-MEM 0.7.12-r1039 (Li 2013) with the default parameters. A read was regarded as mapped to a gene if this read was mapped to any of the Refseq transcripts of this gene. The read pairs where the two ends were mapped to two different protein coding genes were identified. Any duplicated chimeric read pairs were subsequently removed to obtain non-duplicate chimeric read pairs.

Test of association between a gene pair and the chimeric read pairs

A Chi-square test was carried out on every gene pair. The null hypothesis is that the mapping of one end of a chimeric read pair to a gene is independent of the mapping of the other end of this chimeric read pair to the other gene. The contingency table of this association test is given in Figure S4A. FDR computed from the Benjamini-Hochberg procedure was used to control for family-wise errors.

Downloading APID data and its subsets

PPIs were downloaded as a MITAB file from the Agile Protein Interactomes DataServer (APID) at http://cicblade.dep.usal.es:8080/APID/init.action. The AP-MS and co-IP derived PPIs were identified by the corresponding labels in the 'Interaction detection method' column of the downloaded MITAB file. The LC-MS derived PPIs were identified by the label of "biochemistry" in the 'Interaction detection method' column and specifying "Publication first author" as "Wan, C. et al. (2015)" (Wan et al. 2015), "Havugimana, PC. et al. (2012)" (Havugimana et al. 2012) and "Kristensen, AR. et al. (2012)" (Kristensen, Gsponer, and Foster 2012).

Quantifying reproducibility by odds ratio

The odds ratio was used to quantify the degree of overlap between two sets of PPIs. The odds ratio (OR) of Table S1.4 is calculated as OR=(A×D)/(C×B), where A, B, C, and D are numbers of PPIs in the corresponding cell in the contingency table.

Comparison to structurally predicted PPIs

The human prePPIs were downloaded from the PrePPI database (https://honiglab.c2b2.columbia.edu/PrePPI/ref/preppi_final600.txt.tar.gz). The Uniprot protein IDs used in PrePPI were converted to gene symbols using the org.Hs.eg.db Bioconductor package in R.

GO term defined subnetworks

The subnetwork associated with a GO term (Ashburner et al. 2000) was retrieved by the PROPER v1.0 nodes that were annotated by this GO term and all the edges connecting these nodes. GO term enrichment analysis was based on hypergeometric tests between the genes annotated by every GO term and the PROPER v1.0 nodes. FDR computed from the Benjamini-Hochberg procedure was used to control for family-wise errors. The entire PROPER v1.0 was plotted with Gephi (0.9.2, https://gephi.org/) (Bastian, Heymann, and Jacomy 2009). All other network figures were plotted with Cytoscape (Shannon et al. 2003).

Test of cell type association

A Chi-square test was applied to every PPI to test the association of this PPI with a cell type. The null hypothesis is that whether a chimeric read pair is mapped to this gene pair is independent of whether this chimeric read pair was generated from this cell type. A PPI was regarded as attributable to a cell type if Chi-square test FDR < 0.05 and odds ratio > 2, where the odds ratio for Table S1.5 is calculated as OR=(A×D)/(C×B).

A GO term defined subnetwork was included in the analysis of cell type association when this GO term contained at least 50 genes (regardless of whether these genes were included in PROPER v1.0) and this GO term defined subnetwork contained at least 10 edges. The association of a subnetwork to a cell type was quantified by the proportions of PPIs (edges) associated with that cell type among all the PPIs of this subnetwork.

Calculating screening completeness, sampling sensitivity, assay sensitivity, precision, and protein interactome size for PROPER v1.0

Screening completeness, sampling sensitivity, assay sensitivity, precision, and protein interactome size were defined by Yu et al. (Yu et al. 2008) and Venkatesan et al. (Venkatesan et al. 2009). We calculated these metrics for PROPER v1.0 based on the methods described by Venkatesan et al. (Venkatesan et al. 2009) and the following positive reference set (PRS), random reference set (RRS) and orthogonal validation sets.

Positive reference set (PRS)

The CORUM database (Giurgiu et al. 2019) contains 2417 human protein complexes, corresponding to 3433 proteins and 39,103 protein pairs. These 39,103 protein pairs are used as our PRS.

Random reference set (RRS)

Following Venkatesan et al. (Venkatesan et al. 2009), RRS was randomly sampled from PROPER-seq's search space outside the PRS to contain the same number of gene pairs as PRS.

Orthogonal validation assay

Targeted co-IP is used as the orthogonal validation assay. The targeted co-IP data were retrieved from APID based on two MI Ontology terms: Anti bait coimmunoprecipitation (MI:0006) and Anti tag coimmunoprecipitation (MI:0007)).

1. 8 Acknowledgements

Chapter 1, in full, is an adapted reprint of materials that appears in Johnson, Kara L.; Qi, Zhijie; Yan, Zhangmin; Wen, Xingzhao; Nguyen, Tri C.; Zaleta-Rivera, Kathia; Chen, Chien-Ju; Fan, Xiaochen; Sriram, Kiran; Wan, Xueyi; Chen, Zhen Bouman; Zhong, Sheng. "Revealing protein-protein interactions at the transcriptome scale by sequencing", Molecular Cell, Volume 81, Issue 19, 2021. The dissertation author was one of the primary investigators and the co-first author of this paper.

1. 9 Reference

Alonso-Lopez, D., F. J. Campos-Laborie, M. A. Gutierrez, L. Lambourne, M. A. Calderwood, M. Vidal, and J. De Las Rivas. 2019. 'APID database: redefining protein-protein interaction experimental evidences and binary interactomes', *Database (Oxford)*, 2019.

Alonso-Lopez, D., M. A. Gutierrez, K. P. Lopes, C. Prieto, R. Santamaria, and J. De Las Rivas. 2016. 'APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks', *Nucleic Acids Res*, 44: W529-35.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium', *Nature genetics*, 25: 25-29.

Aw, J. G. A., Y. Shen, N. Nagarajan, and Y. Wan. 2017. 'Mapping RNA-RNA Interactions Globally Using Biotinylated Psoralen', *J Vis Exp*.

Bader, G. D., D. Betel, and C. W. Hogue. 2003. 'BIND: the Biomolecular Interaction Network Database', *Nucleic Acids Res*, 31: 248-50.

Barabasi, A. L. 2009. 'Scale-free networks: a decade and beyond', *Science*, 325: 412-3.

Barabasi, A. L., and E. Bonabeau. 2003. 'Scale-free networks', *Sci Am*, 288: 60-9.

Barabasi, A. L., and Z. N. Oltvai. 2004. 'Network biology: understanding the cell's functional organization', *Nat Rev Genet*, 5: 101-13.

Barendt, P. A., D. T. Ng, C. N. McQuade, and C. A. Sarkar. 2013. 'Streamlined protocol for mRNA display', *ACS Comb Sci*, 15: 77-81.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." In *Third international AAAI conference on weblogs and social media*.

Benjamini, Yoav, and Yosef Hochberg. 1995. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, 57: 289-300.

Buldyrev, S. V., R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. 2010. 'Catastrophic cascade of failures in interdependent networks', *Nature*, 464: 1025-8.

Chen, T. S., D. Petrey, J. I. Garzon, and B. Honig. 2015. 'Predicting peptide-mediated interactions on a genome-wide scale', *PLoS Comput Biol*, 11: e1004248.

Consortium, Encode Project. 2004. 'The ENCODE (ENCyclopedia Of DNA Elements) Project', *Science*, 306: 636-40.

Consortium, Encode Project. 2012. 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489: 57-74.

Cusick, M. E., H. Yu, A. Smolyar, K. Venkatesan, A. R. Carvunis, N. Simonis, J. F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D. E. Hill, J. R. Ecker, F. P. Roth, and M. Vidal. 2009. 'Literature-curated protein interaction datasets', *Nat Methods*, 6: 39-46.

Deanfield, J. E., J. P. Halcox, and T. J. Rabelink. 2007. 'Endothelial function and dysfunction: testing and clinical relevance', *Circulation*, 115: 1285-95.

Dekker, J., A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. R. Politz, J. Shendure, S. Zhong, and D. Nucleome Network. 2017. 'The 4D nucleome project', *Nature*, 549: 219-26.

Fornerod, M., M. Ohno, M. Yoshida, and I. W. Mattaj. 1997. 'CRM1 is an export receptor for leucine-rich nuclear export signals', *Cell*, 90: 1051-60.

Garzon, J. I., L. Deng, D. Murray, S. Shapira, D. Petrey, and B. Honig. 2016. 'A computational interactome and functional annotation for the human proteome', *Elife*, 5.

Giurgiu, M., J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and A. Ruepp. 2019. 'CORUM: the comprehensive resource of mammalian protein complexes-2019', *Nucleic Acids Res*, 47: D559-D63.

Gu, Liangcai, Chao Li, John Aach, David E. Hill, Marc Vidal, and George M. Church. 2014. 'Multiplex single-molecule interaction profiling of DNA-barcoded proteins', *Nature*, 515: 554-57.

Gullberg, M., S. M. Gustafsdottir, E. Schallmeiner, J. Jarvius, M. Bjarnegard, C. Betsholtz, U. Landegren, and S. Fredriksson. 2004. 'Cytokine detection by antibody-based proximity ligation', *Proc Natl Acad Sci U S A*, 101: 8420-4.

Havugimana, P. C., G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, M. Babu, S. A. Craig, P. Hu, C. Wan, J. Vlasblom, V. U. Dar, A. Bezginov, G. W. Clark, G. C. Wu, S. J. Wodak, E. R. Tillier, A. Paccanaro, E. M. Marcotte, and A. Emili. 2012. 'A census of human soluble protein complexes', *Cell*, 150: 1068-81.

Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. 2004. 'IntAct: an open source molecular interaction database', *Nucleic Acids Res*, 32: D452-5.

Huang, R., M. Han, L. Meng, and X. Chen. 2018. 'Capture and Identification of RNA-binding Proteins by Using Click Chemistry-assisted RNA-interactome Capture (CARIC) Strategy', *J Vis Exp*.

Jäkel, S., and D. Görlich. 1998. 'Importin beta, transportin, RanBP5 and RanBP7 mediate nuclear import of ribosomal proteins in mammalian cells', *Embo j*, 17: 4491-502.

Jerby-Arnon, L., N. Pfetzer, Y. Y. Waldman, L. McGarry, D. James, E. Shanks, B. Seashore-Ludlow, A. Weinstock, T. Geiger, P. A. Clemons, E. Gottlieb, and E. Ruppin. 2014. 'Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality', *Cell*, 158: 1199-209.

Kawalia, A., S. Motameny, S. Wonczak, H. Thiele, L. Nieroda, K. Jabbari, S. Borowski, V. Sinha, W. Gunia, U. Lang, V. Achter, and P. Nurnberg. 2015. 'Leveraging the power of high performance computing for next generation sequencing data analysis: tricks and twists from a high throughput exome workflow', *PLoS One*, 10: e0126321.

Kovacs, I. A., K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D. K. Kim, N. Kishore, T. Hao, M. A. Calderwood, M. Vidal, and A. L. Barabasi. 2019. 'Network-based prediction of protein interactions', *Nat Commun*, 10: 1240.

Kristensen, A. R., J. Gsponer, and L. J. Foster. 2012. 'A high-throughput approach for measuring temporal changes in the interactome', *Nat Methods*, 9: 907-9.

Kukar, T., S. Eckenrode, Y. Gu, W. Lian, M. Megginson, J. X. She, and D. Wu. 2002. 'Protein microarrays to detect protein-protein interactions using red and green fluorescent proteins', *Anal Biochem*, 306: 50-4.

Lee, J. S., A. Das, L. Jerby-Arnon, R. Arafeh, N. Auslander, M. Davidson, L. McGarry, D. James, A. Amzallag, S. G. Park, K. Cheng, W. Robinson, D. Atias, C. Stossel, E. Buzhor, G. Stein, J. J. Waterfall, P. S. Meltzer, T. Golan, S. Hannenhalli, E. Gottlieb, C. H. Benes, Y. Samuels, E. Shanks, and E. Ruppin. 2018. 'Harnessing synthetic lethality to predict the response to cancer treatment', *Nat Commun*, 9: 2546.

Lewis, J. D., J. Wan, R. Ford, Y. Gong, P. Fung, H. Nahal, P. W. Wang, D. Desveaux, and D. S. Guttman. 2012. 'Quantitative Interactor Screening with next-generation Sequencing (QIS-Seq) identifies Arabidopsis thaliana MLO2 as a target of the Pseudomonas syringae type III effector HopZ2', *BMC Genomics*, 13: 8.

Li, Heng. 2013. 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', *arXiv preprint arXiv:1303.3997*.

Li, S., S. W. Tighe, C. M. Nicolet, D. Grove, S. Levy, W. Farmerie, A. Viale, C. Wright, P. A. Schweitzer, Y. Gao, D. Kim, J. Boland, B. Hicks, R. Kim, S. Chhangawala, N. Jafari, N. Raghavachari, J. Gandara, N. Garcia-Reyero, C. Hendrickson, D. Roberson, J. Rosenfeld, T. Smith, J. G. Underwood, M. Wang, P. Zumbo, D. A. Baldwin, G. S. Grills, and C. E. Mason. 2014. 'Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study', *Nat Biotechnol*, 32: 915-25.

Li, X., B. Zhou, L. Chen, L. T. Gou, H. Li, and X. D. Fu. 2017. 'GRID-seq reveals the global RNA-chromatin interactome', *Nat Biotechnol*, 35: 940-50.

Licata, L., L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni. 2012. 'MINT, the molecular interaction database: 2012 update', *Nucleic Acids Res*, 40: D857-61.

Lievens, S., N. Vanderroost, J. Van der Heyden, V. Gesellchen, M. Vidal, and J. Tavernier. 2009. 'Array MAPPIT: high-throughput interactome analysis in mammalian cells', *J Proteome Res*, 8: 877-86.

Lu, Z., J. Gong, and Q. C. Zhang. 2018. 'PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution', *Methods Mol Biol*, 1649: 59-84.

Luck, K., D. K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, D. Choi, A. G. Cote, M. Daley, S.

Deimling, A. Desbuleux, A. Dricot, M. Gebbia, M. F. Hardy, N. Kishore, J. J. Knapp, I. A. Kovacs, I. Lemmens, M. W. Mee, J. C. Mellor, C. Pollis, C. Pons, A. D. Richardson, S. Schlabach, B. Teeking, A. Yadav, M. Babor, D. Balcha, O. Basha, C. Bowman-Colin, S. F. Chin, S. G. Choi, C. Colabella, G. Coppin, C. D'Amata, D. De Ridder, S. De Rouck, M. Duran-Frigola, H. Ennajdaoui, F. Goebels, L. Goehring, A. Gopal, G. Haddad, E. Hatchi, M. Helmy, Y. Jacob, Y. Kassa, S. Landini, R. Li, N. van Lieshout, A. MacWilliams, D. Markey, J. N. Paulson, S. Rangarajan, J. Rasla, A. Rayhan, T. Rolland, A. San-Miguel, Y. Shen, D. Sheykhkarimli, G. M. Sheynkman, E. Simonovsky, M. Tasan, A. Tejeda, V. Tropepe, J. C. Twizere, Y. Wang, R. J. Weatheritt, J. Weile, Y. Xia, X. Yang, E. Yeger-Lotem, Q. Zhong, P. Aloy, G. D. Bader, J. De Las Rivas, S. Gaudet, T. Hao, J. Rak, J. Tavernier, D. E. Hill, M. Vidal, F. P. Roth, and M. A. Calderwood. 2020. 'A reference map of the human binary protein interactome', *Nature*, 580: 402-08.

Martin, Marcel. 2011. 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *2011*, 17: 3.

McGregor, Lynn M., Tara Jain, and David R. Liu. 2014. 'Identification of Ligand–Target Pairs from Combined Libraries of Small Molecules and Unpurified Protein Targets in Cell Lysates', 136: 3264-70.

Mellacheruvu, D., Z. Wright, A. L. Couzens, J. P. Lambert, N. A. St-Denis, T. Li, Y. V. Miteva, S. Hauri, M. E. Sardiu, T. Y. Low, V. A. Halim, R. D. Bagshaw, N. C. Hubner, A. Al-Hakim, A. Bouchard, D. Faubert, D. Fermin, W. H. Dunham, M. Goudreault, Z. Y. Lin, B. G. Badillo, T. Pawson, D. Durocher, B. Coulombe, R. Aebersold, G. Superti-Furga, J. Colinge, A. J. Heck, H. Choi, M. Gstaiger, S. Mohammed, I. M. Cristea, K. L. Bennett, M. P. Washburn, B. Raught, R. M. Ewing, A. C. Gingras, and A. I. Nesvizhskii. 2013. 'The CRAPome: a contaminant repository for affinity purification-mass spectrometry data', *Nat Methods*, 10: 730-6.

Messner, S., D. Schuermann, M. Altmeyer, I. Kassner, D. Schmidt, P. Schär, S. Müller, and M. O. Hottiger. 2009. 'Sumoylation of poly(ADP-ribose) polymerase 1 inhibits its acetylation and restrains transcriptional coactivator function', *Faseb j*, 23: 3978-89.

Navlakha, S., X. He, C. Faloutsos, and Z. Bar-Joseph. 2014. 'Topological properties of robust biological and computational networks', *J R Soc Interface*, 11: 20140283.

Nguyen, T. C., X. Cao, P. Yu, S. Xiao, J. Lu, F. H. Biase, B. Sridhar, N. Huang, K. Zhang, and S. Zhong. 2016. 'Mapping RNA-RNA interactome and RNA structure in vivo by MARIO', *Nat Commun*, 7: 12023.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W.

Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. 2016. 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Res*, 44: D733-45.

Peri, S., J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey. 2003. 'Development of human protein reference database as an initial platform for approaching systems biology in humans', *Genome Res*, 13: 2363-71.

Petalidis, L., S. Bhattacharyya, G. A. Morris, V. P. Collins, T. C. Freeman, and P. A. Lyons. 2003. 'Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis', *Nucleic Acids Res*, 31: e142.

Roberts, R. W., and J. W. Szostak. 1997. 'RNA-peptide fusions for the in vitro selection of peptides and proteins', *Proc Natl Acad Sci U S A*, 94: 12297-302.

Rolland, T., M. Tasan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A. R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J. C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A. L. Barabasi, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal. 2014. 'A proteome-scale map of the human interactome network', *Cell*, 159: 1212-26.

Rual, J. F., K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. 2005. 'Towards a proteome-scale map of the human protein-protein interaction network', *Nature*, 437: 1173-8.

Saito, T., and M. Rehmsmeier. 2015. 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PLoS One*, 10: e0118432.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Res*, 13: 2498-504.

Sharma, E., T. Sterne-Weiler, D. O'Hanlon, and B. J. Blencowe. 2016. 'Global Mapping of Human RNA-RNA Interactions', *Mol Cell*, 62: 618-26.

Soderberg, O., M. Gullberg, M. Jarvius, K. Ridderstrale, K. J. Leuchowius, J. Jarvius, K. Wester, P. Hydbring, F. Bahram, L. G. Larsson, and U. Landegren. 2006. 'Direct observation of individual endogenous protein complexes in situ by proximity ligation', *Nat Methods*, 3: 995-1000.

Sridhar, B., M. Rivas-Astroza, T. C. Nguyen, W. Chen, Z. Yan, X. Cao, L. Hebert, and S. Zhong. 2017. 'Systematic Mapping of RNA-Chromatin Interactions In Vivo', *Curr Biol*, 27: 610-12.

Stark, C., B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. 2006. 'BioGRID: a general repository for interaction datasets', *Nucleic Acids Res*, 34: D535-9.

Su, Zhenqiang, Paweł P. Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Charles Wang, Gary P. Schroth, Robert A. Setterquist, John F. Thompson, Wendell D. Jones, Wenzhong Xiao, Weihong Xu, Roderick V. Jensen, Reagan Kelly, Joshua Xu, Ana Conesa, Cesare Furlanello, Hanlin Gao, Huixiao Hong, Nadereh Jafari, Stan Letovsky, Yang Liao, Fei Lu, Edward J. Oakeley, Zhiyu Peng, Craig A. Praul, Javier Santoyo-Lopez, Andreas Scherer, Tieliu Shi, Gordon K. Smyth, Frank Staedtler, Peter Sykacek, Xin-Xing Tan, E. Aubrey Thompson, Jo Vandesompele, May D. Wang, Jian Wang, Russell D. Wolfinger, Jiri Zavadil, Scott S. Auerbach, Wenjun Bao, Hans Binder, Thomas Blomquist, Murray H. Brilliant, Pierre R. Bushel, Weimin Cai, Jennifer G. Catalano, Ching-Wei Chang, Tao Chen, Geng Chen, Rong Chen, Marco Chierici, Tzu-Ming Chu, Djork-Arné Clevert, Youping Deng, Adnan Derti, Viswanath Devanarayan, Zirui Dong, Joaquin Dopazo, Tingting Du, Hong Fang, Yongxiang Fang, Mario Fasold, Anita Fernandez, Matthias Fischer, Pedro Furió-Tari, James C. Fuscoe, Florian Caimet, Stan Gaj, Jorge Gandara, Huan Gao, Weigong Ge, Yoichi Gondo, Binsheng Gong, Meihua Gong, Zhuolin Gong, Bridgett Green, Chao Guo, Lei Guo, Li-Wu Guo, James Hadfield, Jan Hellemans, Sepp Hochreiter, Meiwen Jia, Min Jian, Charles D. Johnson, Suzanne Kay, Jos Kleinjans, Samir Lababidi, Shawn Levy, Quan-Zhen Li, Li Li, Li Li, Peng Li, Yan Li, Haiqing

Li, Jianying Li, Shiyong Li, Simon M. Lin, Francisco J. López, Xin Lu, Heng Luo, Xiwen Ma, Joseph Meehan, Dalila B. Megherbi, Nan Mei, Bing Mu, Baitang Ning, Akhilesh Pandey, Javier Pérez-Florido, Roger G. Perkins, Ryan Peters, John H. Phan, Mehdi Pirooznia, Feng Qian, Tao Qing, Lucille Rainbow, Philippe Rocca-Serra, Laure Sambourg, Susanna-Assunta Sansone, Scott Schwartz, Ruchir Shah, Jie Shen, Todd M. Smith, Oliver Stegle, Nancy Stralis-Pavese, Elia Stupka, Yutaka Suzuki, Lee T. Szkotnicki, Matthew Tinning, Bimeng Tu, Joost van Delft, Alicia Vela-Boza, Elisa Venturini, Stephen J. Walker, Liqing Wan, Wei Wang, Jinhui Wang, Jun Wang, Eric D. Wieben, James C. Willey, Po-Yen Wu, Jiekun Xuan, Yong Yang, Zhan Ye, Ye Yin, Ying Yu, Yate-Ching Yuan, John Zhang, Ke K. Zhang, Wenqian Zhang, Wenwei Zhang, Yanyan Zhang, Chen Zhao, Yuanting Zheng, Yiming Zhou, Paul Zumbo, Weida Tong, David P. Kreil, Christopher E. Mason, Leming Shi, and Seqc Maqc-Iii Consortium. 2014. 'A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium', *Nature Biotechnology*, 32: 903-14.

Touchette, M. H., E. R. Van Vlack, L. Bai, J. Kim, A. B. Cognetta, 3rd, M. L. Previti, K. M. Backus, D. W. Martin, B. F. Cravatt, and J. C. Seeliger. 2017. 'A Screen for Protein-Protein Interactions in Live Mycobacteria Reveals a Functional Link between the Virulence-Associated Lipid Transporter LprG and the Mycolyltransferase Antigen 85A', *ACS Infect Dis*, 3: 336-48.

Trigg, Shelly A., Renee M. Garza, Andrew Macwilliams, Joseph R. Nery, Anna Bartlett, Rosa Castanon, Adeline Goubil, Joseph Feeney, Ronan O'Malley, Shao-Shan C. Huang, Zhuzhu Z. Zhang, Mary Galli, and Joseph R. Ecker. 2017. 'CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping', *Nature Methods*, 14: 819-25.

Venkatesan, K., J. F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A. S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A. L. Barabasi, and M. Vidal. 2009. 'An empirical framework for binary interactome mapping', *Nat Methods*, 6: 83-90.

Vermeulen, M., N. C. Hubner, and M. Mann. 2008. 'High confidence determination of specific protein-protein interactions using quantitative mass spectrometry', *Curr Opin Biotechnol*, 19: 331-7.

Walhout, A. J., and M. Vidal. 2001. 'High-throughput yeast two-hybrid assays for large-scale protein interaction mapping', *Methods*, 24: 297-306.

Wan, C., B. Borgeson, S. Phanse, F. Tu, K. Drew, G. Clark, X. Xiong, O. Kagan, J. Kwan, A. Bezginov, K. Chessman, S. Pal, G. Cromar, O. Papoulas, Z. Ni, D. R. Boutz, S.

Stoilova, P. C. Havugimana, X. Guo, R. H. Malty, M. Sarov, J. Greenblatt, M. Babu, W. B. Derry, E. R. Tillier, J. B. Wallingford, J. Parkinson, E. M. Marcotte, and A. Emili. 2015. 'Panorama of ancient metazoan macromolecular complexes', *Nature*, 525: 339-44.

Xenarios, I., D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. 2000. 'DIP: the database of interacting proteins', *Nucleic Acids Res*, 28: 289-91.

Yachie, N., E. Petsalaki, J. C. Mellor, J. Weile, Y. Jacob, M. Verby, S. B. Ozturk, S. Li, A. G. Cote, R. Mosca, J. J. Knapp, M. Ko, A. Yu, M. Gebbia, N. Sahni, S. Yi, T. Tyagi, D. Sheykhkarimli, J. F. Roth, C. Wong, L. Musa, J. Snider, Y. C. Liu, H. Yu, P. Braun, I. Stagljar, T. Hao, M. A. Calderwood, L. Pelletier, P. Aloy, D. E. Hill, M. Vidal, and F. P. Roth. 2016. 'Pooled-matrix protein interaction screens using Barcode Fusion Genetics', 12: 863-63.

Yan, Z., N. Huang, W. Wu, W. Chen, Y. Jiang, J. Chen, X. Huang, X. Wen, J. Xu, Q. Jin, K. Zhang, Z. Chen, S. Chien, and S. Zhong. 2019. 'Genome-wide colocalization of RNA-DNA interactions and fusion RNA pairs', *Proc Natl Acad Sci U S A*, 116: 3328-37.

Yang, F., Y. Lei, M. Zhou, Q. Yao, Y. Han, X. Wu, W. Zhong, C. Zhu, W. Xu, R. Tao, X. Chen, D. Lin, K. Rahman, R. Tyagi, Z. Habib, S. Xiao, D. Wang, Y. Yu, H. Chen, Z. Fu, and G. Cao. 2018. 'Development and application of a recombination-based library versus library high- throughput yeast two-hybrid (RLL-Y2H) screening system', *Nucleic Acids Res*, 46: e17.

Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. 2008. 'High-quality binary protein interaction map of the yeast interactome network', *Science*, 322: 104-10.

Yu, M., W. Yang, T. Ni, Z. Tang, T. Nakadai, J. Zhu, and R. G. Roeder. 2015. 'RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II', *Science*, 350: 1383-6.

Zhang, Q. C., D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, and B. Honig. 2012. 'Structure-based prediction of protein-protein interactions on a genome-wide scale', *Nature*, 490: 556-60.

Zhang, Q. C., D. Petrey, J. I. Garzon, L. Deng, and B. Honig. 2013. 'PrePPI: a structure-informed database of protein-protein interactions', *Nucleic Acids Res*, 41: D828-33.

Zhang, Y., W. L. Ku, S. Liu, K. Cui, W. Jin, Q. Tang, W. Lu, B. Ni, and K. Zhao. 2017. 'Genome-wide identification of histone H2A and histone variant H2A.Z-interacting proteins by bPPI-seq', *Cell Res*, 27: 1258-74.

Zhu, Y. Y., E. M. Machleder, A. Chenchik, R. Li, and P. D. Siebert. 2001. 'Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction', *Biotechniques*, 30: 892-7.

CHAPTER 2 Meta-analysis on human PPI networks

2.1 Abstract

People developed divergent technologies to detect protein-protein interactions on large scale, thus leading the structures of the mapped protein-protein interaction networks to vary from each other. In this study, we utilized multiple human PPI datasets derived from different techniques (PROPER-seq, AP-MS, co-IP, LC-MS, Y2H, etc.) and of different confidence levels to systematically examine architectural characteristics of human PPI networks, including the clustering coefficient distribution, the node degree distribution, the hub protein properties, and the maximal-clique properties. By utilizing the L3 link prediction algorithm, we proposed that a comprehensive human PPI network should be a scale-free network filled with many completed or close-to-completed cliques. We found that in human PPI networks, the hub proteins are often involved in large protein complexes. They are highly inter-connected with each other and serve as basic building blocks of small biological meaningful network motifs. We also found that the proteins in the same maximal cliques of a PPI network are more likely to share similar molecular functions and thus more biologically related.

2.2 Introduction

Protein-protein interactions (PPIs) are essential to almost every biological process and cellular function in humans (Peng et al. 2017). As a modeled representation of the numerous interacted proteins in the human cells, PPI networks serve informative roles in helping us understand multiple aspects of human proteome including putative roles of uncharacterized proteins, the relationships between proteins within the same multi-molecular complexes, as well as probing unknown disease mechanisms, etc. (Jordan, Nguyen, and Liu 2012) Our current understanding of the human protein interactome is incomplete and noisy. Various human PPI datasets exist but with little overlap with each other (Johnson et al. 2021). The divergence across the human PPI datasets may partly emerge from the different PPI screening techniques. For example, HuRI was derived from yeast two-hybrid (Y2H) assay that detects binary interactions (Luck et al. 2020), Bioplex3.0 was derived from AP-MS that detects co-complex interactions (Huttlin et al. 2021), and PROPER v.1.0 was derived from PROPER-seq that detects both binary and co-complex interactions (Johnson et al. 2021). The divergence also emerges from the difference in scale and confidence level when constructing the human PPI network. For example, Lit-NB-13 and Lit-BM both contain around 10,000 multiple-evidence PPIs curated from different literatures (Rolland et al. 2014) while PROPER v.1.0 and HuRI contain over 50,000 PPIs but derived from one single technique. In other words, a high confidence PPI network is always constructed at the cost of being less comprehensive to cover the entire human PPI interactome.

As a result, many of the human PPI network's general characteristics still need further investigation and validation. For example, we observed diverse distribution patterns of clustering coefficients of the protein nodes across different PPI networks (Johnson et al. 2021), which suggests different architectures exist for different PPI networks regarding how proteins are embedded locally. This intrigued us to explore what pattern may better describe the human PPI network in nature. Besides, although the human protein interactome is proposed to be a 'scale free' network (Nacher, Hayashida, and Akutsu 2009), we noticed that for some networks, the relationship between log scaled node frequency and log scaled node degree is not completely linear (Johnson et al. 2021). This raised the question of whether power law is always the best fit for human PPI networks.

Hub proteins are the highly connected central nodes in a PPI network (Fox et al. 2011). They connect different functional modules across the network and make the entire network functionally more robust (He and Zhang 2006). Previously literature had divergent observations and discussions about how hub proteins exist in a PPI network. Some studies found that the inter-connectivity between hub proteins is lower as compared to the non-hub proteins (Vandereyken et al. 2018) while some other studies proposed that hub proteins are together in a larger complex so that they are highly inter-connected (Batada, Hurst, and Tyers 2006). Some studies proposed a bimodality distribution regarding the hub proteins' co-expression level with their interaction partners which further classifies hub proteins into party hubs and date hubs (Fraser 2005) while some other studies observed a continuous distribution of the co-expression levels (Agarwal et al. 2010). Besides, lots of these related studies and analyses were conducted based on non-human PPI networks (Vandereyken et al. 2018; Fraser 2005; Chang et al. 2013). So, whether any of these aforementioned observations or conclusions still remain robust when applied to human PPI networks is questionable.

Another important component of PPI networks is clique. The cliques, which are the complete subgraphs in a PPI network, are often considered to be tightly associated with protein complexes and functional modules (Wang et al. 2010), leading these cliques to serve as important references for understanding disease related mechanisms (Yang, Zhao, and Tang 2014). Many PPI prediction algorithms involve completing the cliques of a network that was derived from biological experiments (Yu et al. 2006). Some of them also adopt GO ontology to help filter the false positive predictions (Yang and Tang 2014). These metrics are all based on the empirical observation that PPIs from cliques are usually biologically related so that they have common terms in GO annotations of biological process (BP), cellular component (CC), or molecular function (MF) (Ashburner et al. 2000). However, the number and the scale of the PPI networks on which such observation was based are limited. Whether the cliques within the more updated and comprehensive human PPI networks nowadays still inherit the same properties remains in doubt.

In this study, we collected 11 human PPI datasets of different techniques and different confidence levels (Table S2.1). In detail, we selected PPI datasets named as AP-MS, co-IP, and LC-MS from Agile Protein Interactome Dataserver (APID) (Alonso-Lopez et al. 2019) where the PPIs were derived from co-complex PPI detection technologies like affinity purification mass spectrometry (AP-MS) (Morris et al. 2014), co-immunoprecipitation (co-IP) (Free, Hazelwood, and Sibley 2009), and liquid chromatography-mass spectrometry (LC-MS) (Pitt 2009). We acquired binary PPI datasets including HuRI (Luck et al. 2020), HI-II-14 (Rolland et al. 2014), H-I-05 (Rual et al. 2005), and APID-binary (Alonso-Lopez et al. 2019). We acquired PROPER v.1.0 from PROPER-seq which detects both binary and co-complex PPIs (Johnson et al. 2021). We also included literature-curated multiple-evidence PPI datasets like Lit-BM (Luck et al. 2020), Lit-NB-13 (Rolland et al. 2014), and CORUM (Giurgiu et al. 2019) into our study. We performed systematic analysis on the PPI networks from these datasets to infer the bona fide architecture and characteristics of human PPI networks.

2.3 Results

2.3.1 Clustering coefficient distribution

The clustering coefficient distribution of the nodes is a major descriptive statistic of networks as it quantifies the degree to which the nodes in a network tend to cluster together (Watts and Strogatz 1998). While in general, the clustering coefficient exhibits a reverse correlation to the degree, the exact distribution patterns vary across different human PPI networks (Figure S2.1). One pattern occurs for PROPER v.1.0 and CORUM PPI networks where there is a long plateau region at the higher clustering coefficient regions and then extended with a negative correlation tail at the higher node degree regions (Figure S2.1A, I). Such a pattern suggests that the corresponding PPI network is a dense network filled with many completed or close-to-completed cliques. Another pattern occurs for co-complex (AP-MS) and binary (HuRI) PPI networks where there is no obvious plateau at the high clustering coefficient region but instead, there is a negative correlation slope between clustering coefficients and node degrees (Figure S2.1B, E). Such a pattern suggests that the corresponding PPI network is with much fewer completed cliques. There are some groups of nodes that are densely connected but there are fewer connections between the groups. The two patterns are not separated by PPI detection methods, as they are exhibited in both binary and co-complex PPI networks.

We developed a two-dimensional metric to quantitatively describe the clustering coefficient distribution patterns. In one dimension, we computed the Pearson correlation coefficient (PCC) (Schober, Boer, and Schwarte 2018) between a PPI network's log-transformed clustering coefficients and log-transformed node degrees by Equation 1. This dimension evaluates the extent of the negative correlation between clustering coefficients and node degrees of the network. A larger PCC value means the corresponding PPI network adopts a less hierarchical architecture.

$$PCC(C,D) = \frac{cov\left(log_{10}(C), log_{10}(D)\right)}{\sigma\left(log_{10}(C)\right) * \sigma\left(log_{10}(D)\right)} \quad (Eq1)$$

Where:

- C: clustering coefficients of the nodes
- D: degrees of the nodes
- cov: covariance
- σ: standard deviation

In the other dimension, we computed the clustering coefficient vs. degree distribution (CCDG) score by Equation 2. For each node, we first computed the product of its clustering coefficient and its square of log-transformed degree ($C * \left(log_{10}(D)\right)^2$). We log-transformed the node degree to restrict the score from being too large. We did not log transform the clustering coefficient so as to keep the score positive. We squared the log-transformed degree to reward more credit to those highly connected but still closely clustered nodes. We then summed up the products of all the nodes and normalized it by the square root of the number of proteins (nodes) and by the square root of the number of PPIs (edges) of the network. We applied normalization to the number of nodes in the network because we summed the computed products for each node. The number of edges in the network should also be normalized against because the node degrees will be inflated by the number of edges. Since we only summed up the nodes once, we took the square root of both the number of edges and the number of nodes to keep the normalization balanced. This dimension evaluates the expansion of the distribution's plateau from low node degrees to high node degrees. A larger CCDG score indicates the network has more nodes clustered together to form completed or close-to-completed cliques.

$$CCDG\ score = \frac{\sum_n \left(C * \left(log_{10}(D)\right)^2\right)}{\sqrt{N} * \sqrt{P}} \quad (Eq2)$$

Where:

- n: index of the nodes
- C: clustering coefficient of node n
- D: degree of node n
- N: total number of PPIs (edges)
- P: total number of proteins (nodes)

We applied this two-dimensional metric to the human PPI networks and to a random network. We found that the PPI networks with different clustering coefficient distribution patterns are well separated from each other, suggesting the validity of this metric modeling the clustering coefficient distribution patterns (Figure 2.1A).

We hypothesized that the clustering coefficient distribution of the bona fide human PPI network should have a large PCC value and a high CCDG score. Some PPI networks having either small PCC values or low CCDG scores are due to the limitation of the PPI mapping technology used. The authors of the L3 algorithm used this hypothesis as an assumption (Kovacs et al. 2019). Based on this assumption, the L3 algorithm predicts the chance of two proteins to form a PPI by the connectivity of the respective immediate neighbors of each protein. Hereafter, we will leverage the L3 algorithm to test our hypothesis. We applied the L3 link prediction algorithm to various human PPI networks. We selected high-confidence predicted PPIs by the requirements of either these predicted PPIs having an L3-prediction score of 99th percentile or higher, or the number of predicted PPIs with the highest L3-prediction score being greater than 1/3 of its original size. The second requirement ensures the minimum size increase of those small PPI networks like H-I-05 after L3 predictions. We then combined these high-confidence predicted PPIs with the original PPI network to get the L3-predicted PPI network. The L3-predicted networks exhibited 1.38 to 4.96-fold more edges than their original networks (Table S2.2). For all these L3-predicted PPI networks, their clustering coefficient distribution against node degrees exhibits larger PCC values and higher CCDG scores than their origins (Figure 2.1B, Figure S2.2). This is consistent with the assumption used by the L3 algorithm.

We then plotted the precision-recall (PR) curves (Saito and Rehmsmeier 2015) of all the L3-predicted PPI networks by varying the L3 prediction score percentile of selecting predicted PPIs. We used CORUM as the reference set for PROPER v.1.0 and for co-complex PPI networks and used HuRI as the reference set for binary PPI networks. To extend the PR curves into the range of the greater recall values, we simulated the higher recall part of the curves by fitting a reciprocal function ($y = \frac{a}{x+k}$) to the empirical curves (Figure S2.3). This is because the range of the L3-predicted PPI dataset's recall values is always limited by the fact that the L3-predicted PPI networks will always have a larger size than their origins. We also applied the L3 prediction to randomly permutated networks to generate background PR curves. We found that the PR curve of the L3-predicted network overrides the PR value of the original network and outperforms the background for all the human PPI datasets tested (Figure S2.3). This strongly supports our hypothesis that the clustering coefficient distribution of human PPI networks should exhibit large PCC values and higher CCDG scores by nature. The human PPI network, regardless of binary or co-complex, is a dense network filled with completed and close-to-complete cliques.

2.3.2 Degree distribution

Many biological networks exhibit a scale-free property with most of the nodes having a low degree of interactions while a few nodes having a high connectivity with the others (Przytycka and Yu 2004). However, by examining the degree distributions of various human PPI networks, we found most of the human PPI networks are not perfectly scale-free (Figure S2.4). Some of the curves (AP-MS, Figure S2.4B) bend down when the node degree is high, making the log-scaled relationship between node frequency and node degree not linear. Since an exponential function may also yield similar distribution patterns (Ciavolella et al. 1991), we asked whether the power-law function is the better fit to human PPI networks' degree distribution than the exponential function. We performed linear regression on both log-scale transformed degree distributions (power-law function) and semi-log transformed degree distributions (exponential function) of different human PPI networks. We computed their resulting mean square errors (MSE) to evaluate which function is the better fit. We found that for all the human PPI networks tested, the MSEs of the exponential fit are larger than the MSEs of the power-law fit (Figure 2.1C, Figure S2.5). This suggests that human PPI networks are closer to scale-free than to other architectural properties.

We then asked whether the L3-predicted PPI networks also inherit the scale-free property. We applied linear regression to the log-transformed node degree distributions of all the L3-predicted PPI networks and computed the resulting MSEs. We found that compared with the original networks, the MSEs of the L3-predicted networks are lower for all the PPI networks (Figure 2.1D). This further implies that when human PPI networks are getting more comprehensive, regardless of binary or co-complex, they are also getting closer to the ideal scale-free architecture.

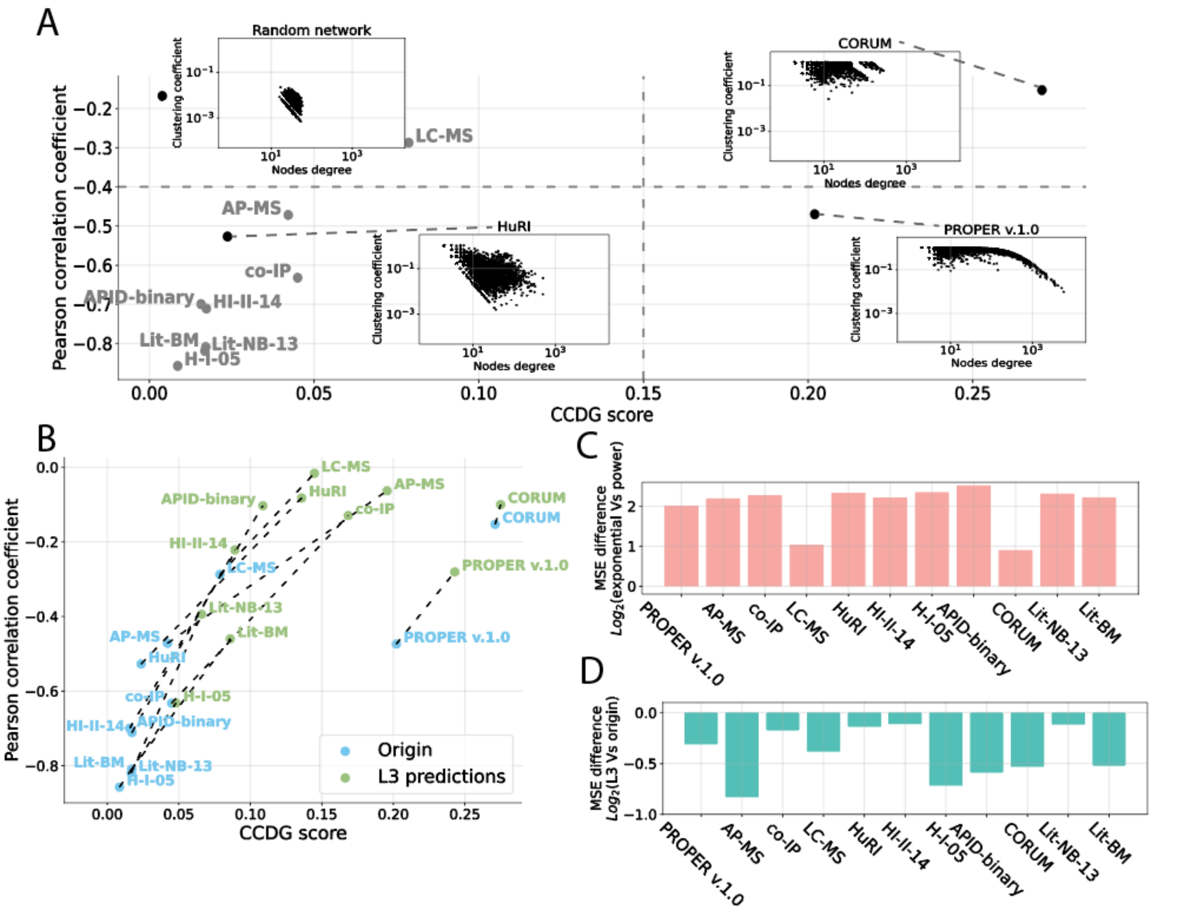


Figure 2.1: Clustering coefficient and degree distribution of PPI networks

(A) The two-dimensional metric to describe the pattern of clustering coefficient against nodes degree of PPI networks.

(B) Comparison of PPI networks before (origin, blue) and after L3 predictions (green) regarding the two-dimensional metric.

(C) The log fold change of mean square errors (MSE) derived from exponential fit to PPI networks over MSE derived from power-law fit to PPI networks.

(D) The log fold change of mean square errors (MSE) derived from power-law fit to L3-predicted PPI networks over MSE derived from power-law fit to original PPI networks.

2.3.3 Hub proteins

Interconnectivity of hub proteins

Hub proteins are the highly-connected nodes in a PPI network that are essential to the biological functions and pathways in the network (He and Zhang 2006). In this study, hub proteins of a PPI network are identified as the top 10% of proteins with the highest number of node degrees in that network. We identified the hub proteins for various human PPI networks as well as for their L3-predicted networks. We found that most of the hub proteins remain the same between original and L3-predicted networks (Figure 2.2A), supporting the functional necessity of these hub proteins in PPI networks.

As previous literatures have controversial observations regarding the interconnectivity of hub proteins (Vandereyken et al. 2018; Batada, Hurst, and Tyers 2006), we here explored the distribution of shortest path length of between-hub proteins and of between-non-hub proteins in different human PPI networks to evaluate the interconnectivity of hub proteins. We found for all the networks tested, the shortest path length of between-hub proteins is significantly shorter than that of between-non-hub proteins ($p\text{-value}<0.05$, student's t-test, Figure S2.6). Hub proteins of PROPER v.1.0 are the most connected, with all the shortest path lengths shorter than a length of 3. Hub proteins of co-complex and binary PPI networks are also highly connected with each other, with those of co-complex PPI networks being less connected than those of binary PPI networks ($p\text{-value}<0.05$, student's t-test). We repeated the analysis with the L3-predicted networks and the observations above stay the same for all the L3-predicted PPI networks (Figure S2.7).

Applying L3 to PPI networks decreased the average shortest path length of both between-hub and between-non-hub proteins. We proceeded to ask whether L3 attenuates or enhances the shortest path length difference between between-hub and between-non-hub proteins. We used t-statistics values of between-non-hub vs. between-hub proteins from the student's t-test to measure the shortest path length difference. We found that L3-predicted networks always have the larger t-statistics than their original networks ($p\text{-value}=7.26e\text{-}4$, paired t-test, Figure 2.2B), meaning that the hub proteins become relatively more deeply connected with each other than the non-hub proteins after applying the L3 algorithm. In other words, despite L3 shortening the shortest path length of the entire network, this shortening effect is more pronounced between two hubs than between two non-hubs. Having established that L3-predicted networks have higher precision and recall than the original networks, these observations serve as a piece of evidence to support that hub proteins are often highly connected with each other in large subgraphs in both binary and co-complex human PPI networks.

Distribution modality of hub proteins

Previous studies proposed that the co-expression level of hub proteins and their interaction partners in PPI networks present a bimodality distribution (Fraser 2005). Based on the bimodality distribution, hub proteins that have weak co-expressions are considered as date hubs that interact transiently with other proteins. Hub proteins with strong co-expressions are considered as party hubs that are involved in large functional protein complexes and interact with other proteins simultaneously and continuously (Fraser 2005). We asked whether such co-expression bimodality distribution exists in human PPI networks. We used mutual rank scores derived from the integrated human gene co-expression data from COXPRESdb (Obayashi et al. 2019) to quantify gene co-expression levels in this analysis. We found that for all the human PPI networks tested except LC-MS, the average mutual rank score of hub proteins with their interaction partners expresses a unimodal distribution ($p\text{-value}>0.05$, Hartigan's dip test for unimodality, Figure S2.8) while the distribution of LC-MS exhibits a bi-modal distribution according to the SkinnyDip algorithm (Maurus and Plant 2016). The average mutual rank score of hub protein interactions is significantly lower than that of all the interactions in the network for all the human PPI networks ($p\text{-value}<0.05$, student's t-test), indicating a stronger co-expression between hub proteins and their interacted partners than the background. We repeated the analysis with the L3-predicted networks and found that the observations above stayed the same for all the L3-predicted PPI networks (Figure S2.9).

We then compared the distribution of the mutual rank score of hub proteins before and after applying L3 predictions. We found for all the networks except for LC-MS and CORUM, the average mutual rank score of hub proteins of the L3-predicted network is significantly smaller than that of the original network (p-value<0.05, student's t-test, Figure 2.2C). This general decrease in the mutual rank score, in other words, an increase in the co-expression level, of hub proteins after applying L3 predictions further suggests that in a more comprehensive human PPI network regardless of detection methods, hub proteins are often contained in large protein complexes that interact with many other proteins simultaneously.

Hub proteins in network motifs

To further understand the role of hub proteins, we asked if hub proteins are enriched in the network motifs of PPI networks. We identified small network motifs of size-3, size-4, and size-5 for all the PPI networks (Figure S2.10). We found the average percentages of hub proteins in these network motifs are all higher than 40% (Figure 2.3A). Given that hub proteins only consist of 10% of the proteins in each PPI network, hub proteins are thus enriched in the small PPI network motifs. Besides, the average percentage of hub proteins in the network motif increases as the size of the network motif increases. Applying L3 predictions to the PPI network also significantly increases the average percentage of hub proteins in the size-3 and size-4 motifs (p-value<0.05, student's t-test, Figure 2.3A), further emphasizing hub proteins' significant enrichment in human PPI network motifs.

For all the PPI networks, we found over 50% of the size-4 motif occurrences contain size-3 motif occurrences and over 50% of the size-5 motif occurrences contain either size-3 or size-4 motif occurrences. This percentage further increases after applying L3 predictions to the PPI networks (Figure S2.11A). Out of the motifs that contain smaller motifs, over 95% involve at least 1 hub protein for all the PPI networks. These observations suggest that the small network motifs with hub proteins may serve as the basic building blocks in PPI networks that can further combine with each other to form larger complexes that perform biological functions. For example, we noticed HDAC1, HDAC2, and SKP1, three hub proteins in PROPER v.1.0, form two size-3 network motifs with VRK1, a non-hub protein. They then recruit CDY1 and further merge to form a size-5 network motif that is related to histone modification function (Figure 2.3B). In another example in Lit-BM, U2AF1, U2AF2, and SRPK2 form a hub-protein size-3 motif and then grow into a size-5 motif by incorporating two other hub proteins to perform RNA splicing function (Figure 2.3C).

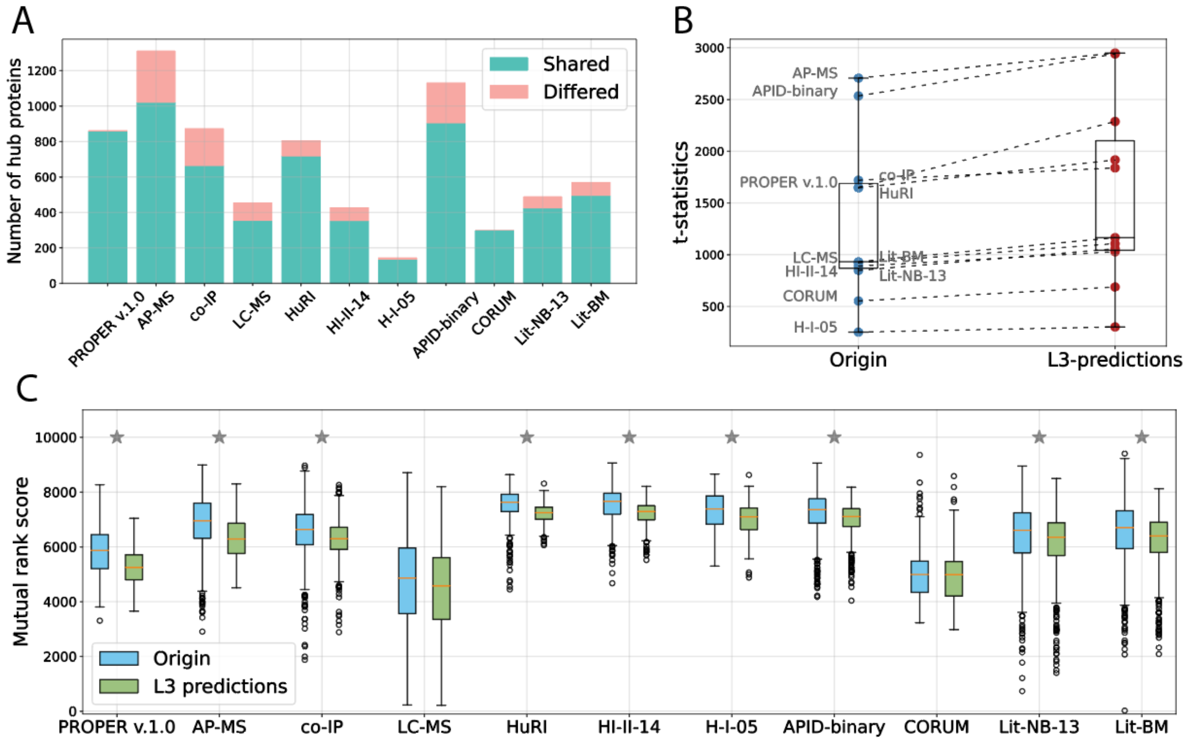


Figure 2.2: Hub proteins in PPI networks

(A) Number of hub proteins shared (green) and differed (pink) in PPI networks before and after applying L3 predictions.

(B) Comparison of t-statistics derived from the student's t-test comparing the shortest path length of between-non-hub proteins against that of between-hub proteins of PPI networks before (origin, blue) and after L3-predictions (L3-predictions, red). A larger t-statistic suggests a larger shortest path length difference for between-non-hub proteins and between-hub proteins.

(C) Comparison of co-expression values of hub proteins in PPI networks before (Origin, blue) and after L3-predictions (green).

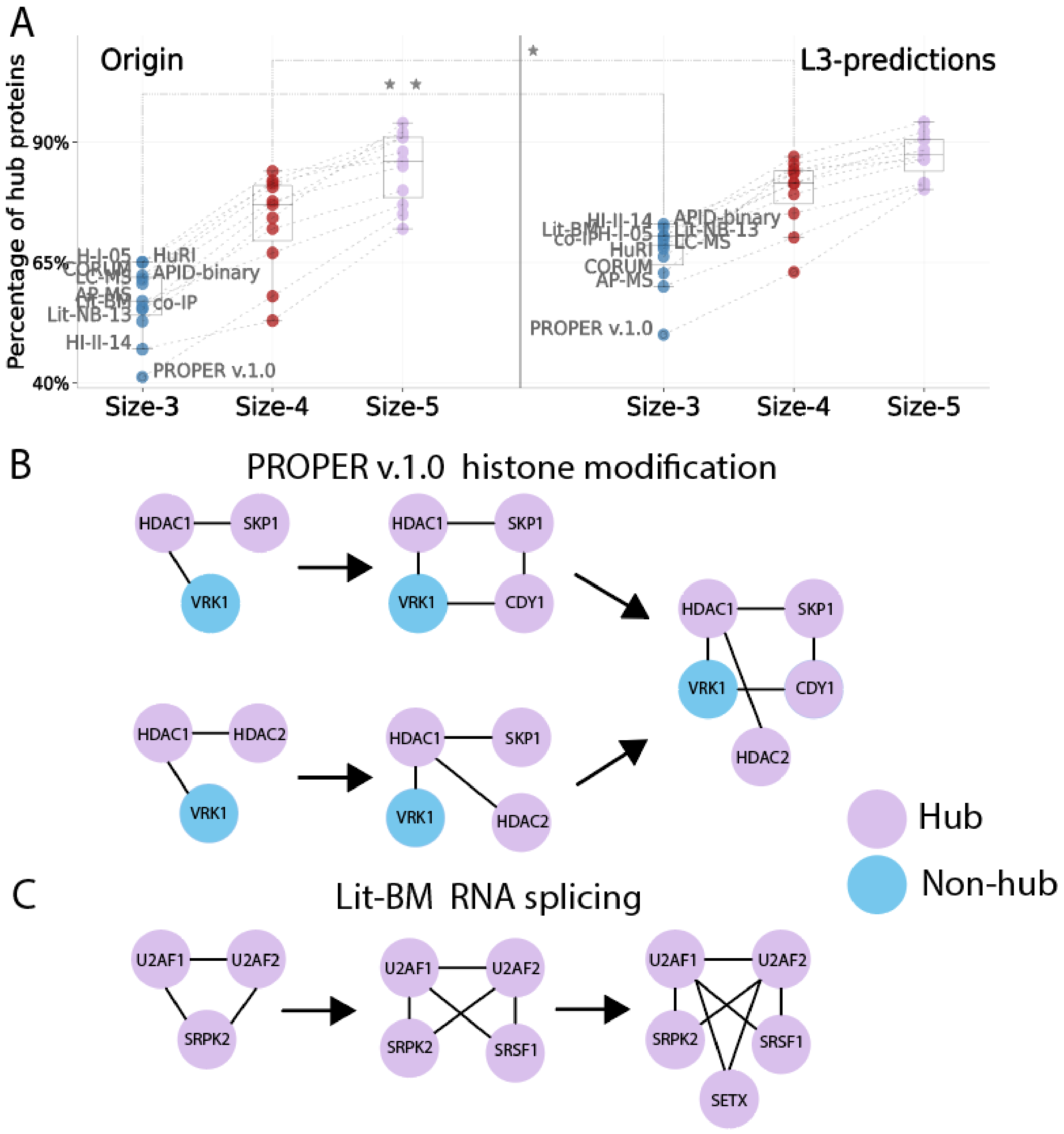


Figure 2.3: Hub proteins in the network motifs of PPI networks

(A) Percentage of hub proteins in size-3 (blue), size-4 (red), and size-5 (purple) network motifs in PPI networks before and after L3 predictions.

(B) Size-3, size-4, and size-5 network motifs in PROPER v.1.0 that are related to histone modification function.

(C) Size-3, size-4, and size-5 network motifs in Lit-BM that are related to RNA splicing function.

2.3.4 Maximal cliques

Hub proteins in M-cliques

The maximal clique for a node is the largest complete subgraph of the node within a network (Darehmiraki 2009). We identified the maximal cliques (M-cliques) for each node in various human PPI networks as well as in their L3-predicted networks. L3-predicted networks shared a similar number of M-cliques with their original networks (p-value=0.98, student's t-test, Table S2.2), while the average size of the M-cliques in all the L3-predicted networks is significantly larger than that in the original networks (p-value<0.05, student's t-test, Figure S2.12, S2.13). We observed that the number of proteins (nodes) in the human PPI networks is just slightly larger than the number of M-cliques in the same network, meaning that only a minor portion of proteins in the network share the same M-cliques (Table S2.2). We asked if these M-clique-sharing proteins are more likely to be hub proteins or not. We found that for all the human PPI networks and their L3-predicted networks tested, the average percentage of hub proteins in M-clique-sharing proteins is significantly larger than 10% (p-value<0.05, student's t-test, Figure 2.4A). As 10% is the percentage of hub proteins within all the proteins in the PPI networks in this study, hub proteins are thus more likely to be within the same M-cliques than non-hub proteins in human PPI networks. PROPER v.1.0 has the highest percentage of M-clique-sharing proteins being hub proteins both before and after L3 predictions. The M-clique-sharing proteins in co-complex PPI networks and binary PPI networks share a similar chance of being hub proteins both before and after L3 predictions (p-value=0.36 for original networks, p-value=0.78 for L3-predicted networks, student's t-test).

It is expected that in general, the size of M-cliques with hub proteins will be larger than that of M-cliques without hub proteins, and applying L3-prediction will increase the size of almost all M-cliques. We asked whether such size increase from applying L3-prediction is more effective on M-cliques with hub or without hub protein. We measured the size increase of the M-clique of a certain protein before and after L3 predictions by computing the fold change of the between pair-wise nodes number of proteins in the corresponding M-clique. We found that the size of M-cliques with hub proteins increases significantly more than that of M-cliques without hub proteins after L3 predictions for all the human PPI networks tested (p-value<0.05, student's t-test, Figure 2.4B). This suggests that M-cliques with hub proteins are more connected to each other, and the interconnectivity of hub proteins is relatively high in human PPI networks.

Biological association between proteins in M-cliques

We asked in human PPI networks, whether the proteins in the same M-clique are more biologically related. We measured the biological similarity between a pair of proteins based on the semantic similarity of their GO terms. For each pair of proteins, we applied GOGO (Zhao and Wang 2018) to achieve three similarity scores with each corresponding to Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) GO terms. We evaluated the biological closeness of the proteins within an M-clique by computing the average BP, CC, and MF similarity scores of all the protein pairs within that clique. We also computed the background average BP, CC, and MF similarity scores by generating random cliques of the same number and size as the M-cliques in the original network using the PPIs from that network. We found that for all the PPI networks and their L3-predicted networks, their average BP, CC, and MF similarity score of M-cliques is significantly higher than that of the background (p-value<0.05, student's t-test, Figure S2.14, S2.15), with MF having the highest similarity scores on average. This suggests that proteins within the same M-cliques are more likely to serve similar biological functions.

For most of the human PPI networks and their L3-predicted networks, we observed a weak positive correlation between the similarity score of M-cliques and the size of M-cliques (Figure S2.16, S2.17). We conjured that large M-cliques may have a higher similarity score than small M-cliques on average as large M-cliques are less likely formed by false positives in the PPI network. An M-clique is identified as a large M-clique if its size is either among the top 10% of the largest M-cliques in the network or larger than 5 proteins. The rest of the non-large M-cliques are classified as small M-cliques. We found that large M-cliques have a higher average BP, CC, and MF similarity score than small M-cliques for all the PPI networks tested, with MF score being the most separated between large and small M-cliques. The observations above also hold the same for all the L3-predicted PPI networks (Figure 2.4C).

Applying the L3-prediction algorithm will add edges to the original PPI networks, which may cause some of the M-cliques to merge into a large M-clique. We asked if M-cliques with higher biological similarities between each other are more likely to merge after L3 predictions. In this analysis, we only focused on M-cliques with a minimum size of 3 proteins. We first identified all the merged M-cliques in the L3-predicted networks and their source M-cliques (cliques that are to be merged after L3-predictions) in the original networks (Table S2.3). Then we measured the biological similarity across the source M-cliques by computing the average GO similarity scores of their merged version with regard to BP, CC, and MF. We also computed the background distribution of similarity scores by randomly merging the same number and the same size of source M-cliques. We found that for all the human PPI networks, the merged cliques have a higher average BP, CC, and MF similarity score than the background (p-value<0.05, student's t-test, Figure S2.18). The MF similarity scores are mostly separated from the background (Figure 2.4D). These results suggest that proteins in the M-cliques of human PPI networks are mostly associated in the aspect of molecular functions. In other words, proteins with similar molecular functions are more likely to form interaction pairs and to cluster together to serve as a functional module in PPI networks.

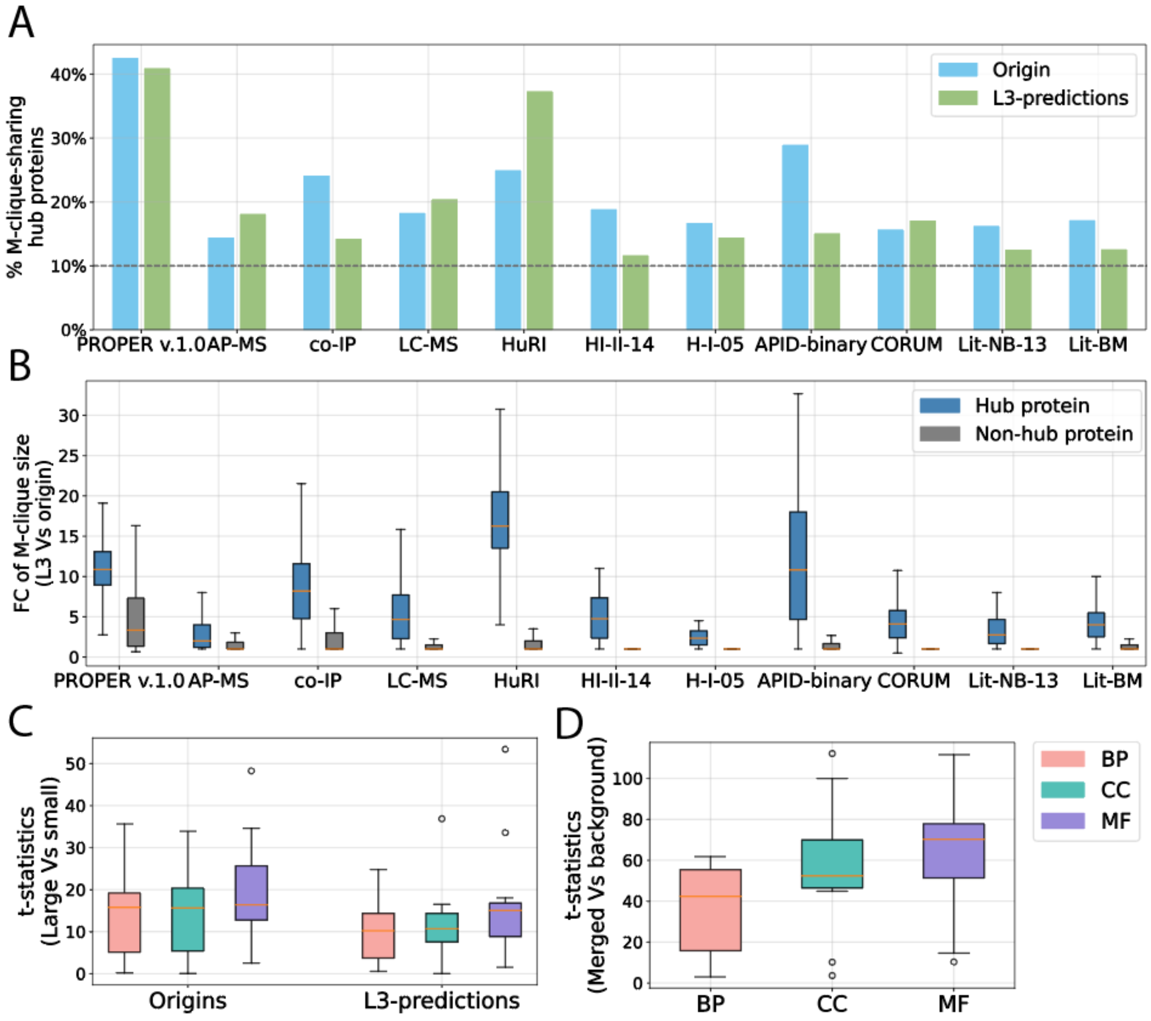


Figure 2.4: Maximal cliques in PPI networks

(A) Percentage of hub proteins that share the same maximal cliques (M-cliques) in PPI networks before (origin, blue) and after L3 predictions (green). The grey dotted line indicates the percentage of proteins that are defined as hub proteins in PPI networks in this study.

(B) Comparison of the fold change of M-clique size (number of proteins) between hub proteins (blue) and non-hub proteins (grey) in PPI networks. The fold change was the M-clique size of each node in the L3-predicted PPI networks over that in the original PPI networks.

(C) Comparison of t-statistics derived from the student's t-test comparing the BP (pink), CC (green), and MF (purple) similarity scores of large M-cliques over that of small M-cliques in origin and L3-predicted PPI.

(D) Comparison of t-statistics derived from the student's t-test comparing the BP (pink), CC (green), and MF (purple) similarity scores of the merged M-cliques after L3 predictions over that of the background in PPI networks.

2.4 Discussion

By analyzing the characteristics of various human PPI networks, we made observations in consensus to suggest that a comprehensive human PPI network, regardless of detection methods, should be a scale-free network filled with many completed or close-to-completed cliques. The hub proteins of the human PPI network are highly inter-connected with each other and meanwhile, are centered with other non-hub proteins of similar molecular functions. By repeating the analysis with the L3-predicted networks, which exhibited higher precision and recall values than the original networks, we found more evidence to further support the conclusions above. The human PPI network features inducted from this study may serve as a useful reference for researchers to develop PPI mapping techniques and to understand human proteome in the future.

Within the analysis scope of this study, co-complex PPIs and binary PPI networks share similar network characteristics. On average, co-complex PPI networks possess more cliques while the hub proteins of binary PPI networks are more inter-connected and co-expressed. But none of the network features analyzed can perfectly separate the individual co-complex PPI datasets from the binary PPI datasets. This study does not take into account the difference in search space or in cell lines that may exist in detecting PPIs through different experiments. We anticipate that future work is required to consider these variations with statistical rigor to investigate the difference between co-complex and binary PPI networks. We anticipate the analysis to be extended into specific functional modules of the networks so that certain disease mechanisms may be well explained by either co-complex or binary PPIs.

2.5 Supplementary information

Supplementary figures

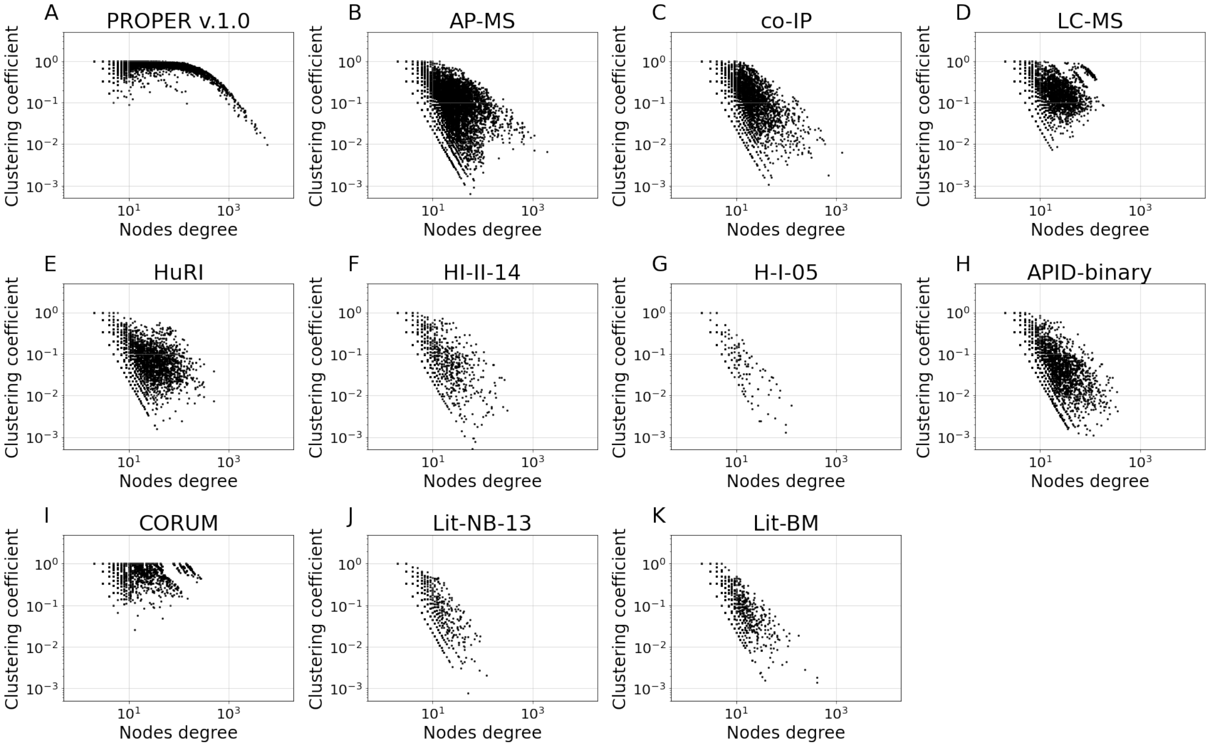


Figure S2.1: Clustering coefficient distribution of human PPI networks
(A)-(K) Scattering plots of log-scaled clustering coefficient as a function of log-scaled node degrees for each node in the PPI networks.

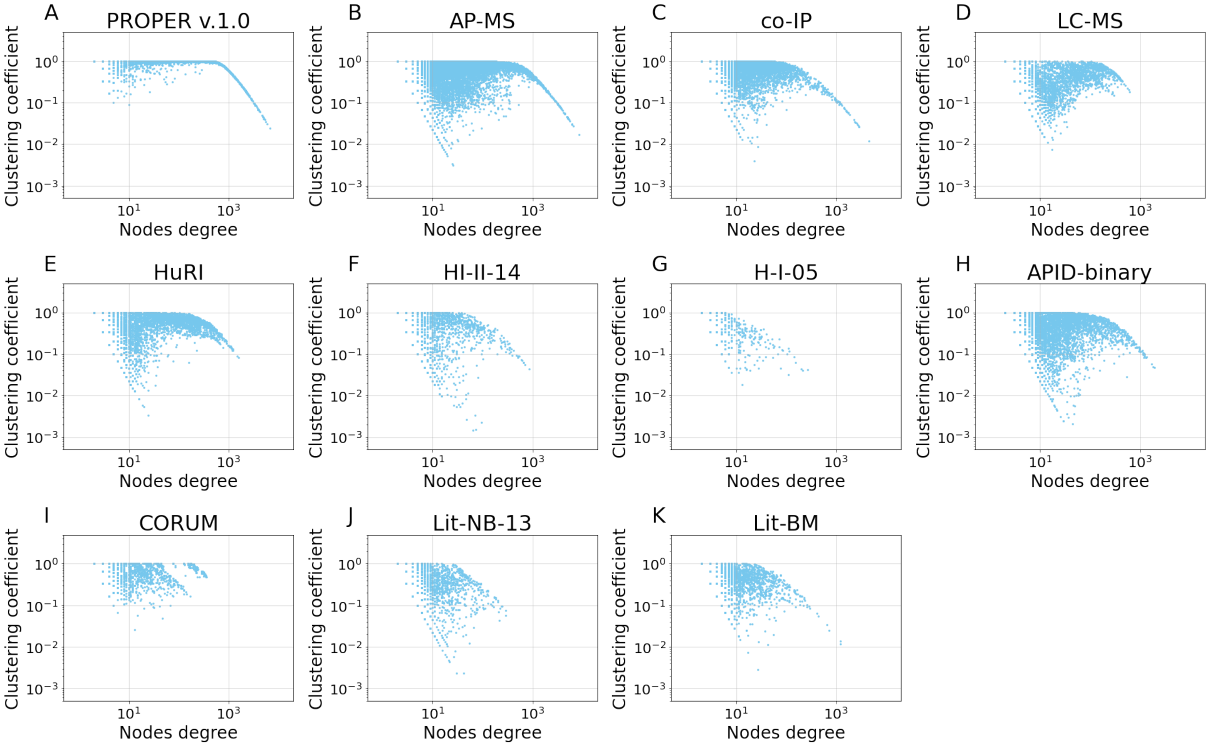


Figure S2.2: Clustering coefficient distribution of L3-predicted PPI networks
(A)-(K) Scattering plots of log-scaled clustering coefficient as a function of log-scaled node degrees for each node in the PPI networks after applying L3 predictions.

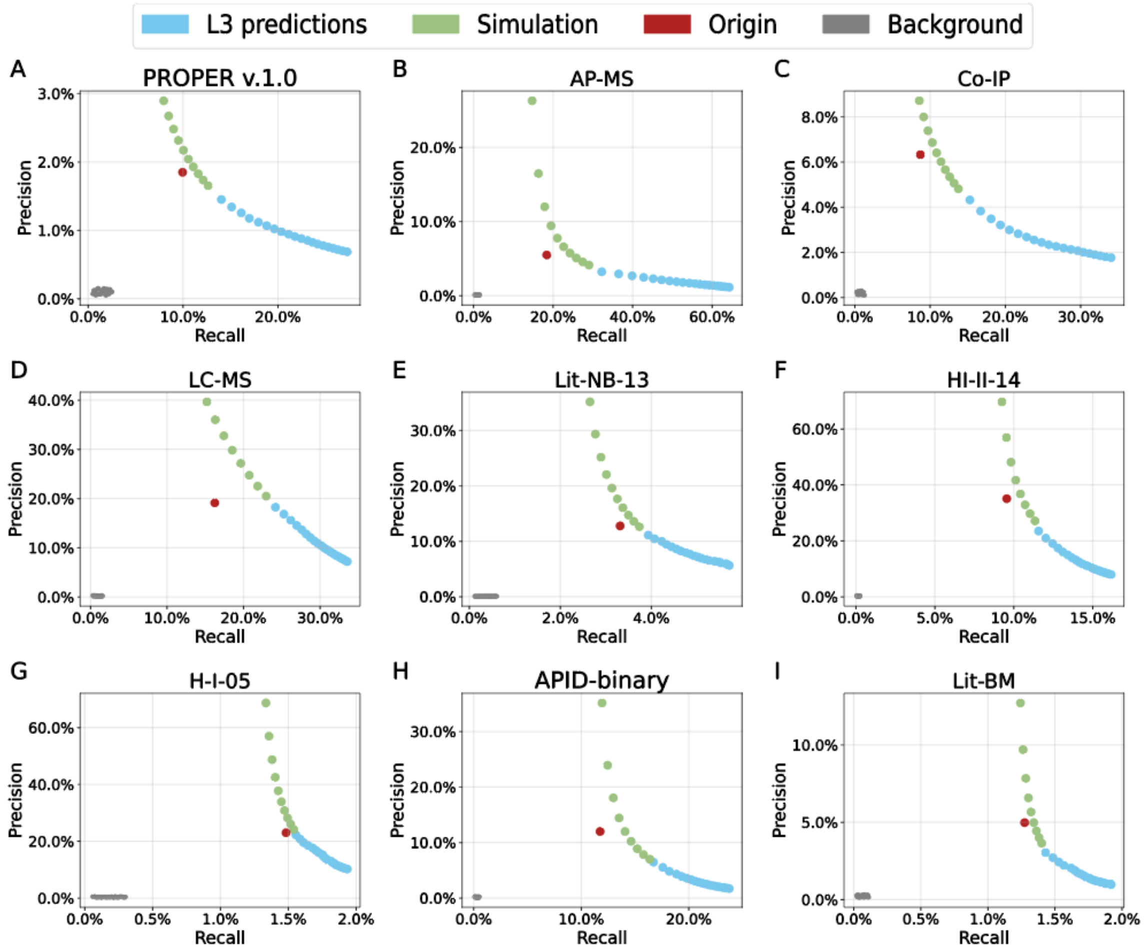


Figure S2.3: Precision-recall curves of L3-predicted PPI networks

(A)-(I) Precision-recall curves of L3-predicted PPI networks (blue and green) compared with the precision-recall value of the corresponding original PPI networks (red). The blue dots are real precision-recall values computed from L3-predicted PPI networks by varying the percentile of selecting L3-predicted PPIs. The green dots are simulated precision-recall values of L3-predicted PPI networks by fitting a reciprocal function to the blue dots. The grey dots are background curves by applying L3 predictions to the random network formed by permutating gene pairs from the PPIs of the original network.

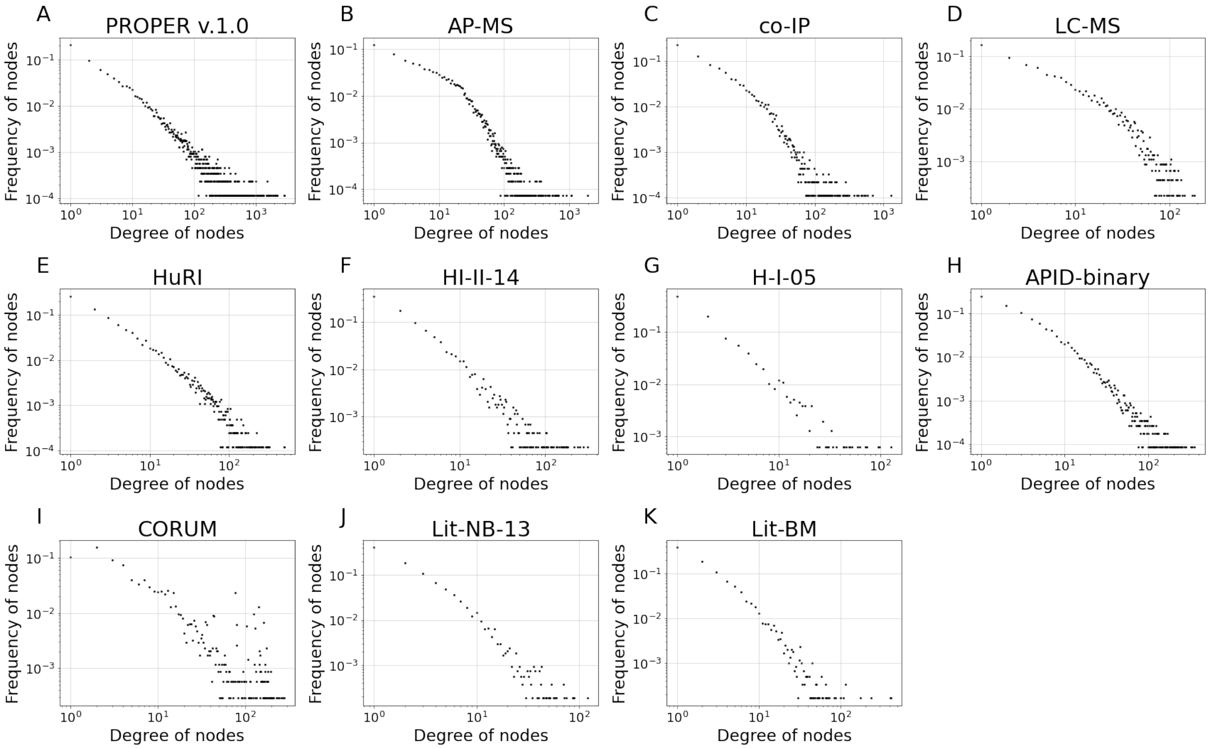


Figure S2.4: Node degree distribution of human PPI networks

(A)-(K) Scattering plots of log-scaled frequency as a function of log-scaled degree for each node in the PPI networks.

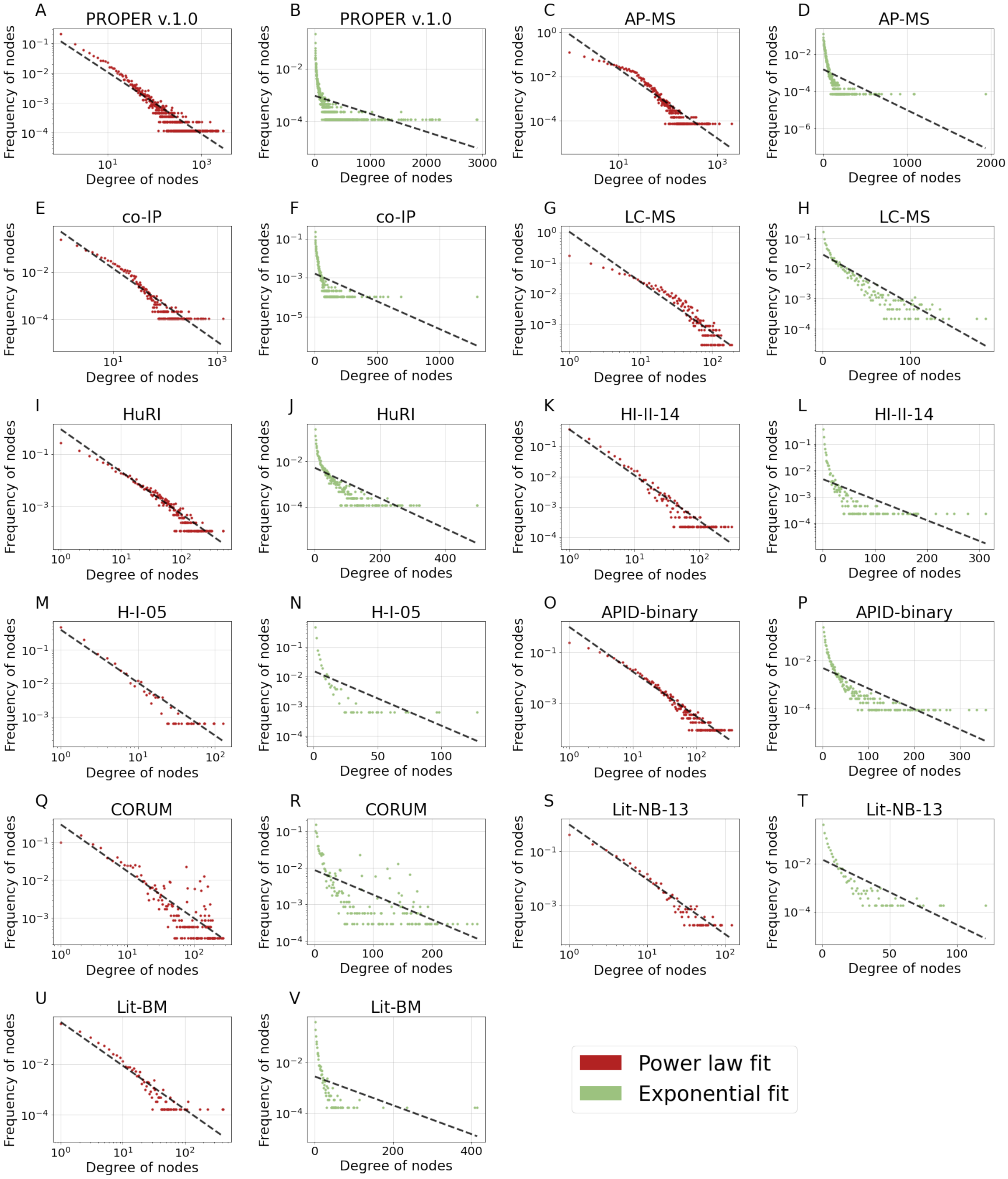


Figure S2.5: Linear regression on node degree distributions of PPI networks
(A)-(V) Linear regressions on log-scaled degree distributions (red, power-law fit) and on semi-log-scaled node degree distributions (green, exponential fit) of PPI networks.

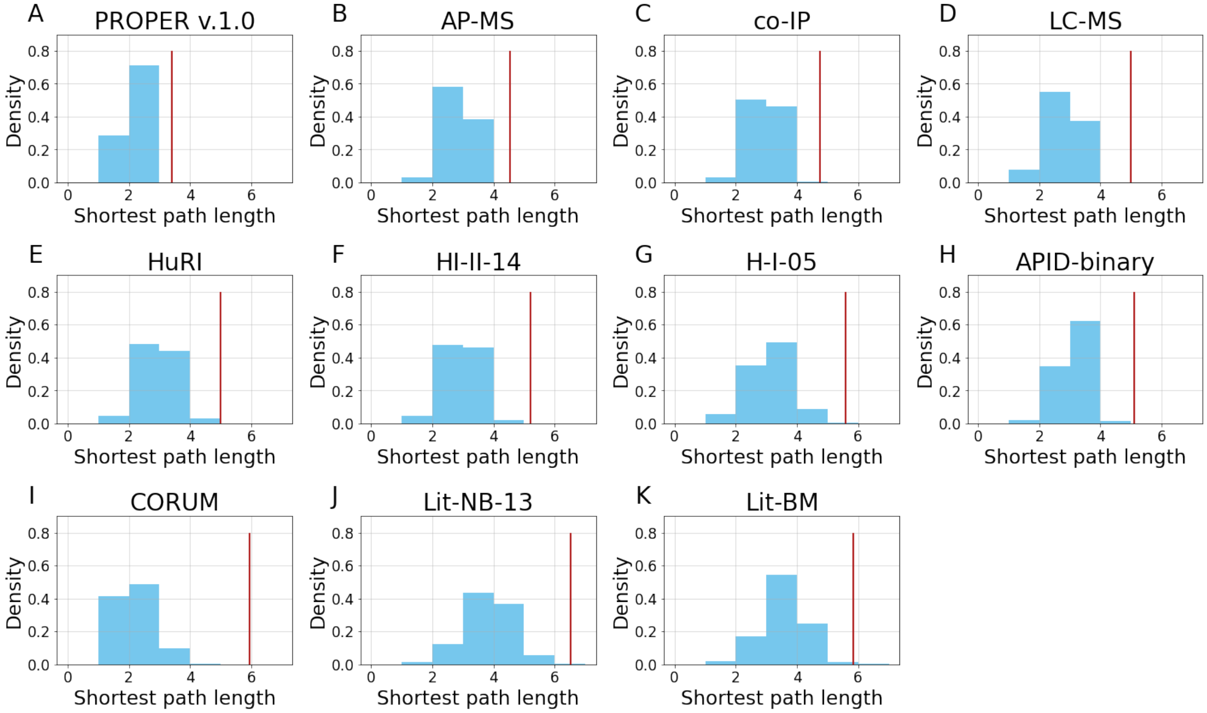


Figure S2.6: Shortest path length distribution of PPI networks

(A)-(K) Distributions of the shortest path length of between-hub proteins of PPI networks. The red line marks the average shortest path length of between-non-hub proteins of the network.

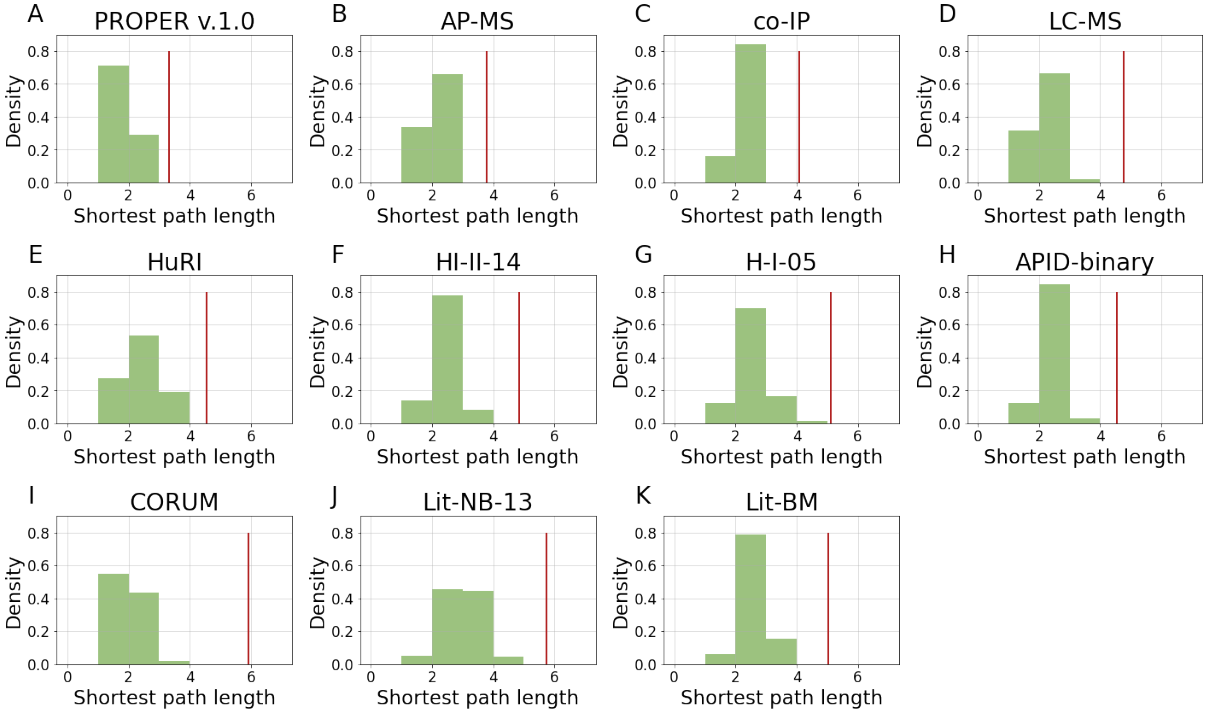


Figure S2.7: Shortest path length distribution of L3-predicted PPI networks
(A)-(K) Distributions of the shortest path length of between-hub proteins of PPI networks after applying L3 predictions. The red line marks the average shortest path length of between-non-hub proteins of the network after applying L3 predictions.

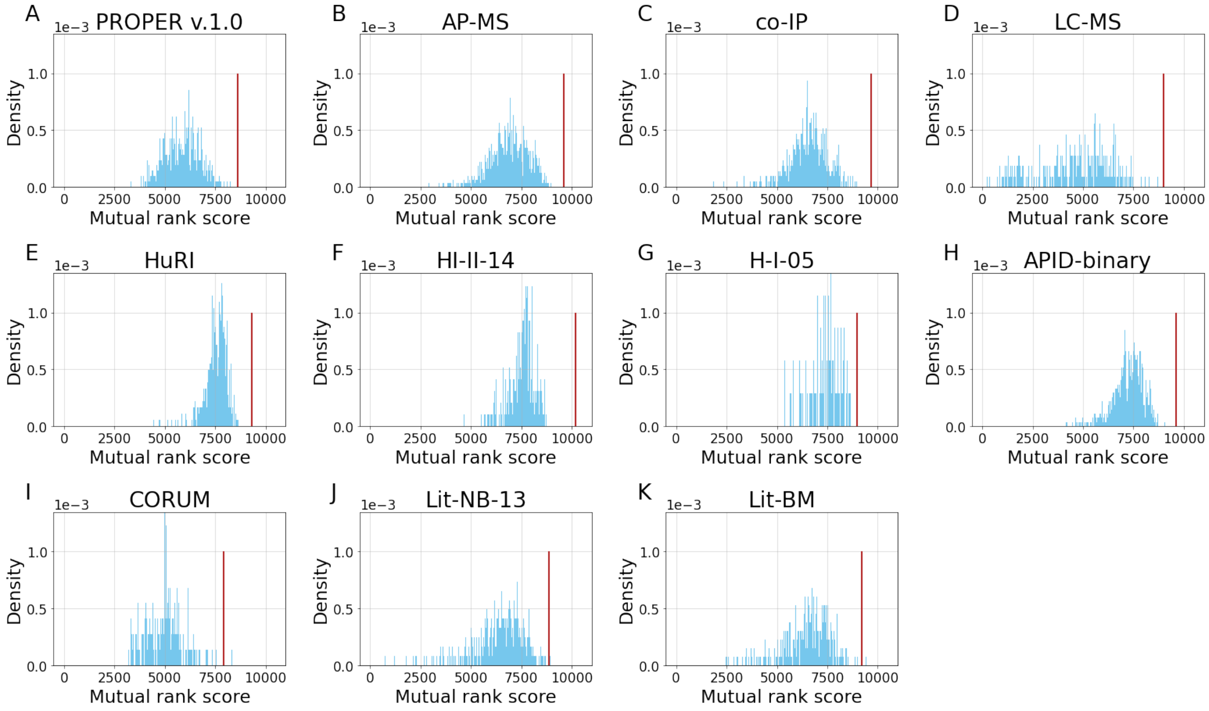


Figure S2.8: Co-expression level distribution of PPI networks
(A)-(K) Distributions of hub proteins' mutual rank score with their interaction partners in PPI networks. The red line marks the average mutual rank score of all the PPIs in the original network.

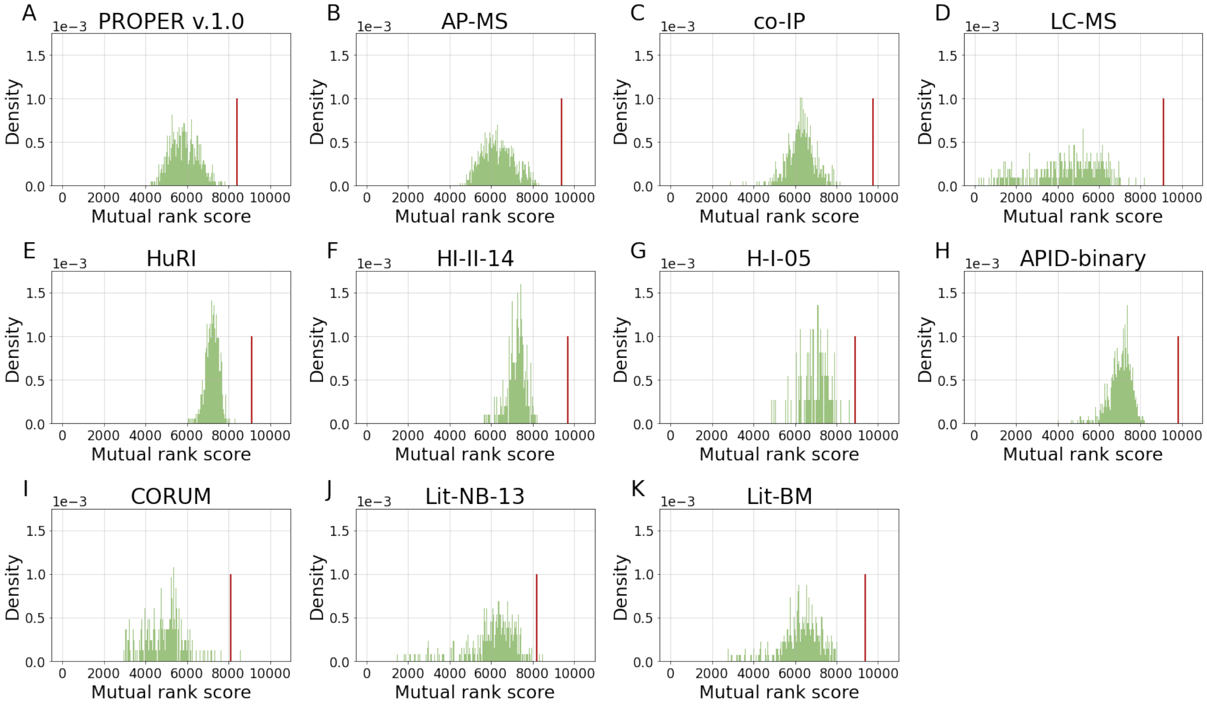


Figure S2.9: Co-expression level distribution of L3-predicted PPI networks
(A)-(K) Distributions of hub proteins' mutual rank score with their interaction partners in PPI networks after applying L3 predictions. The red line marks the average mutual rank score of all the PPIs in the L3-predicted network.

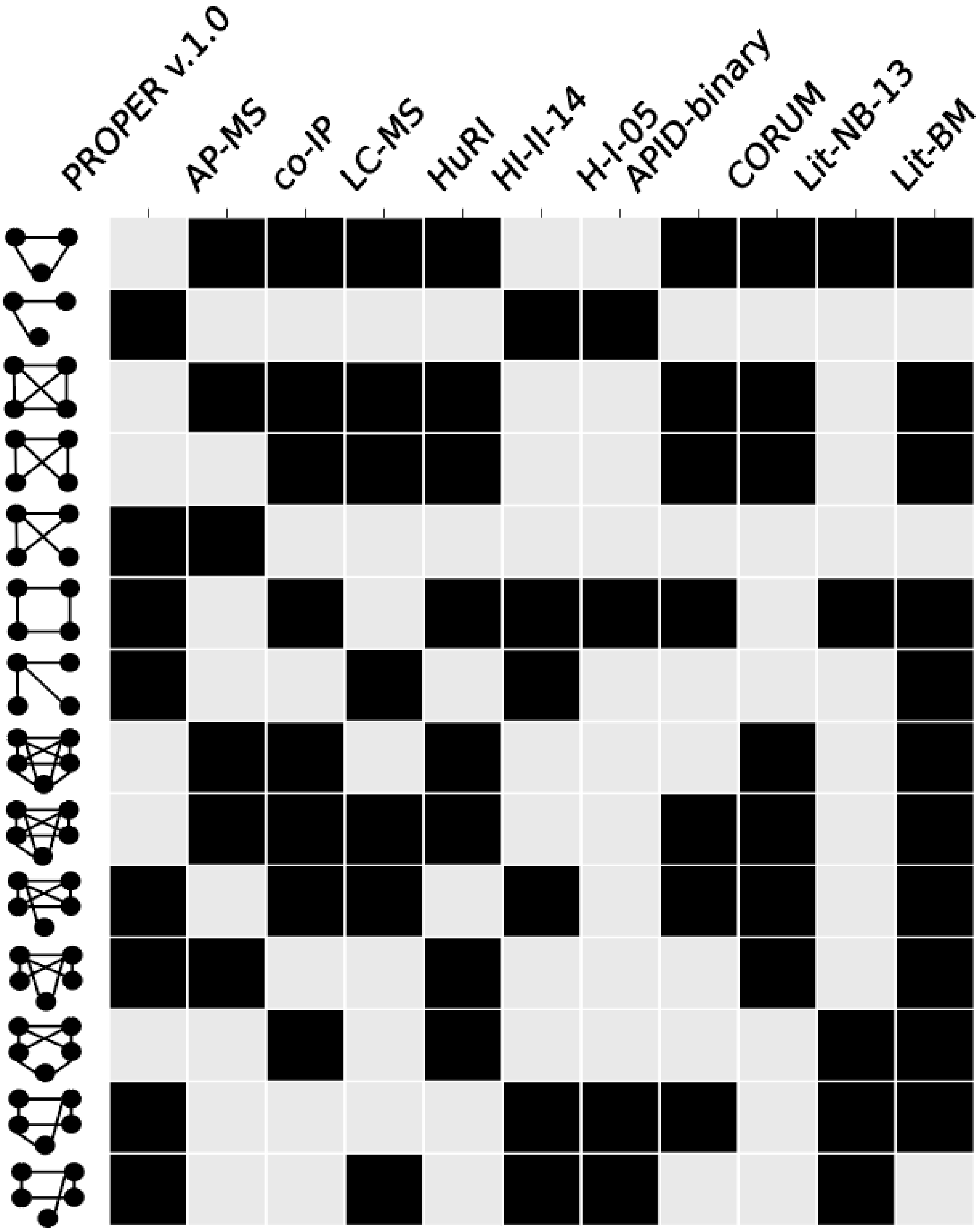


Figure S2.10: Network motifs of PPI networks
Size-3, size-4, and size-5 network motifs discovered in PPI networks.

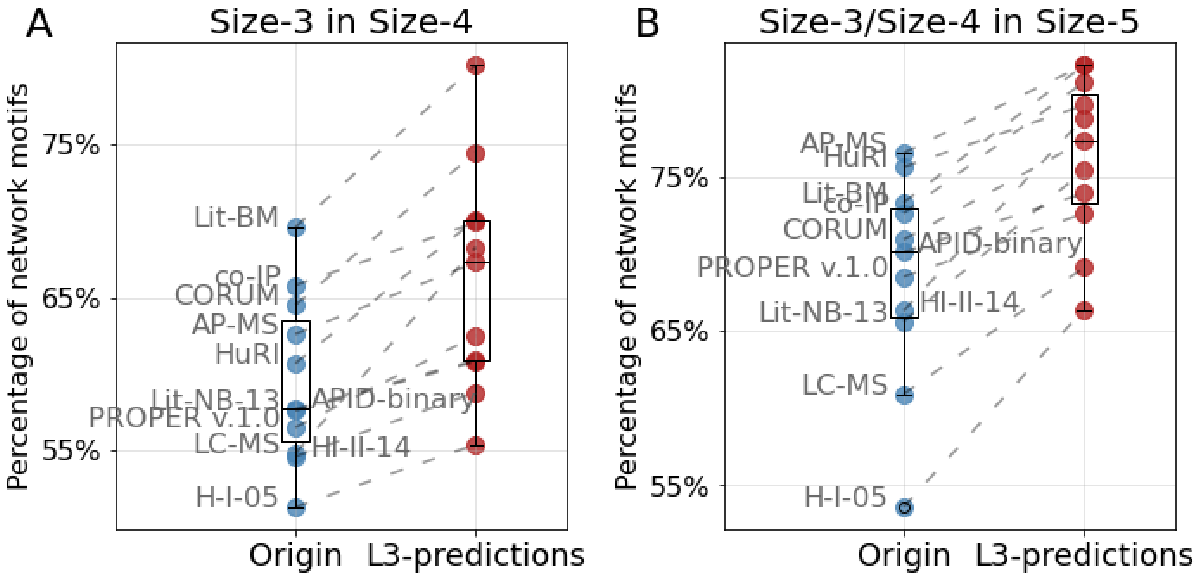


Figure S2.11: Sub-motifs of PPI network
(A) Percentage of size-4 network motifs that contain size-3 network motifs in origin and L3-predicted PPI networks. (B) Percentage of size-5 network motifs that contain size-3 or size-4 network motifs in origin and L3-predicted PPI networks.

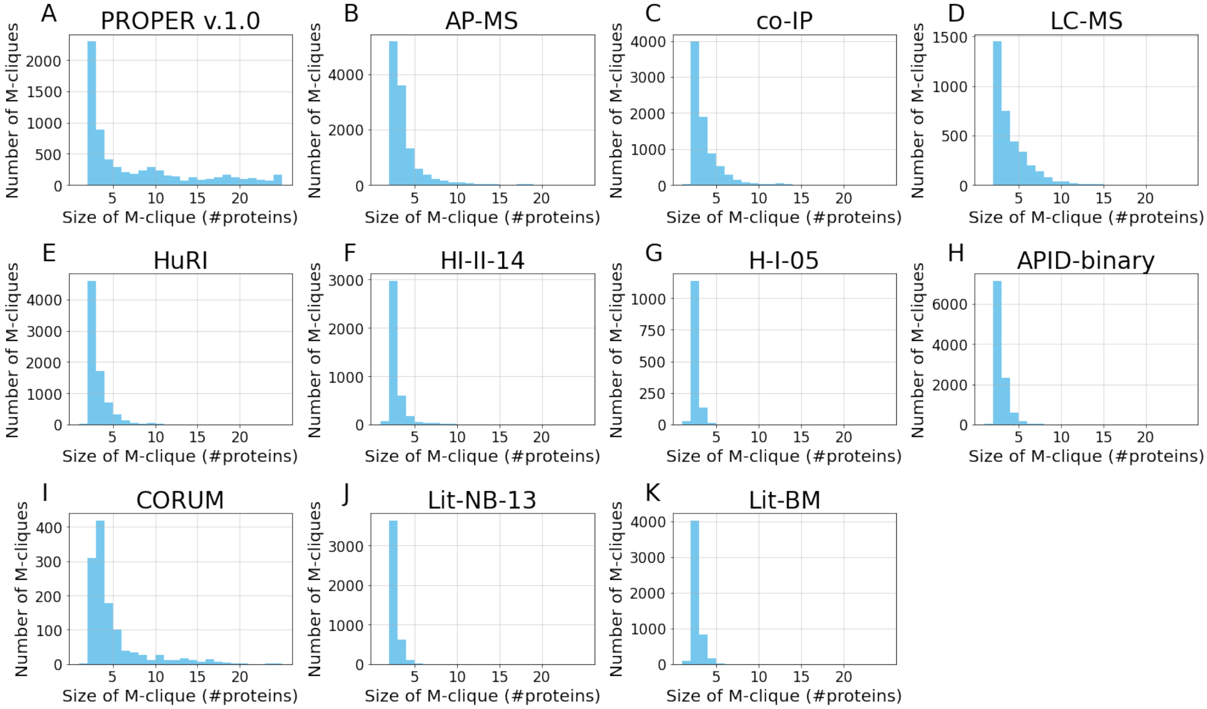


Figure S2.12: M-clique size distribution of PPI networks
(A)-(K) Distributions of M-clique size (number of proteins in the M-clique) in PPI networks.

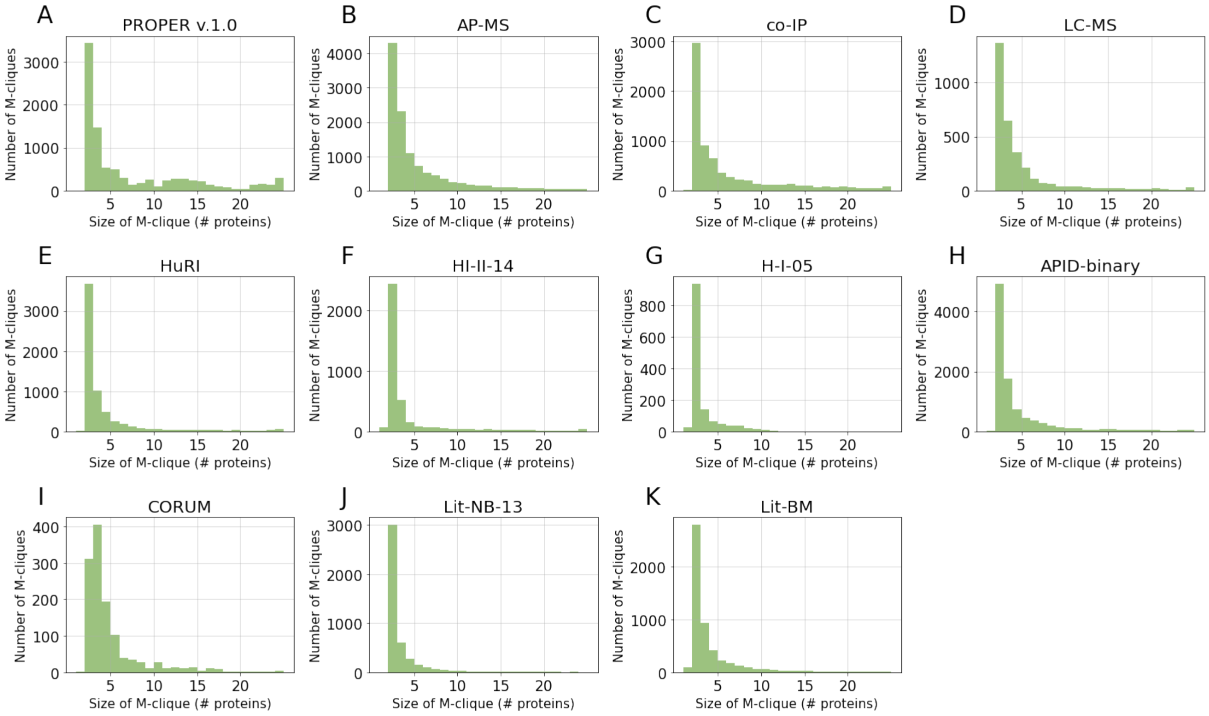


Figure S2.13: M-clique size distribution of L3-predicted PPI networks
(A)-(K) Distributions of M-clique size (number of proteins in the M-clique) in PPI networks after applying L3 predictions.

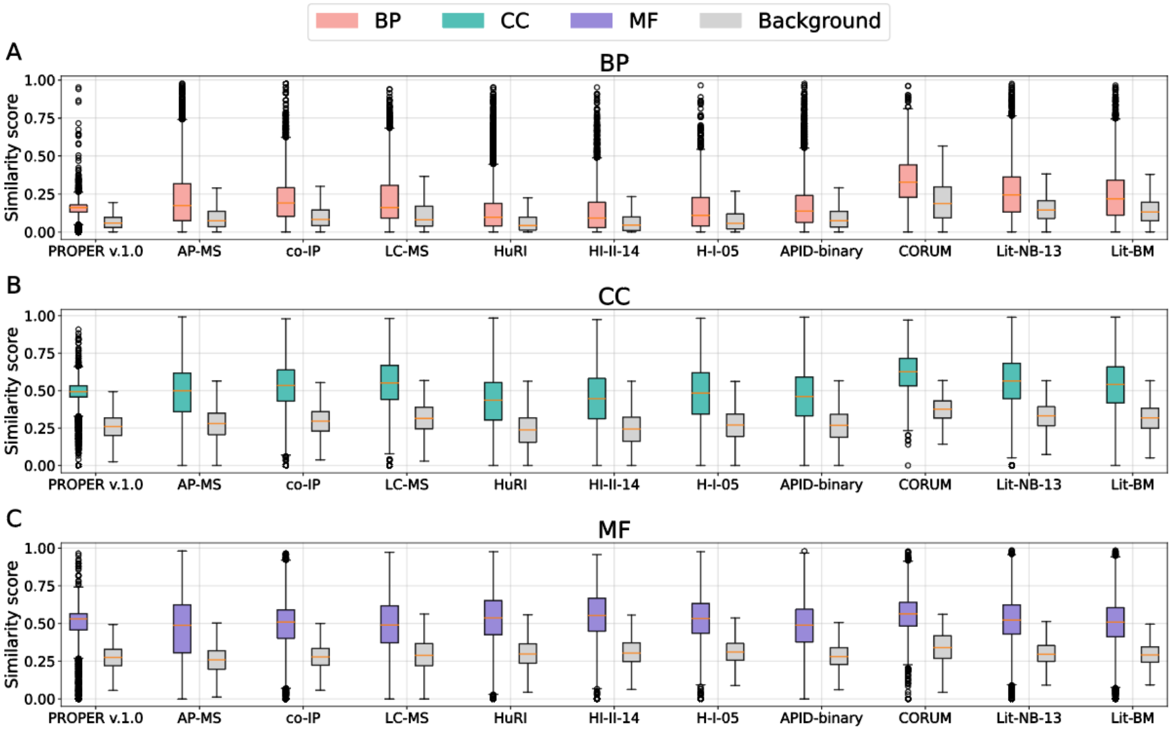


Figure S2.14: GO similarity score of M-cliques in PPI networks
Distributions of average BP (A), CC (B, green), and MF (C, purple) GO similarity score of M-cliques and distributions of their corresponding average background score in PPI networks.

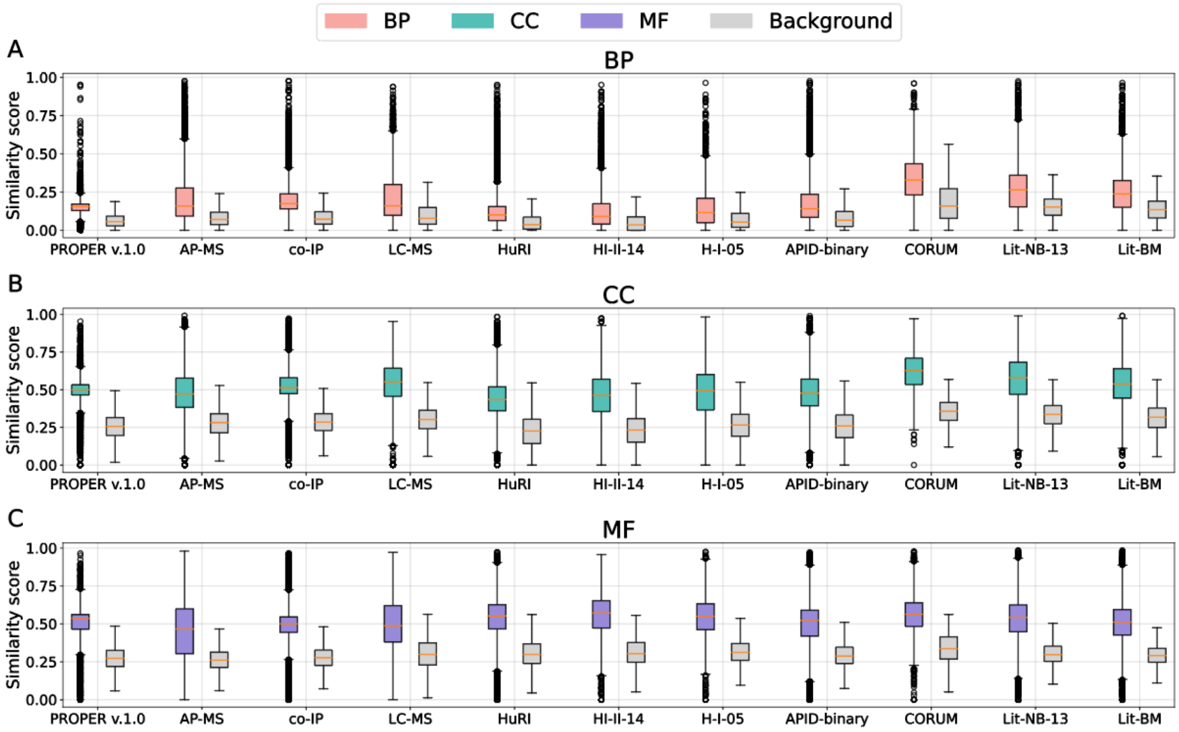


Figure S2.15: GO similarity score of M-cliques in L3-predicted PPI networks Distributions of average BP (A), CC (B, green), and MF (C, purple) GO similarity score of M-cliques and distributions of their corresponding average background score in PPI networks after applying L3 predictions.

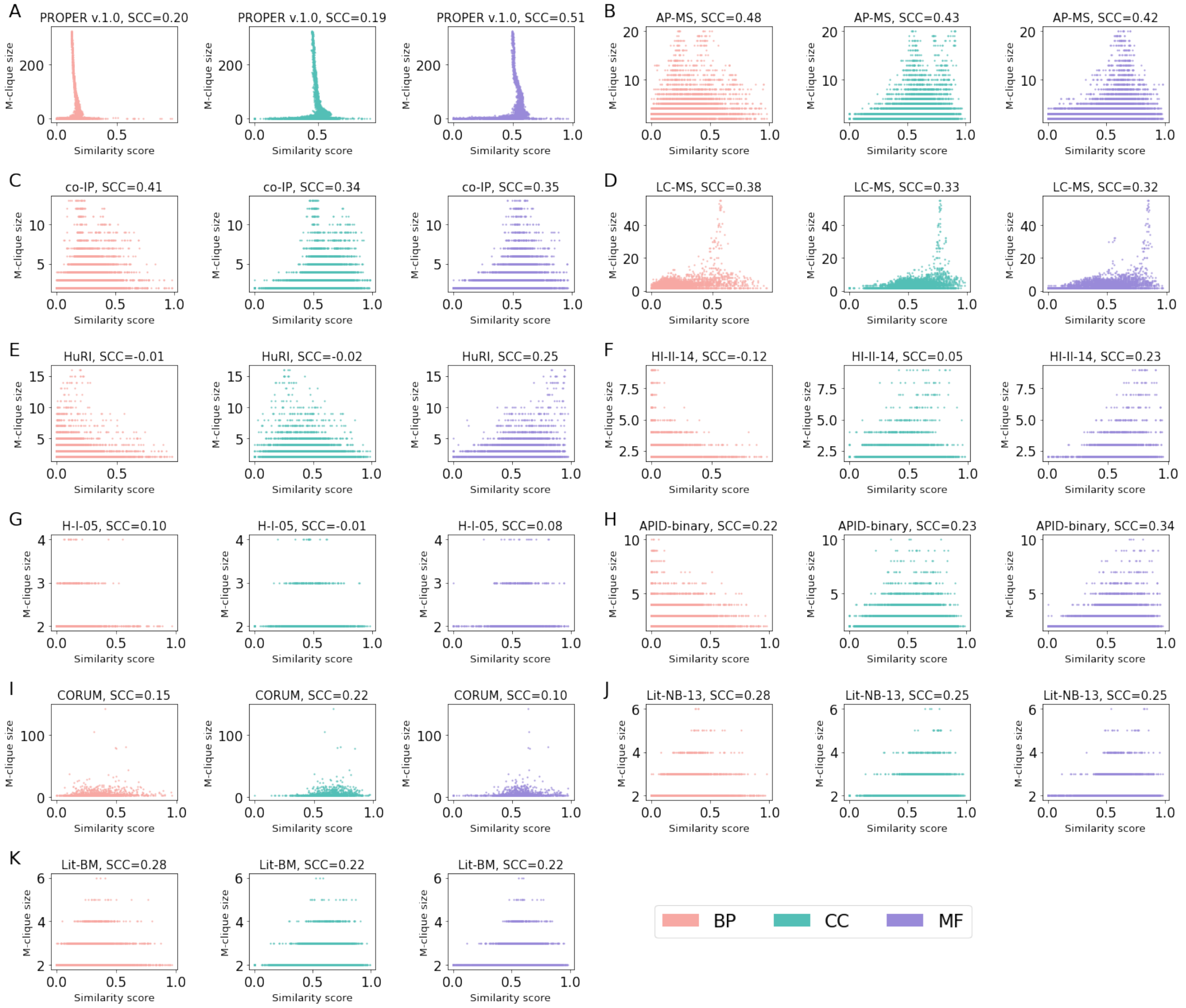


Figure S2.16: Correlation between M-clique features in PPI networks
(A)-(K) Correlation between M-clique size and each of the M-clique BP (pink), CC (green), and MF (purple) similarity scores in PPI networks.

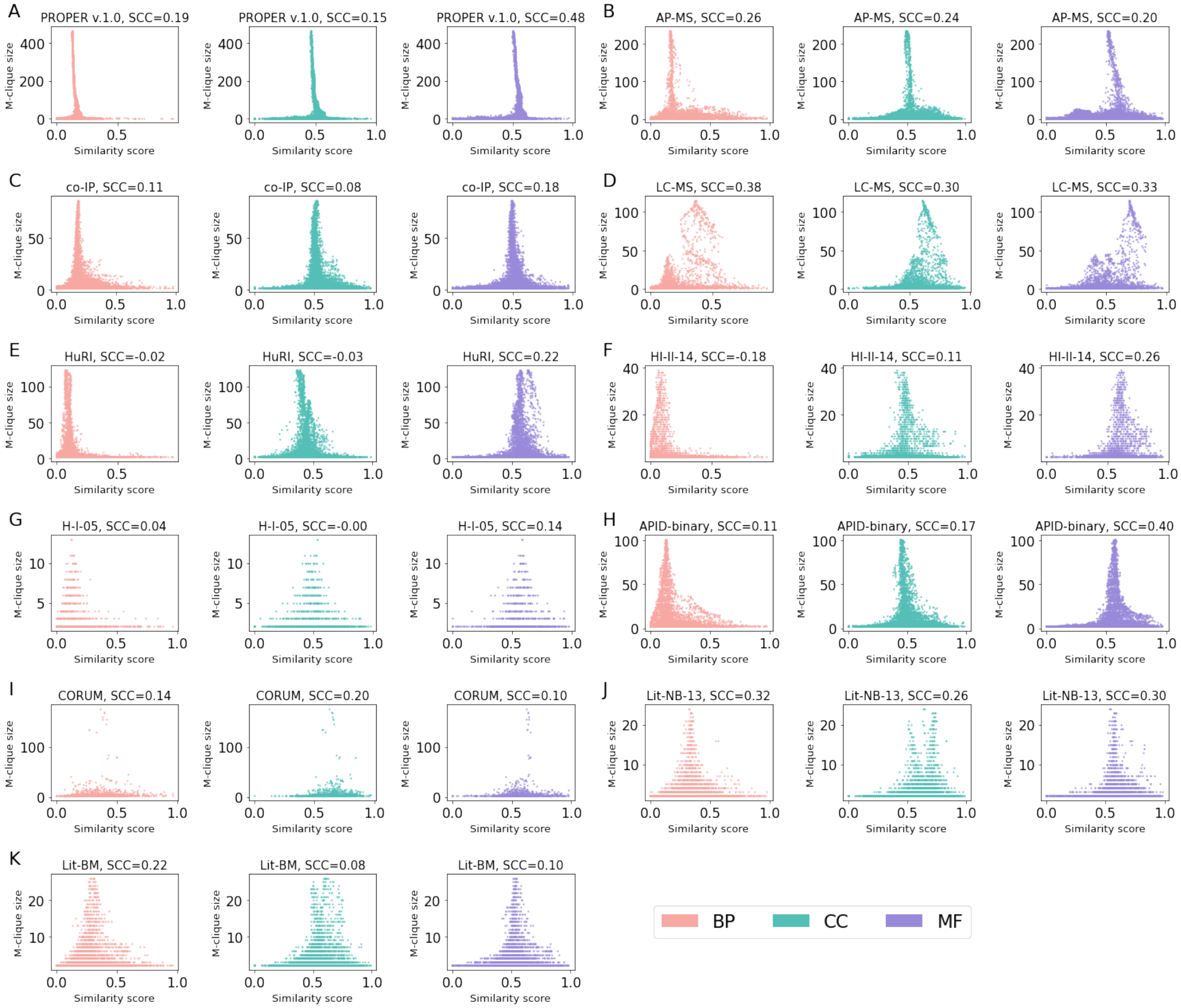


Figure S2.17: Correlation between M-clique features in L3-predicted PPI networks
(A)-(K) Correlation between M-clique size and each of the M-clique BP (pink), CC (green), and MF (purple) similarity scores in PPI networks after applying L3 predictions.

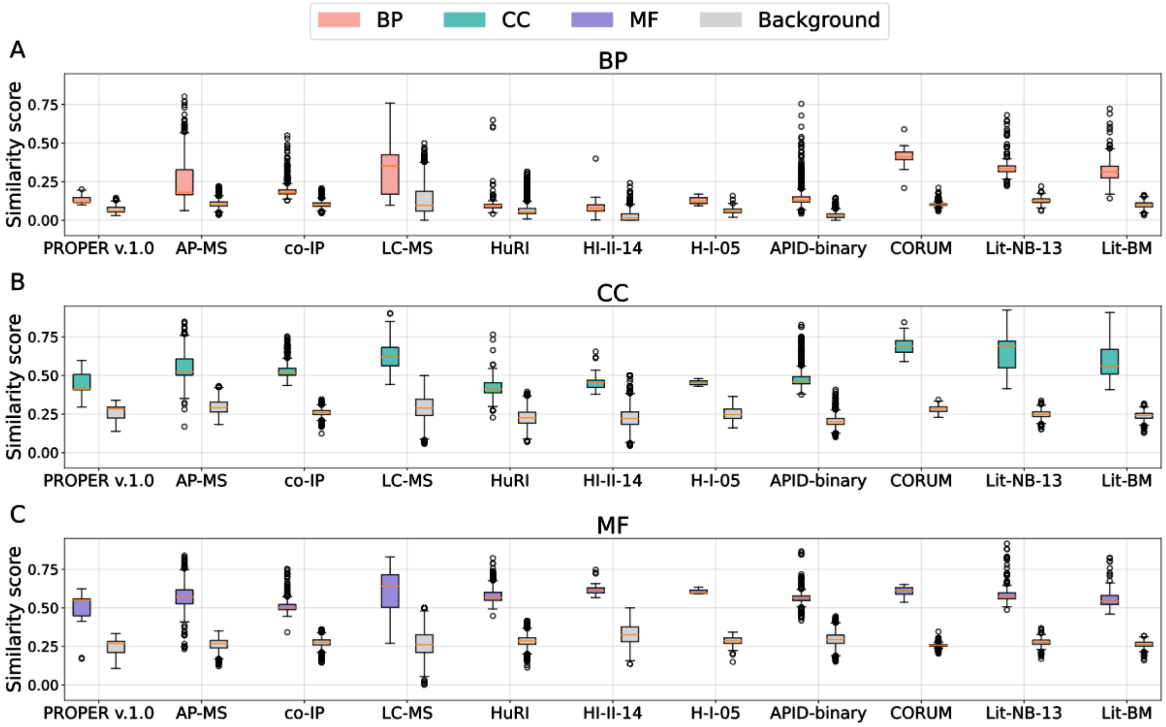


Figure S2.18: Go similarly score of merged M-cliques in L3-predicted PPI networks Distributions of average BP (A), CC (B, green), and MF (C, purple) GO similarity score of merged M-cliques and distributions of their corresponding average background score in PPI networks after applying L3 predictions.

Supplementary tables

Table S2.1: The human PPI datasets used
The human PPI datasets used in this work, including APID's subsets, collections of literature reported binary and co-complex PPIs, and PROPER-seq derived PPIs

Name	Description	# PPIs	PPI type
PROPER v.1.0	PROPER-seq detected PPIs by Johnson et al., 2021	210,518	binary, co-complex
AP-MS	Affinity purification-mass spec detected PPIs that are included in APID	131,224	co-complex
Co-IP	Co-IP detected PPIs that are included in APID	50,290	co-complex
LC-MS	Liquid chromatography–mass spec detected PPIs that are included in APID	33,195	co-complex
HuRI	Y2H detected PPIs by Luck et al., 2020	52,516	binary
HI-II-14	Y2H detected PPIs by Rolland et al., 2014	14,308	binary
H-I-05	Y2H detected PPIs by Rual et al., 2005	2,781	binary
APID-binary	Binary PPIs curated into the APID database	51,466	binary
CORUM	PPIs in protein complex curated by Giurgiu et al., 2019	39,103	co-complex
Lit-NB-13	Non-binary PPIs curated by Rolland et al., 2014	10,152	co-complex
Lit-BM	Binary PPIs curated by Luck et al., 2020	13,441	binary

Table S2.2: Summary of PPI networks

Number of PPIs, number of proteins, and number of M-cliques in PPI networks before (origin) and after L3 predictions.

Name		# PPIs	# Proteins	# M-cliques
PROPER v1.0	Origin	210,518	8,635	7,178
	L3 predictions	546,782		7,048
AP-MS	Origin	131,224	13,650	11,900
	L3 predictions	650,691		12,081
Co-IP	Origin	50,290	9,088	7,948
	L3 predictions	138,301		8,194
LC-MS	Origin	33,195	4,548	3,557
	L3 predictions	70,437		3,645
HuRI	Origin	52,516	8,267	7,682
	L3 predictions	167,105		7,463
HI-II-14	Origin	14,308	4,386	3,961
	L3 predictions	25,862		3,894
H-I-05	Origin	2,781	1,556	1,314
	L3 predictions	3,849		1,349
APID-binary	Origin	51,466	12,572	10,406
	L3 predictions	167,105		10,462
CORUM	Origin	39,103	3,435	2,264
	L3 predictions	52,287		2,278
Lit-NB-13	Origin	10,152	5,382	4,390
	L3 predictions	17,557		4,505
Lit-BM	Origin	13,441	6,047	5,114
	L3 predictions	24,716		5,120

Table S2.3: Number of source and merged M-cliques in PPI networks

Name	# Source M-cliques	# Merged M-cliques
PROPER v.1.0	797	552
AP-MS	1,082	698
Co-IP	427	278
LC-MS	514	409
HuRI	658	460
HI-II-14	78	51
H-I-05	26	10
APID-Y2H	455	305
CORUM	22	14
Lit-NB-13	190	113
Lit-BM	236	151

2.6 Methods and materials

Downloading PPI datasets

PROPER v.1.0 was downloaded from https://genemo.ucsd.edu/proper/. PPIs were downloaded as a MITAB file from the Agile Protein Interactomes DataServer (APID) at http://cicblade.dep.usal.es:8080/APID/init.action. AP-MS and Co-IP consist of PPIs identified by the corresponding labels in the 'Interaction detection method' column of the downloaded MITAB file. LC-MS consists of PPIs identified by the label of "biochemistry" in the 'Interaction detection method' column and specifying "Publication first author" as "Wan, C. et al. (2015)" (Wan et al. 2015), "Havugimana, PC. et al. (2012)" (Havugimana et al. 2012) and "Kristensen, AR. et al. (2012)" (Kristensen, Gsponer, and Foster 2012). APID-binary consists of PPIs identified by the label of "two-hybrid" in the "Interaction detection method" column of the MITAB file. HuRI, HI-II-14, H-I-05, and Lit-BM were downloaded from http://www.interactome-atlas.org/download. Human protein complexes were downloaded as a TXT file from http://mips.helmholtz-muenchen.de/corum/#download. CORUM consists of PPIs that are pair-wise formed by the proteins within the same protein complex according to the TXT file. Lit-NB-13 was downloaded from http://interactome.dfci.harvard.edu/H_sapiens/.

Calculating clustering coefficient, node degree, shortest path length, maximal cliques of PPI networks

The clustering coefficient of node, the degree of node, the shortest path length between pairwise nodes, and the maximal cliques of PPI networks were calculated and detected by NetworkX (Hagberg 2008), implemented in Python.

Implementation of the L3 algorithm

L3 link prediction algorithm, as described by Kovacs et al. (Kovacs et al. 2019), was implemented in Python to predict PPIs from the experimentally derived PPI networks.

Calculating protein co-expression levels

We used the human gene co-expression data to calculate protein co-expression levels in this work. The data was downloaded from COXPRESdb at https://coxpresdb.jp/download/ (Obayashi et al. 2019).

Identification of network motifs

MFinder 1.21(Ciriello and Guerra 2008) was used to find all occurrences of size-3, size-4, and size-5 networks in the PPI networks. Sub-networks with a z-score over 20 were identified as the network motifs in this study.

Implementation of SkinnyDip algorithm

SkinnyDip algorithm, as described by Maurus and Plant (Maurus and Plant 2016), was implemented in Python to determine the modularity type of protein co-expression distributions of PPI networks.

Calculating Gene Ontology similarity score

We used GOGO software downloaded from http://dna.cs.miami.edu/GOGO/ (Zhao and Wang 2018) to calculate Biological Process, Cellular Component, and Molecular Function similarity score of PPIs.

2.7 Acknowledgement

Chapter 2, in full, is currently being prepared for submission for publication of the material. Qi. Zhijie; Zhong, Sheng. The dissertation/thesis author was the primary investigator and author of this paper.

2.8 Reference

Agarwal, S., C. M. Deane, M. A. Porter, and N. S. Jones. 2010. 'Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks', *PLoS Comput Biol*, 6: e1000817.

Alonso-Lopez, D., F. J. Campos-Laborie, M. A. Gutierrez, L. Lambourne, M. A. Calderwood, M. Vidal, and J. De Las Rivas. 2019. 'APID database: redefining protein-protein interaction experimental evidences and binary interactomes', *Database (Oxford)*, 2019.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium', *Nat Genet*, 25: 25-9.

Batada, N. N., L. D. Hurst, and M. Tyers. 2006. 'Evolutionary and physiological importance of hub proteins', *PLoS Comput Biol*, 2: e88.

Chang, X., T. Xu, Y. Li, and K. Wang. 2013. 'Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs', *Sci Rep*, 3: 1691.

Ciavolella, M., P. E. Puddu, M. Schiariti, C. Ciani, E. Cerquetani, D. Scali, C. Giannitti, and A. Reale. 1991. 'Exponential fit of QT interval-heart rate relation during exercise used to diagnose stress-induced myocardial ischemia', *J Electrocardiol*, 24: 145-53.

Ciriello, G., and C. Guerra. 2008. 'A review on models and algorithms for motif discovery in protein-protein interaction networks', *Brief Funct Genomic Proteomic*, 7: 147-56.

Darehmiraki, M. 2009. 'A new solution for maximal clique problem based sticker model', *Biosystems*, 95: 145-9.

Fox, A. D., B. J. Hescott, A. C. Blumer, and D. K. Slonim. 2011. 'Connectedness of PPI network neighborhoods identifies regulatory hub proteins', *Bioinformatics*, 27: 1135-42.

Fraser, H. B. 2005. 'Modularity and evolutionary constraint on proteins', *Nat Genet*, 37: 351-2.

Free, R. B., L. A. Hazelwood, and D. R. Sibley. 2009. 'Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy', *Curr Protoc Neurosci*, Chapter 5: Unit 5 28.

Giurgiu, M., J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and A. Ruepp. 2019. 'CORUM: the comprehensive resource of mammalian protein complexes-2019', *Nucleic Acids Res*, 47: D559-D63.

Hagberg, Aric and Swart, Pieter and S Chult, Danie. 2008. 'Exploring network structure, dynamics, and function using networkx'.

Havugimana, P. C., G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, M. Babu, S. A. Craig, P. Hu, C. Wan, J. Vlasblom, V. U. Dar, A. Bezginov, G. W. Clark, G. C. Wu, S. J. Wodak, E. R. Tillier, A. Paccanaro, E. M. Marcotte, and A. Emili. 2012. 'A census of human soluble protein complexes', *Cell*, 150: 1068-81.

He, Xionglei, and Jianzhi Zhang. 2006. 'Why do hubs tend to be essential in protein networks?', *PLoS genetics*, 2: e88-e88.

Huttlin, E. L., R. J. Bruckner, J. Navarrete-Perea, J. R. Cannon, K. Baltier, F. Gebreab, M. P. Gygi, A. Thornock, G. Zarraga, S. Tam, J. Szpyt, B. M. Gassaway, A. Panov, H. Parzen, S. Fu, A. Golbazi, E. Maenpaa, K. Stricker, S. Guha Thakurta, T. Zhang, R. Rad, J. Pan, D. P. Nusinow, J. A. Paulo, D. K. Schweppe, L. P. Vaites, J. W. Harper, and S. P. Gygi. 2021. 'Dual proteome-scale networks reveal cell-specific remodeling of the human interactome', *Cell*, 184: 3022-40 e28.

Johnson, K. L., Z. Qi, Z. Yan, X. Wen, T. C. Nguyen, K. Zaleta-Rivera, C. J. Chen, X. Fan, K. Sriram, X. Wan, Z. B. Chen, and S. Zhong. 2021. 'Revealing protein-protein interactions at the transcriptome scale by sequencing', *Mol Cell*, 81: 3877.

Jordan, F., T. P. Nguyen, and W. C. Liu. 2012. 'Studying protein-protein interaction networks: a systems view on diseases', *Brief Funct Genomics*, 11: 497-504.

Kovacs, I. A., K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D. K. Kim, N. Kishore, T. Hao, M. A. Calderwood, M. Vidal, and A. L. Barabasi. 2019. 'Network-based prediction of protein interactions', *Nat Commun*, 10: 1240.

Kristensen, A. R., J. Gsponer, and L. J. Foster. 2012. 'A high-throughput approach for measuring temporal changes in the interactome', *Nat Methods*, 9: 907-9.

Luck, K., D. K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, D. Choi, A. G. Cote, M. Daley, S. Deimling, A. Desbuleux, A. Dricot, M. Gebbia, M. F. Hardy, N. Kishore, J. J. Knapp, I. A. Kovacs, I. Lemmens, M. W. Mee, J. C. Mellor, C. Pollis, C. Pons, A. D. Richardson, S. Schlabach, B. Teeking, A. Yadav, M. Babor, D. Balcha, O. Basha, C. Bowman-Colin, S. F. Chin, S. G. Choi, C. Colabella, G. Coppin, C. D'Amata, D. De Ridder, S. De Rouck, M. Duran-Frigola, H. Ennajdaoui, F. Goebels, L. Goehring, A. Gopal, G. Haddad, E. Hatchi, M. Helmy, Y. Jacob, Y. Kassa, S. Landini, R. Li, N. van Lieshout, A. MacWilliams, D. Markey, J. N. Paulson, S. Rangarajan, J. Rasla, A. Rayhan, T. Rolland, A. San-Miguel, Y. Shen, D. Sheykhkarimli, G. M. Sheynkman, E. Simonovsky, M. Tasan, A. Tejeda, V. Tropepe, J. C. Twizere, Y. Wang, R. J. Weatheritt, J. Weile, Y. Xia, X. Yang, E. Yeger-Lotem, Q. Zhong, P. Aloy, G. D. Bader, J. De Las Rivas, S. Gaudet, T. Hao, J. Rak, J. Tavernier, D. E. Hill, M. Vidal, F. P. Roth, and M. A. Calderwood. 2020. 'A reference map of the human binary protein interactome', *Nature*, 580: 402-08.

Maurus, Samuel, and Claudia Plant. 2016. "Skinny-dip: Clustering in a Sea of Noise." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1055–64. San Francisco, California, USA: Association for Computing Machinery.

Morris, J. H., G. M. Knudsen, E. Verschueren, J. R. Johnson, P. Cimermancic, A. L. Greninger, and A. R. Pico. 2014. 'Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions', *Nat Protoc*, 9: 2539-54.

Nacher, J. C., M. Hayashida, and T. Akutsu. 2009. 'Emergence of scale-free distribution in protein-protein interaction networks based on random selection of interacting domain pairs', *Biosystems*, 95: 155-9.

Obayashi, T., Y. Kagaya, Y. Aoki, S. Tadaka, and K. Kinoshita. 2019. 'COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference', *Nucleic Acids Res*, 47: D55-D62.

Peng, X., J. Wang, W. Peng, F. X. Wu, and Y. Pan. 2017. 'Protein-protein interactions: detection, reliability assessment and applications', *Brief Bioinform*, 18: 798-819.

Pitt, J. J. 2009. 'Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry', *Clin Biochem Rev*, 30: 19-34.

Przytycka, T. M., and Y. K. Yu. 2004. 'Scale-free networks versus evolutionary drift', *Comput Biol Chem*, 28: 257-64.

Rolland, T., M. Tasan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A. R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J. C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A. L. Barabasi, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal. 2014. 'A proteome-scale map of the human interactome network', *Cell*, 159: 1212-26.

Rual, J. F., K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. 2005. 'Towards a proteome-scale map of the human protein-protein interaction network', *Nature*, 437: 1173-8.

Saito, Takaya, and Marc Rehmsmeier. 2015. 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PLoS One*, 10: e0118432-e32.

Schober, Patrick, Christa Boer, and Lothar A. Schwarte. 2018. 'Correlation Coefficients: Appropriate Use and Interpretation', *Anesthesia & Analgesia*, 126.

Vandereyken, K., J. Van Leene, B. De Coninck, and B. P. A. Cammue. 2018. 'Hub Protein Controversy: Taking a Closer Look at Plant Stress Response Hubs', *Front Plant Sci*, 9: 694.

Wan, C., B. Borgeson, S. Phanse, F. Tu, K. Drew, G. Clark, X. Xiong, O. Kagan, J. Kwan, A. Bezginov, K. Chessman, S. Pal, G. Cromar, O. Papoulas, Z. Ni, D. R. Boutz, S. Stoilova, P. C. Havugimana, X. Guo, R. H. Malty, M. Sarov, J. Greenblatt, M. Babu, W. B. Derry, E. R. Tillier, J. B. Wallingford, J. Parkinson, E. M. Marcotte, and A. Emili. 2015. 'Panorama of ancient metazoan macromolecular complexes', *Nature*, 525: 339-44.

Wang, J., B. Liu, M. Li, and Y. Pan. 2010. 'Identifying protein complexes from interaction networks based on clique percolation and distance restriction', *BMC Genomics*, 11 Suppl 2: S10.

Watts, Duncan J., and Steven H. Strogatz. 1998. 'Collective dynamics of ‘small-world’ networks', *Nature*, 393: 440-42.

Yang, L., and X. Tang. 2014. 'Protein-protein interactions prediction based on iterative clique extension with gene ontology filtering', *ScientificWorldJournal*, 2014: 523634.

Yang, L., X. Zhao, and X. Tang. 2014. 'Predicting disease-related proteins based on clique backbone in protein-protein interaction network', *Int J Biol Sci*, 10: 677-88.

Yu, H., A. Paccanaro, V. Trifonov, and M. Gerstein. 2006. 'Predicting interactions in protein networks by completing defective cliques', *Bioinformatics*, 22: 823-9.

Zhao, C., and Z. Wang. 2018. 'GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms', *Sci Rep*, 8: 15107.

CHAPTER 3 HIGH THROUGHPUT MAPPING OF RNA-PROTEIN INTERACTIONS

3.1 Abstract

We describe PRIM-seq (protein-RNA interaction mapping by sequencing) to systematically map RNA-protein interactions (RPIs) *in vitro*. PRIM-seq utilizes an RNA-barcoded protein library to interact with RNAs and converts the interaction pairs into chimeric DNA sequences, which are further decoded by sequencing and mapping. We applied PRIM-seq to human embryonic kidney cells and identified 1,175,516 human RPIs (collected as the PRIM v.1.0 database). PRIM v.1.0 supports 4,418 RPIs curated in RNAInter and 1,569 RNA-binding proteins captured by pCLAP and RBDmap. PRIM-seq revealed PHGDH as an RNA binding protein. 5 previously uncharacterized interactions of PHGDH with PTMA, HNRNPA2B1, ATF4, BCLAF1, and BECN1 as the RNAs are validated by RIP-qPCR. PRIM-seq presents a time-effective technology to massively map RPIs and provides an extensive RPI network for studying RNA binding proteins, RNA binding domains, and RNA motifs with their secondary structures that bind the protein domains.

3.2 Introduction

Many cellular regulatory processes like RNA processing, mRNA localization, and protein synthesis require RNA-proteins interactions (RPIs) (Balcerak et al. 2019). However, mapping the human RPI network remains technically challenging. Two classes of RPI mapping methods have been pursued. The first class of methods, like CLIP-seq (Stork and Zheng 2016), RIP-seq (Zhao et al. 2010), and PAR-CLIP (Danan, Manickavel, and Hafner

2016), pulls down a protein of interest and examines its bound RNAs by sequencing. The second class of methods pulls down an RNA of interest and reveals the protein partners using mass spectrometry (RAP-MS (McHugh and Guttman 2018), PAIR (Zeng et al. 2006), MS2-BioTRAP (Tsai et al. 2011), ChiRP (Chu et al. 2011), CHART(Simon et al. 2011)), enzymatic labeling (TRIBE (McMahon et al. 2016), RaPID (Ramanathan et al. 2018)), or protein microarrays (ProtoArray (Syu, Dunn, and Zhu 2020)). As not all RNA binding proteins (RBPs) have been reported as RBPs and not all reported RBPs have high-affinity antibodies (Wheeler, Van Nostrand, and Yeo 2018), these methods are hard to be scaled up to generate a reference network for human RPIs. In other words, it would require significant time and resources to identify *de novo* RBPs and to resolve what RNAs that every RBP can bind with.

Here, we describe PRIM-seq (Protein-RNA Interaction Mapping by sequencing) for systematic identification of RBPs and the RNAs bound by every RBP in one single experiment. PRIM-seq utilizes the RNA barcodes of the proteins in the SMART-display library (Johnson et al. 2021) and ligates them with the proteins' interacted RNAs. The ligated pairs are then subjected to high-throughput sequencing for the identification of RPIs. We applied PRIM-seq to human embryonic kidney (HEK) cells to yield a human RPI network (PRIM v.1.0) involving 7,691 RBPs and 1,175,516 RPIs.

3.3 Design

3.3.1 Sequencing library preparation

PRIM-seq converts RPIs into chimeric sequences with the structure as cDNA1-linker-cDNA2 with one of the cDNAs derived from the RNA barcode of the interacted protein and the other cDNA derived from the interacted RNA (Figure 3.1). PRIM-seq starts with using SMART-display (Johnson et al. 2021) to generate a protein library from a certain cell type in which the proteins are attached with mRNA barcodes. The SMART-display library is then immobilized on streptavidin beads through the biotin on the puromycin linker sequence and incubated with the total RNAs extracted from the same cell type to allow for RNA-protein interactions. After removal of spurious interactions, the RNA is then ligated to the mRNA barcode of its bound protein to create a chimeric sequence in the form of cDNA1-linker-cDNA2 to represent the original RPI. These chimeric sequences are subsequently selected for and subjected to paired-end sequencing.

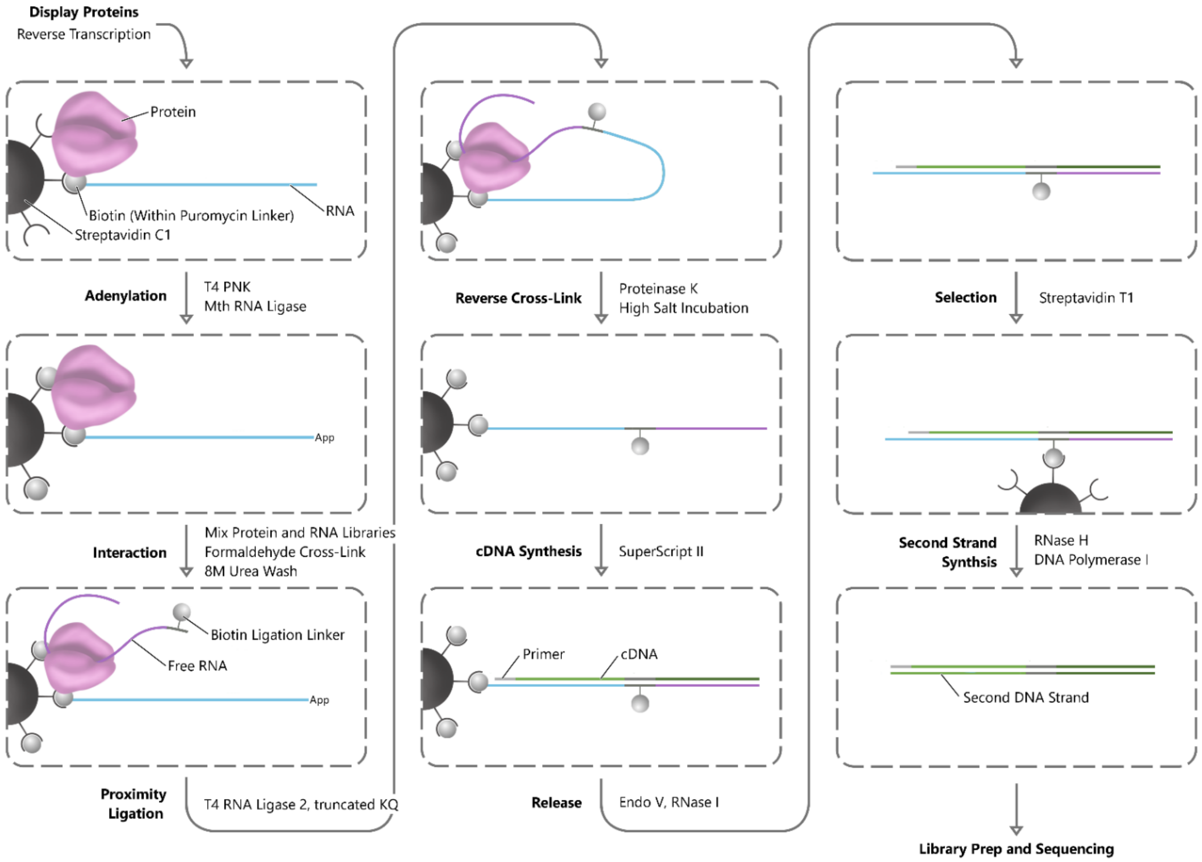


Figure 3.1: Sequencing library preparation.
Steps are indicated in bold font to the left of each process arrow, and primary enzymes or reagents used to accomplish each step are indicated to the right of the process arrow. The process begins with the stabilization of the display complexes on streptavidin magnetic beads. The RNA library ligated with the biotin ligation linker is combined with the immobilized display complexes to perform the RNA protein interaction step. The beads are washed to remove non-specific interactions. The DNA from the protein is digested with a non-palindromic restriction enzyme, and proximity ligation between the nucleic acids from the interacted pair is performed. The interacting pair is then reverser crosslinked by proteinase K digestion, and the ligated RNA is converted to double-stranded DNA. The DNA is then fragmented, and adaptor ligation for sequencing is performed before final streptavidin selection for the biotin-containing interaction linker and library amplification.

3.3.2 Identification of RPIs

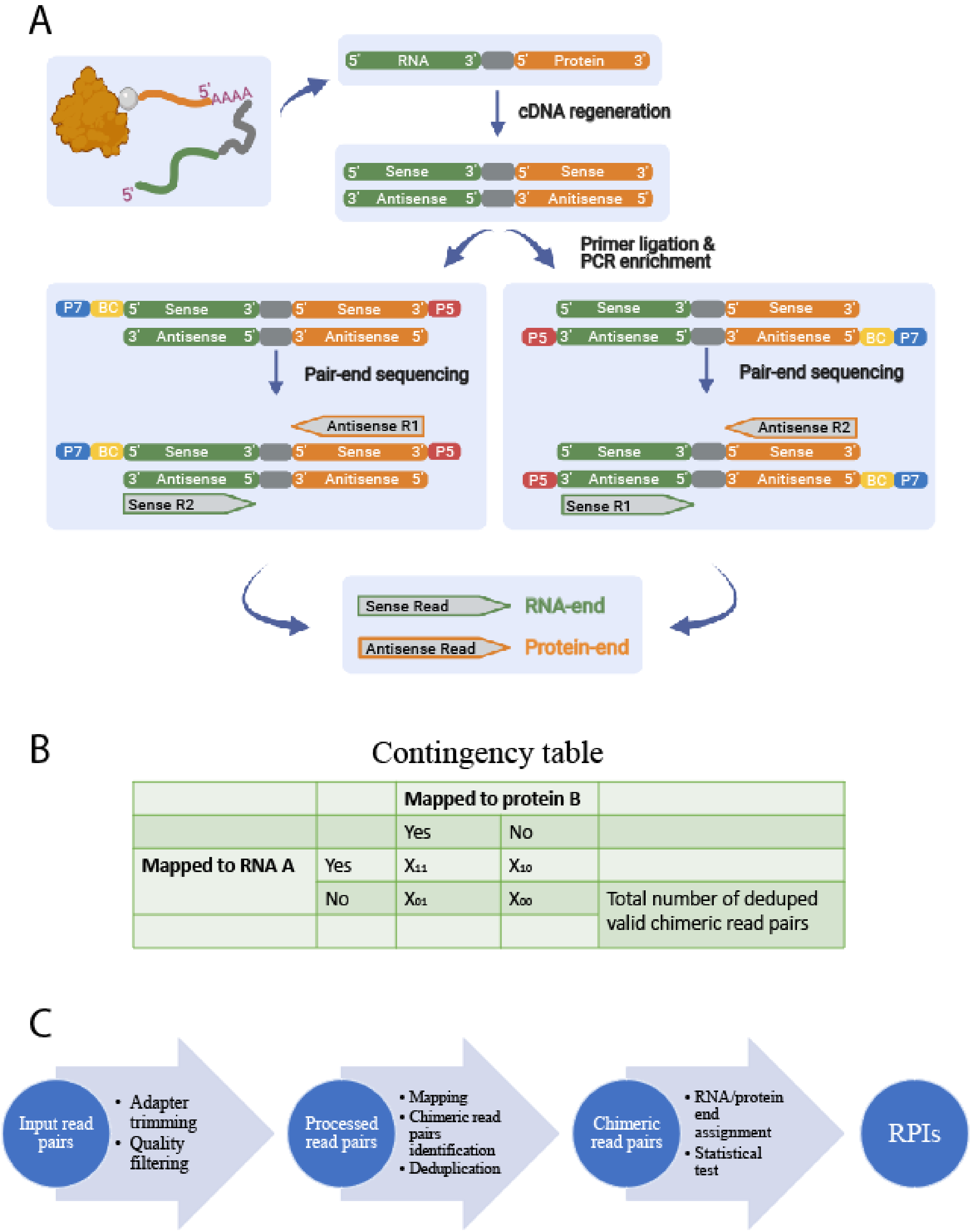
The next step is to identify RNA-protein pairs from the mapped read pairs of the sequencing library. We utilized the mapping strandness to first distinguish the RNA-end from the protein-end for each chimeric read pair. According to the experiment design and sequencing mechanism, the read-end of a chimeric read pair that was sensely mapped to a gene was derived from RNA and the read-end of the pair that was antisensely mapped to a gene was derived from protein (Figure 3.2A). We dropped the chimeric read pairs that were mapped both sensely or both antisensely to two genes. Then we subjected the valid chimeric read pairs on each RNA-protein pair to an association test. The null hypothesis is that the mapping of a read pair to one RNA is independent to the mapping of this read pair to its paired protein (Figure 3.2B). We used Bonferroni-Hochberg (BH) correction to account for multiple hypothesis tests (Benjamini and Hochberg 1995). An RNA-protein pair was identified (i.e., an RPI) by two criteria. First, the BH-corrected p-value derived from the association (Chi-square) test is smaller than 0.05. Second, the number of the chimeric read pairs mapped to this gene pair is no less than 2 times the average number of valid chimeric reads mapped to any RNA-protein pair (2 × number of all valid chimeric read pairs / number of all mapped RNA-protein pairs). Hereafter, we call these the default threshold, denoted as BH-corrected $p < 0.05$ and number of read pairs $> 2X$, where X is the expected number of read pairs mapped on a randomly chosen RNA-protein pair. Unless otherwise specified, all RPIs presented in the rest of this chapter were identified based on this default threshold. We implemented all data processing and statistical test steps into an open-source software package called PRIMseqTools (Figure 3.2C).

Figure 3.2: Identification of RPIs

(A) Identification of RNA-end and Protein-end from a chimeric read pair yielded from RPI-sequencing. Single strand chimeric RNA is formed by linking the RNA to the RNA barcode of the protein. After cDNA generation, two primers (P5 and P7) and a barcode (BC) are added to the DNA. After PCR enrichment, the primers together with barcode can only exist on one strand of the amplified DNA fragments, either the top or the bottom strand as shown in the figure. P7 primer, which will ligate to the flow cell during the sequencing process, is always at the 5'end of the strand. P5 primer, on the other hand, is always at the 3'end of the strand. The two possible DNA fragments get sequenced by pair-end sequencing. Based on the position of the primers, read1 can be sequenced either from the bottom strand or from the top strand. The same applies to read2. Two possible sequencing orientations are illustrated in the figure. The read that is sequenced sensely is always sequenced from the RNA end. The other read that is sequenced antisensely is always sequenced from the protein end.

(B) A continency table for the read pairs mapped to RNA A (rows) and protein B (columns). Every mapped read pair is assigned to one and only one cell in this contingency table. The null hypothesis is that the mapping of a read pair to one RNA is independent to the mapping of this read pair to the protein.

(B) Flowchart of PRIMseqTools for processing PRIM-seq data. Linker sequence and adaptor sequences were trimmed (Adaptor trimming). Low quality reads and reads that were too short were removed (Quality filtering). The resulting read pairs were mapped to Refseq genes (Mapping), and those with the two ends mapped to two different genes were obtained (Identification of chimeric read pairs). Non-redundant chimeric read pairs with one end sensely mapped to a gene and the other end antisensely mapped to a protein-coding gene (RNA/protein end assignment) were used as the input for the test of association (Statistical test).



		Mapped to protein B		
		Yes	No	
Mapped to RNA A	Yes	X_{11}	X_{10}	
	No	X_{01}	X_{00}	Total number of deduped valid chimeric read pairs

3.4 Results

3.4.1 Evaluations of PRIM-seq identified RPIs

We evaluated PRIM-seq identified RPIs based on their precision, recall, and their robustness against the subsampling process. We generated one PRIM-seq library from HEK293T cells which consist of 409,132,179 read pairs (Table S3.1). The library was named HEK1. HEK1 yielded approximately 6 million non-duplicate chimeric read pairs with one end sensely mapped to a gene (RNA-end) and the other end antisensely mapped to a protein-coding gene (protein-end). By using PRIMseqTools with the default threshold, we identified 1,175,516 RPIs from HEK1.

Precision and recall of PRIM-seq identified RPIs

We evaluated the precision and recall of the HEK1 derived RPIs with reference to previously characterized human RPIs (Saito and Rehmsmeier 2015). We obtained RPIs from RNA Interactome Database (RNAInter) (Kang et al. 2022), a most up-to-date repository of experimentally validated RPIs integrated from databases including starBase v2.0 (Li et al. 2014), ChiPBase v2.0 (Zhou et al. 2017), POSTAR2 (Zhu et al. 2019), TransmiR v2.0 (Tong et al. 2019) and miRTarBase (Huang et al. 2022). According to RNAInter, three types of experiments self-alone yielded more than 100,000 human RPIs. These are enhanced UV crosslinking and immunoprecipitation (eCLIP) (Van Nostrand et al. 2016), individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) (Huppertz et al. 2014), and crosslinking immunoprecipitation associated to high-throughput sequencing (CLIP-seq) (Stork and Zheng 2016), which have reported 301,020, 126,345 and 109,706 human RPIs respectively (Table S3.2). We then obtained 529 previously characterized RBPs from RBDmap (Castello et al. 2016) and 2,043 RBPs from pCLAP (Mullari et al. 2017)

(Table S3.2). 316 RBPs were characterized by both RBDmap and pCLAP. We further stratified the three reference sets of RPIs derived from eCLIP, iCLIP, and CLIP-seq into three RBP levels by limiting the search space of the interacted proteins. RBP-0 is the first level where the protein search space is the collection of all human protein coding genes. RBP-1 is the second level where the protein search space is the collection of RBPs captured by either RBDmap or pCLAP. RBP-2 is the strictest level where the protein search space is the collection of RBPs captured by both RBDmap and pCLAP. We plotted the precision and recall curves by comparing the HEK1 identified RPIs with the three reference sets under the three RBP levels. We also generated random datasets for each reference set by permutating the assignment of the chimeric read pairs within the corresponding protein search space. In all analyses, PRIM-seq-identified RPIs presented larger precisions and smaller recalls when the thresholds increased and exhibited better precisions and recalls than the permutation data (Figure S3.1). These suggest that RPIs derived from PRIM-seq are supported by previous literature and are well distinguished from the background of randomly sampled RNA-protein pairs.

Identification of RPIs from PRIM-seq with subsampling

We also subsampled 75%, 50% and 25% of the input reads from HEK1 PRIM-seq library to ask whether varying the sequencing depth of PRIM-seq library will affect the scale, precision, and recall of the identified RPIs. We found the number of identified RPIs decreases as the subsampling rate decreases (Figure S3.2A, E, I). We then computed precision and recall values of the HEK1 libraries at different subsampling rates using eCLIP, iCLIP, and CLIP-seq as the reference datasets under the protein search spaces defined by RBP-0, RBP-1, and RBP-2. We found that the resulting precision-recall curves overlap with each other for each individual comparison (Figure S3.2). These results suggest that the sequencing depth of PRIM-seq libraries only affects the scale of identified RPIs but does not affect their precision and recall. In other words, increasing the sequencing depth of PRIM-seq libraries will reveal more RNA-protein interactions of equal validity.

3.4.2 PRIM v.1.0: An extensive human RPI network

HEK1 library from PRIM-seq revealed 117,516 pairwise PRIs involving 8,440 RNAs and 7,691 proteins, which are collectively termed the RPIM v.1.0 network (Figure 3.3A). We developed a web interface to download, search, and visualize PRIM v.1.0 (https://genemo.ucsd.edu/prim). PRIM v.1.0 consists of RPIs of different RNA species including mRNA, lncRNA, rRNA, ncRNA, tRNA, snRNA, snoRNA, miRNA, etc., with RPIs of lncRNAs and rRNAs being the most abundant after mRNA (Figure 3.3B). Altogether, PRIM v.1.0 supports 4,481 previously experimentally characterized RPIs according to RNAInter. Adapted from Yu et al. (2008) and Venkatesan et al. (2009) (Venkatesan et al. 2009; Yu et al. 2008), we calculated RPI framework parameters to evaluate PRIM v.1.0. PRIM v.1.0 presents a screening completeness of 42.18%, a sampling sensitivity of 56.69%, an assay sensitivity of 2.91%, and a precision of 9.37% (Table S3.3).

We asked whether any functional groups of proteins are enriched in PRIM v.1.0. We applied GO enrichment analysis to RPIs of different RNA species and focused on GO terms that contained no more than 300 genes. We found ‘cytoplasmic translation, GO:0002181’is the most enriched GO term for proteins that interacted with rRNAs and tRNAs (Figure 3.3C, D), and ‘mRNA splicing, GO:0000398’is the most enriched GO term for proteins that interacted with snRNAs and miRNAs (Figure 3.3E, F). These subnetworks suggested the possibility of using RRIM v.1.0 to reveal cellular processes and signaling pathways that involve RNA-protein interactions.

Enrichment of known RBPs in RPIM v.1.0

We asked whether previously characterized RBPs and RPIs with the characterized RBPs are enriched in PRIM v.1.0. We used the proteins captured by RBDmap and pCLAP as the reference RBPs. In PRIM v.1.0, 1,410 proteins were detected by pCLAP (odds ratio=5.1, p-value<10^{-32}, Fisher's exact test) (Figure 3.4A), 438 proteins were detected by RBDmap (odds ratio=8.9, p-value<10^{-32}, Fisher's exact test) (Figure 3.4B), and 279 proteins were detected by both pCLAP and RBDmap (odds ratio=12.9, p-value<10^{-32}, Fisher's exact test) (Figure 3.4C). As the threshold of calling RPIs from PRIM-seq increases, we observed increases in the odds ratios (Figure 3.4D). These results suggest previously characterized RBPs are enriched in PRIM v.1.0 and PRIM-seq derived RPIs with a higher confidence level are more likely to contain these RBPs.

In PRIM v.1.0, 58,204 RPIs contain proteins captured by pCLAP (p-value<10^{-32}, binomial test), 38,329 RPIs contain proteins captured by RBDmap (p-value<10^{-32}, binomial test), and 32,876 RPIs contain proteins captured by both pCLAP and RBDmap (p-value<10^{-32}, binomial test) (Figure 3.4E). We also found that the highly interacted proteins in PRIM v.1.0 are more likely to be the RBPs detected by pCLAP and RBDmap (Figure 3.4F). By assuming that RPIs with or without the previously characterized RBPs share the same probability of being detected, this suggests that PRIM-seq is more likely to detect RPIs that involve the previously characterized RBPs, further supporting that PRIM v.1.0 is enriched with the known RBPs as well as RPIs that involve these RBPs.

We proceeded to ask whether the proteins in PRIM v.1.0 are enriched with reads that are mapped to the RNA binding domains (RBDs) of the corresponding proteins. We obtained 5,624 RBDs from pCLAP and RBDmap (Table S3.2). Out of the 1,278,091 read pairs from which the RPIs contain RNA binding proteins characterized by pCLAP or RBDmap, 319,951 (25.03%) read pairs have their protein-end aligned to the RBDs on the proteins (p-value<10^{-32}, one-sided binomial test). Meanwhile, on individual protein level, we found 787 RBPs in PRIM v.1.0 being significantly enriched with reads mapped to RBDs (BH corrected p-value<0.05, one-sided binomial test, Figure 3.4G). Assuming the exon regions of proteins share equal probabilities of being translated and interacting with RNA, these results suggest PRIM-seq is more likely to capture RPIs with RNAs interacting with the RNA binding domains of the proteins.

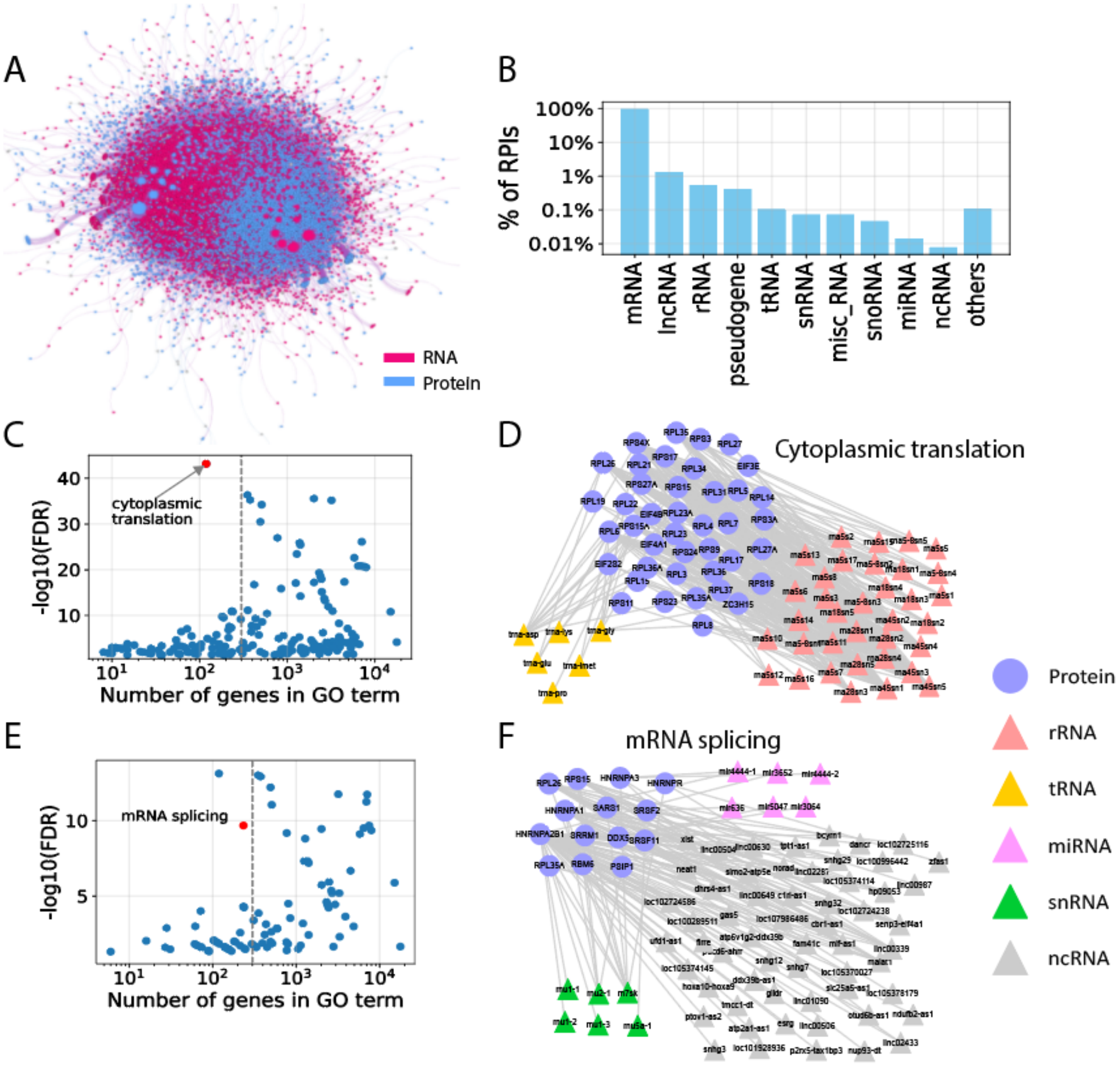


Figure 3.3: PRIM v.1.0

(A) Entire PRIM v.1.0 network with proteins and RNA as nodes and RPIs as edges. RNAs are colored in red, and proteins are colored in blue.

(B) Distribution of RNA-protein interactions of different RNA species in PRIM v.1.0

(C) Number of genes (x axis) of each GO term (dot) versus the enrichment level (y axis) of this GO term for RPIs that involve rRNAs and tRNAs in PRIM v.1.0.

(D) Cytoplasmic translation RPI network that involves rRNAs and tRNAs in PRIM v.1.0.

(E) Number of genes (x axis) of each GO term (dot) versus the enrichment level (y axis) of this GO term for RPIs that involve snRNAs and miRNAs in PRIM v.1.0.

(F) mRNA splicing RPI network that involves snRNAs, miRNAs, and lncRNAs in PRIM v.1.0.

Figure 3.4: Enrichment of characterized RBPs in PRIM v.1.0

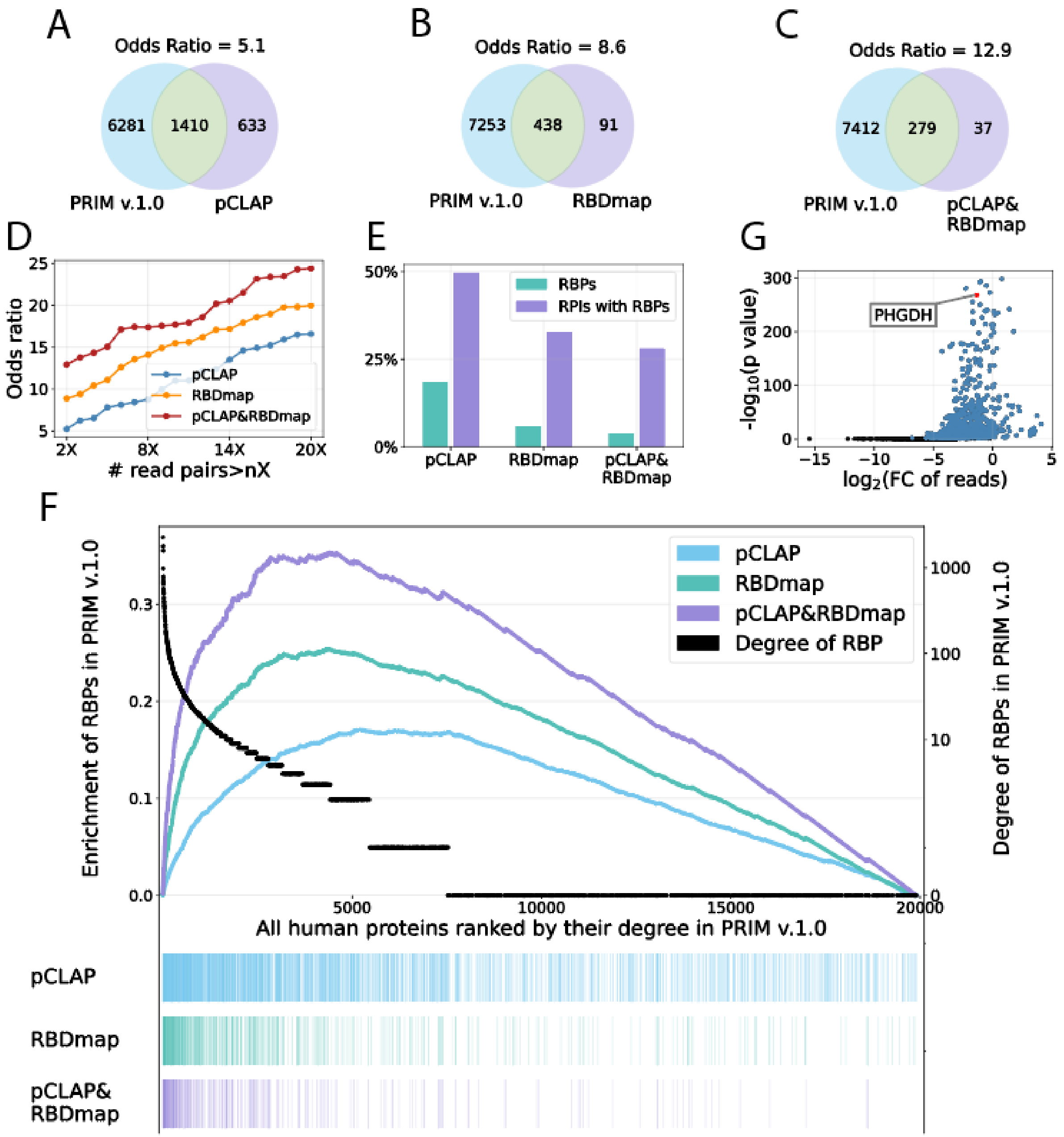
(A)-(C) Venn diagrams of comparing the RBPs captured by PRIM v.1.0 and the RBPs captured by pCLAP (A), by RBDmap (B), and by both pCLAP and RBDmap (C).

(D) The odds ratio (y axis) resulted from comparing the RBPs captured in PRIM v.1.0 with previously characterized RBPs with respect to nX (x axis).

(E) Percentages of previously characterized RBPs and percentages of RPIs with previously characterized RBPs in PRIM v.1.0.

(F) Enrichment (y axis, left) and degree in PRIM v.1.0 (y axis right, black) of previously characterized RBPs detected by pCLAP (blue), RBDmap (green), and both pCLAP and RBDmap (purple) with regards to all human proteins ranked by their degrees in PRIM v.1.0 (x axis). The vertical colored bar indicates the protein's presence in pCLAP (blue), RBDmap (green), and both pCLAP and RBDmap (purple).

(G) Semi-volcano plot of previously characterized RBPs in PRIM v.1.0. The colored dots are RBPs that are significantly enriched with RBD-aligned protein-end reads in PRIM v.1.0. Y axis represents the negative log BH-corrected p-values derived from one-sided binomial test. X axis represents the log fold change of the number of RBD-aligned reads over the number of the rest of the reads.


A
Odds Ratio = 5.1
6281
1410
633
PRIM v.1.0
pCLAP
B
Odds Ratio = 8.6
7253
438
91
PRIM v.1.0
RBDmap
C
Odds Ratio = 12.9
7412
279
37
PRIM v.1.0
pCLAP&
RBDmap
D
Odds ratio
25
20
15
10
5
2X
8X
14X
20X
# read pairs>nX
pCLAP
RBDmap
pCLAP&RBDmap
E
50%
25%
0%
RBPs
RPIs with RBPs
pCLAP
RBDmap
pCLAP&
RBDmap
G
-log10(p value)
300
200
100
0
PHGDH
−15
−10
−5
0
5
log2(FC of reads)
F
Enrichment of RBPs in PRIM v.1.0
0.3
0.2
0.1
0.0
pCLAP
RBDmap
pCLAP&RBDmap
Degree of RBP
1000
100
10
0
Degree of RBPs in PRIM v.1.0
5000
10000
15000
20000
All human proteins ranked by their degree in PRIM v.1.0
pCLAP
RBDmap
pCLAP&
RBDmap

Identification of RBD-bound RNA motifs from PRIM v.1.0

In PRIM v.1.0, 319,951 RPI read pairs have their protein-end aligned to RBDs captured by either RBDmap or pCLAP. For these RBD-aligned read pairs, we asked if we could identify any RNA motifs from their RNA ends. In PRIM v.1.0, 106,636 read pairs have their protein-end mapped to RNA Recognition Motif (RRM), 27,689 read pairs mapped to S1 domain, 13,899 read pairs mapped to PseudoUridine synthase and Archaeosine transglycosylase (PUA),10,771 read pairs mapped to K Homology (KH), 3,930 read pairs mapped to DEAD domain and 1,211 read pairs mapped to Cold-shock Domain (CSD) (Mistry et al. 2021). We grouped the RNA-end reads by their protein-end aligned RBDs and identified 30 RNA motifs with a length of 10 nucleotides for each RBD using all the RNA-end reads from PRIM v.1.0 as the background (Heinz et al. 2010). We asked if the RNA motifs derived from PRIM-seq are consistent with any of the previously characterized RBD-bound RNA motifs and if these RNA motifs can form any potential RNA secondary structures. We obtained experimentally derived RNA motif consensus sequences from ATrRACT (Giudice et al. 2016) and compared them with the PRIM-seq derived RNA motifs. As a result, 15 PRIM-seq derived RNA motifs overlapped with 55 known consensus sequences for RRM (Figure 3.5A, B), 12 PRIM-seq derived motifs overlapped with 19 known consensus sequences for KH (Figure 3.5D, E) and 5 PRIM-seq derived motifs overlapped with 4 known consensus sequences for CSD (Figure 3.5G, H). We applied RNAstructure to these consensus-sequence-overlapped RNA motifs to look for potential RNA secondary structures. We found 1 RRM-binding RNA motif (GACCAGTGGT) (Figure 3.5C) and 1 KH-binding RNA motif (GCGCAAGCGC) (Figure 3.5F) capable of forming high confidence RNA secondary structures (Reuter and Mathews 2010). Altogether, PRIM v.1.0 contains RBD-binding RNA motifs that are supported by previous literatures and can form possible RNA secondary structures. This implied a strong potential of applying PRIM-seq to study the specificity of RNA second structures on different RBDs in the future.

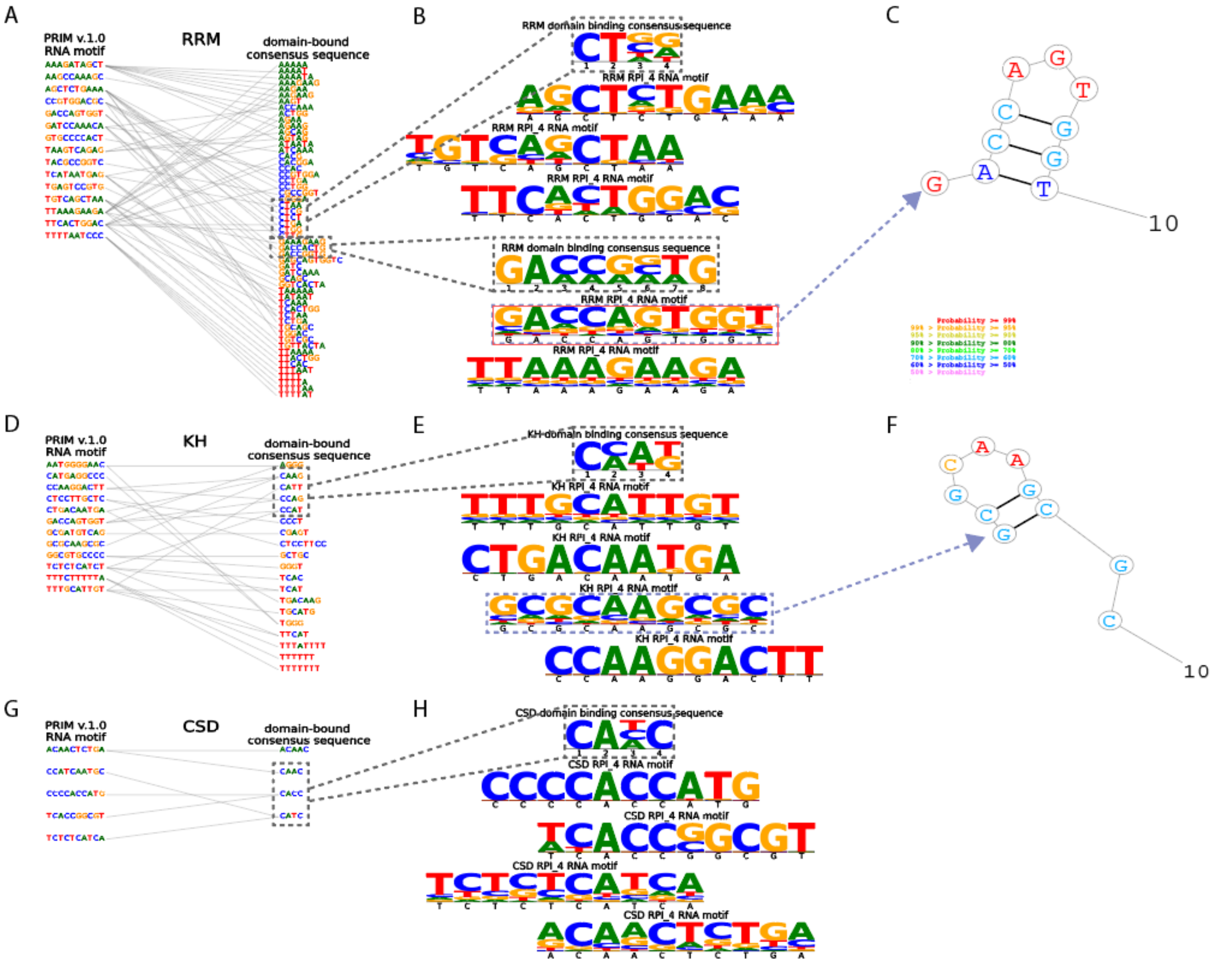


Figure 3.5: RBD-bound RNA motifs from PRIM v.1.0

(A) 15 PRIM v.1.0 RNA motifs overlapping with 55 RNA consensus sequences that are known to bind to RRM.

(B) Examples of PRIM v.1.0 RNA motifs aligning with motifs derived from subsets of known RNA consensus sequences for RRM.

(C) The predicted RNA secondary structure that can be formed by one RRM-bound PRIM v.1.0 RNA motif (GACCAGTGGT).

(D) 12 PRIM v.1.0 RNA motifs overlapping with the 19 RNA consensus sequences that are known to bind to KH.

E) Examples of PRIM v.1.0 RNA motifs aligning with motifs derived from subsets of known consensus sequences for KH.

F) The predicted RNA secondary structure that can be formed by one KH-bound PRIM v.1.0 RNA motif (GCGCAAGCGC).

G) 5 PRIM v.1.0 RNA motifs overlapping with the 4 RNA consensus sequences that are known to bind to CSD.

H) Examples of PRIM v.1.0 RNA motifs aligning with motifs derived from subsets of known consensus sequences for CSD.

3.4.3 Validation of PHGDH as an RNA-binding protein

Although PHGDH has been captured by both pCLAP and RBDmap, it is not widely characterized as an RBP by other literature curated RBP databases like RBPDB (Cook et al. 2011), starBASE (Li et al. 2014), hRBPome (Ghosh, Murugavel, and Sowdhamini 2018), ATtRACT (Giudice et al. 2016) and RBP2GO (Caudron-Herger et al. 2021). Yet, PRIM-seq identified 728 RPIs that involve PHGDH as the interacted protein (Figure 3.6A) and PHGDH is one of the proteins that is mostly enriched with RBD-aligned protein-end reads. We obtained 4 RBDs from pCLAP and RBDmap and compared their genomic positions with those of the protein-end reads that were mapped to PHGDH in PRIM v.1.0. 4,146 out of 14,672 (28.26%) reads were aligned to PHGDH's RBDs (BH corrected p-value<0.05, one-sided binomial test) (Figure 3.6B). Within the 728 PRIM-seq identified PHGDH-RNA interactions, we found 132 interactions have more protein-end reads mapped to RBDs than to non-RBD regions on PHGHD in PRIM v.1.0. GO enrichment analysis was applied to the 132 involved RNAs and found 'DNA conformation change (GO:0071103)' and 'ribosome biogenesis (GO:0007046)' are the two most enriched GO terms containing no more than 300 genes (Figure 3.6C).

To further validate PHGDH as an RNA binding protein, we carried out one RIP-seq experiment on PHGDH with HEK293T cells (Zhao et al. 2010). RIP-seq detected 801 RNAs that bind PHGDH using IgG as the background (CPM>100, fold change>2). Among them, 113 RNAs were also identified by PRIM-seq as PHGDH's binding target (odds ratio=9.9, p-value$<10^{-36}$, Fisher's exact test) (Figure 3.6D). This suggests the enrichment of RIP-seq identified PHGDH-bound RNAs in PRIM v.1.0 (Figure 3.6E). We also selected several previously uncharacterized RNA-PHGDH interactions from PRIM v.1.0 for further experimental validation with RIP-qPCR (Marmisolle, García, and Reyes 2018). The selected interactions are PTMA-PHGDH, HNRNPA2B1-PHGDH, BECN1-PHGDH, BCLAF1-PHGDH, and ATF4-PHGDH. PTMA and HNRNPA2B1 are two of the most abundant RNAs that interact with PHGDH in PRIM v.1.0. BECN1, BCLAF1, and ATF4 are genes that play important roles in the regulation of the cell apoptotic process. For all the RNAs tested, we detected significantly more signals for their interactions with PHGDH than with the IgG background ($p\text{-value}<0.05$, student's t-test) (Figure 3.6F). In other words, 5 of the 5 previously uncharacterized RNA-PHGDH interactions have been validated by RIP-qPCR. Altogether, PHGDH's role as an RNA binding protein was confirmed by RIP-seq and RIP-qPCR.

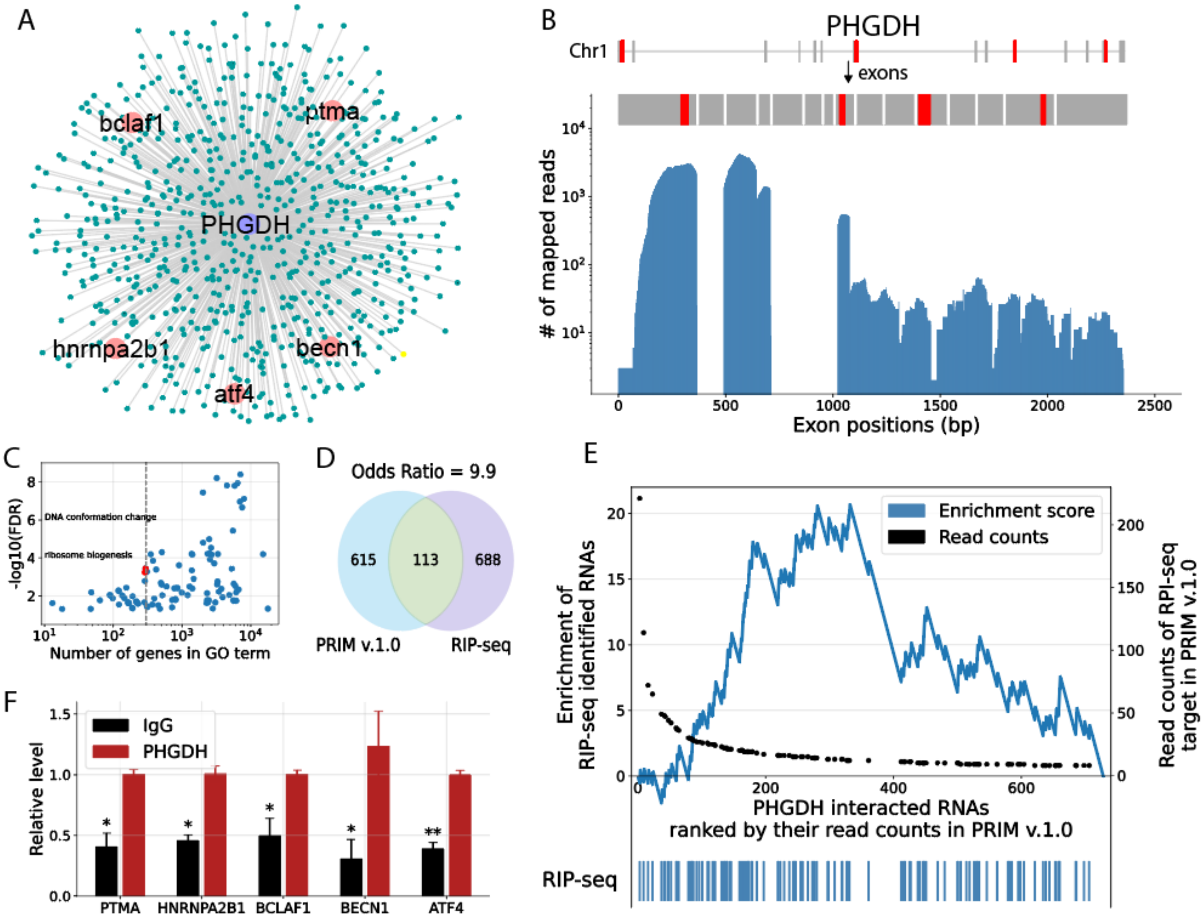


Figure 3.6: PHGDH as an RNA-binding protein

(A) 728 RPIs in PRIM v.1.0 that involve PHGDH as the protein. The 5 RPIs tested by RIP-qPCR are labeled.

(B) Protein-end reads (blue) from PHGDH's RPIs in PRIM v.1.0 aligned to the exons (grey) of PHGDH. RNA-binding domains on PHGDH are colored in red.

(C) Number of genes (x axis) of each GO term (dot) versus the enrichment level (y axis) of this GO term in the 132 RNAs that interact with PHGDH and have more protein-end reads mapped to RBDs than to non-RBD regions in PRIM v.1.0.

(D) Venn diagrams of comparing PHGDH-interacted RNAs captured by PRIM v.1.0 and by RIP-seq.

(E) Enrichment (y axis, left) and degree in PRIM v.1.0 (y axis right, black) of RIP-seq identified PHGDH-interacted RNAs with regards to all the PRIM-seq identified PHGDH-interacted RNAs ranked by the number of read pairs mapped to the interactions in PRIM v.1.0 (x axis). The vertical colored bar indicates the RNAs' presence in RIP-seq identified targets.

(F) Bar plots comparing the signal level of PHGDH's interactions with 5 RNAs (PTMA, HNRNPA2B1, BCLAF1, BECN1, and ATF4) captured by RIP-qPCR against IgG as the background. ∗: $p < 0.05$, student's t-test.

3.5 Discussion

PRIM-seq provides a time-effective approach to identify RNA-binding proteins and to map RNA-protein interactions at the transcriptome scale in a single experiment. It does not require specialized antibodies to capture certain proteins or RNAs. Thus, PRIM-seq may be a useful profiling tool to assist users to study RNA-protein interactome and its associated fields.

PRIM-seq provides evidence to support PHGDH as an RNA-binding protein. It adds several hundred interacting RNAs to PHGDH with 5 of them validated in another orthogonal experiment. The GO enrichment analysis on PHGDH's interacting RNAs implies PHGDH's role in regulating cell death as dysfunctions in two of the most enriched terms, ribosome biogenesis and DNA conformation change, can both induce the apoptosis process (Stedman et al. 2015; De Zio, Cianfanelli, and Cecconi 2013) The RNA binding peaks on PHGDH, as revealed by RPIM-seq, cover both known RNA-binding domains and unknown regions. We anticipate further experiments to confirm PHGDH's function as an RBP as well as to characterize new RBDs on PHGDH. A similar analysis may also be extended to study other uncharacterized RNA binding proteins in PRIM v.1.0.

Regarding the limitation of PRIM-seq, first, we did not experimentally label either the protein-end or the RNA-end of the interaction pairs in the current protocol. This may yield valid chimeric read pairs that are false positives. Second, we did not consider the cases where RNA interacts with its own translated protein. Besides, PRIM-seq is not designed to specifically capture the RNA sequences that attach to the RBDs. This sets limitations on the study of RBD-bound RNA motifs. Since the fragmentation of the chimeric sequences occurred randomly during library preparation, in theory, the RNA sequences with RBD-bound motifs share the same probability of being captured by PRIM-seq as any other RNA sequences. This is different from the enrichment of RBDs from the protein-end. Because the protein can be partially translated from the RNA sequence that contains the RBD information so that they are more likely to be bound by free RNA and thus to be detected by RPIM-seq. We anticipate an improvement in the experiment protocol where the specific RNA sequence that binds to the RBDs can be tagged and sequenced to enable a better study on RNA motifs.

3.6 Supplementary information

Supplementary figures

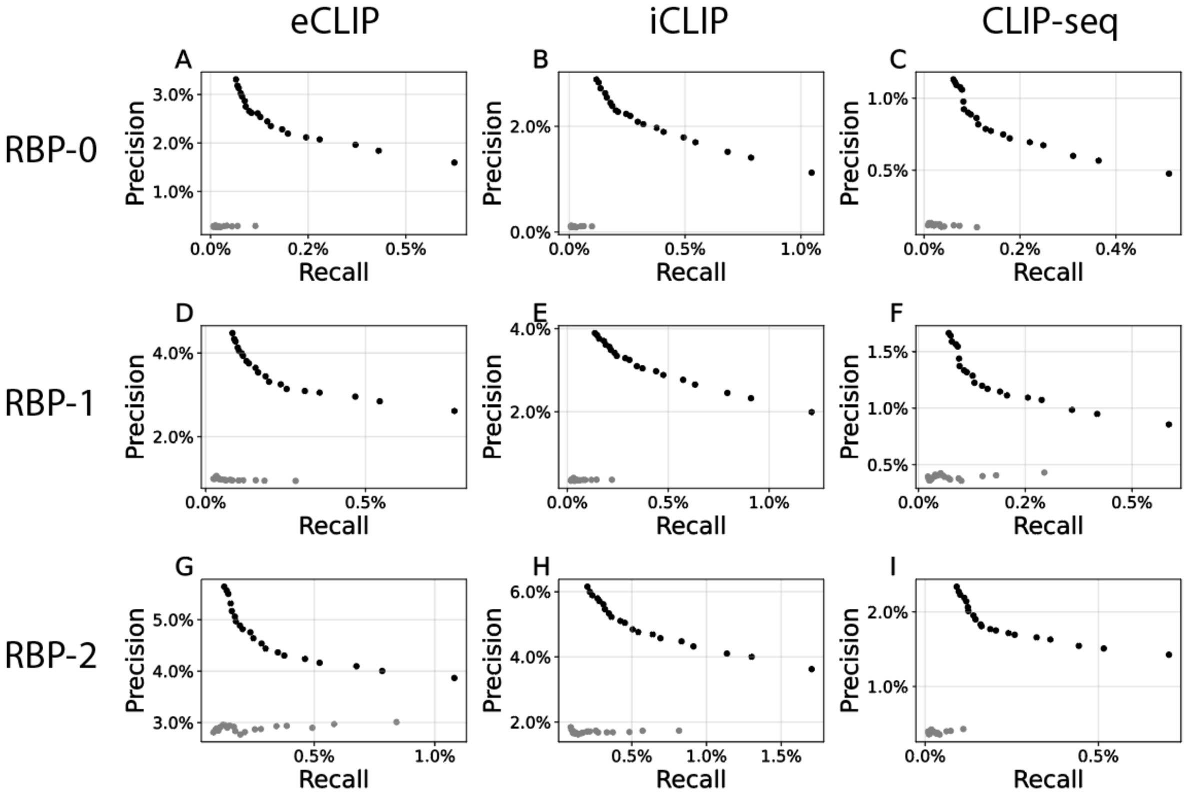


Figure S3.1: Precisions and recalls of PRIM-seq identified RPIs
PRIM-seq derived RPIs from HEK1 under the protein search space defined by RBP-0 (A-C), RBP-1 (D-F) and RBP-2 (G-I) were compared to three types of known RPIs that were retrieved from RNAInter, including all the RPIs that were identified by eCLIP, iCLIP, and CLIP-seq (columns). The precisions of recalls of the RPIs identified from PRIM-seq's permutation dataset are marked in grey dots. The permutations were based on only the genes involved in PRIM-seq detected RPIs under the respective protein search space.

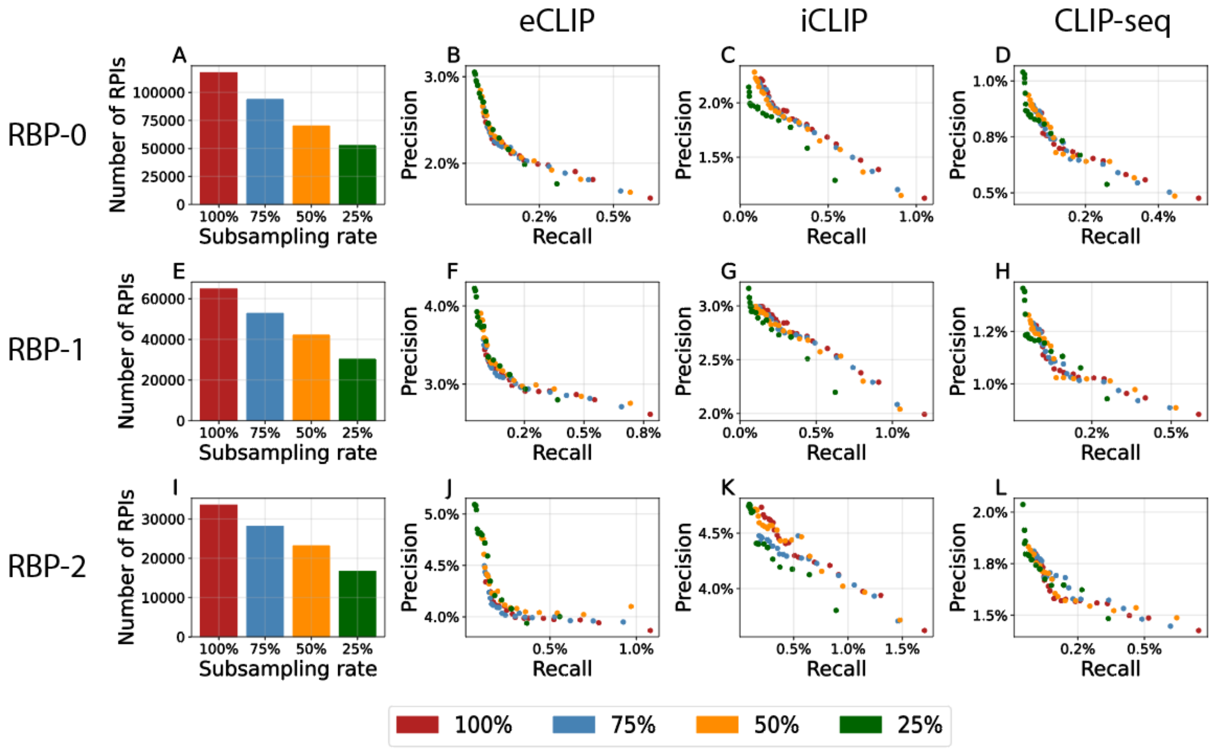


Figure S3.2: PRIM-seq identified RPIs with subsampling

Number of PRIM-seq derived RPIs from HEK1 under the protein search space defined by RBP-0 (A), RBP-1 (E) and RBP-2 (I) at a subsampling rate of 100% (red), 75% (blue), 50% (orange) and 25% (green). Precision-recall curves of RPIs identified from HEK1 under the protein search space defined by RBP-0 (B-D), RBP-1 (F-H), and RBP-2 (J-L) at a subsampling rate of 100% (red), 75% (blue), 50% (orange) and 25% (green), compared to three types of PPIs that are derived from other experimental methods, including all the RNAInter RPIs that are detected by eCLIP, iCLIP, and CLIP-seq under the respective protein search space.

Supplementary tables

Table S3.1: Summary of PRIM-seq library

The total number of read pairs, the number of read pairs mapped to genes and the number of non-duplicate valid chimeric read pairs with one end sensely mapped to a gene and one end antisensely mapped to a protein-coding gene were listed in the last three columns.

Library ID	**HEK1**
Cell line	HEK293T
Number of read pairs	305,357,086
# of mapped read pairs	158,352,224
# of non-duplicate valid chimeric read pairs	5,932,641

Table S3.2: The datasets used

The datasets used in this work, including PRIM-seq derived RPIs, RPIs curated in RNAInter, and RNA-binding proteins (RBPs) with captured RNA-binding domains (RBDs).

Name	**Description**	**# RPIs**	**# RNAs**	**# proteins**
PRIM v.1.0	The RPIs derived from PRIM-seq library of HEK1	117,516	8,440	7,691
RNAInter	All the experimentally-derived human RPIs in RNAInter, downloaded from http://www.rnainter.org/download/	1,342,821	33,674	13,538
eCLIP	Enhanced UV crosslinking and immunoprecipitation detected RPIs that are included in RNAInter	301,020	14,655	133
iCLIP	Individual-nucleotide resolution UV crosslinking and immunoprecipitation detected RPIs that are included in RNAInter	126,345	15,669	37
CLIP-seq	Crosslinking immunoprecipitation associated to high-throughput sequencing detected RPIs that are included in RNAInter	109,706	15,474	46
			#RBPs	**#RBDs**
pCLAP	Peptide crosslinking and affinity purification detected RBPs and RBDs, by Mullari, Lyon, Jensen, & Nielsen, 2017		2,043	4,751
RBDmap	RBDmap detected RBPs and RBDs, Castello, by Fischer et al. 2016		529	1,611

Table S3.3: PRIM v.1.0 framework

Estimated screening completeness, sampling sensitivity, assay sensitivity, precision for PRIM v.1.0, and estimated RNA-protein interactome size based on PRIM v.1.0. The background consists of random RPIs formed by permutating the match of RNAs and proteins in PRIM v.1.0.

	PRIM v.1.0	**Background**
Screening completeness	**42.18%**	N/A
Sampling sensitivity	**56.69%**	N/A
Assay sensitivity	**2.91%**	0.16%
Precision	**9.37%**	0.98%
Human RNA-protein interactome size	**$9.5*10^6$**	N/A

3.7 Materials and methods

SMART-display

The SMART-display library from HEK 293T cells was prepared following the same method as described in section 1.7.

Purification and Immobilization of Display Products

75 uLs of Dynabeads™ MyOne™ Streptavidin T1 (Thermo Fisher Scientific, 65601) were prepared by washing twice in an equivalent volume of 1x PBS pH 7.4 (Thermo Fisher Scientific, 70011044). The IVT reaction was added to the suspended beads in 1.8 mLs of 1x PBS pH 7.4 (Thermo Fisher Scientific, 70011044) with 0.1% Triton™ X-100 (Sigma-Aldrich, T8787-50ML) and incubated for 1 hour with rotation at room temperature. D-Biotin (Ivitrogen, B20656) was added to 2.25 uM and incubated at room temperature for 10 minutes with rotation. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100 (Sigma-Aldrich, T8787-50ML).

DNA Synthesis

50 uLs of first strand reaction was mixed per sample containing 500 U of SuperScript II Reverse Transcriptase (Thermo Scientific, 18064014), 1x SuperScript II FS Buffer, 5 mM DTT, 1 uM dNTP mix (NEB, N0447S), 1 M Betaine (Sigma-Aldrich, 61962), 6 mM MgCl2, 500 pmol of End Capture TSO (5' /5dSp/AGT AAA GGA GAC CTC AGC TTC ACT GGA rGrGrG 3'), and 40 U of SUPERase• In™ RNase Inhibitor. The mix was added to the beads and incubated at 42°C for 50 minutes with agitation, and then cycled 10 times at 50°C for 2 minutes followed by 42°C for 2 minutes. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

RNA-protein Interaction

The bead bound display proteins were suspended in 200 uLs RNA Binding Buffer (10 mM HEPES (Fisher Scientific, BP299100), 50 mM KCl , 4 mM MgCl2, 4 mM DTT, 0.2 mM EDTA, 7.6% glycerol (Invitrogen, 15514011)). 2 ugs of total RNA, prepared as described above, was added the display protein samples with the following conventions: positive reaction: no treatment display proteins and linker ligated total RNA, no linker Control: no treatment display proteins and no linker total RNA, and no bait control: Proteinase K digested display proteins and linker ligated total RNA. The mixtures were incubated at room temperature with rotation for 1 hour. 800 uLs of Binding Buffer was added to each reaction to bring the volume to 1 mL, and they were rotated for an additional 10 minutes at room temperature.

Crosslinking and Washing

Crosslinking was performed at room temperature for 10 minutes at a final concentration of 1% formaldehyde (Thermo Fisher Scientific, 28906). The reaction was quenched with 125 mM glycine (Sigma-Aldrich, 67419-1ML-F) with rotation for 5 minutes. The beads were washed 2 times each for 5 minutes with: 500 uLs Urea wash buffer [50 mM Tris-Cl pH 7.5, 1% NP-40, 0.1% SDS, mM EDTA, 1 M NaCl, 4 M Urea (Sigma-Aldrich, U5378-1KG)], Low Salt wash buffer [0.1% SDS (Invitrogen, AM9820), 0.1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCL ph 8 (Invitrogen, 15568025), 150 mM NaCl], and 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Second Strand Synthesis (Display Complex)

100 uLs of first strand reaction was mixed per sample containing 20 U DNA Polymerase I (NEB, M0209S), 1x NEBuffer 2, 2.4 mM DTT, and 0.25 mM dNTP mix. The mix was added to the beads and incubated at 37°C for 30 minutes with agitation. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Restriction Digestion

All samples were digested with 10 U of BbvCI (NEB, R0601S) in 1x CutSmart Buffer at 500 uLs. The digestion was incubated at 37°C for 1 hour with agitation. All samples were then washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Proximity Ligation

Proximity ligation was performed with 20,000 U of T4 DNA Ligase in 1 mL of 1x T4 DNA Ligase Buffer (NEB, M0202M). The reaction was incubated with constant rotation for 30 minutes at room temperature. The enzyme was inactivated before the beads were gathered by heating to 65°C for 10 minutes. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Protein Digestion and Reverse Crosslinking

The streptavidin beads were suspended in 200 uLs TAE buffer (Invitrogen™, AM9869) with 0.8 U of Proteinase K (NEB, P8107S) and incubated at 70°C for 30 minutes. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

cDNA Synthesis (RNA End)

50 uLs of first strand reaction was mixed per sample containing 500 U of SuperScript II Reverse Transcriptase, 1x SuperScript II FS Buffer, 5 mM DTT, 1 uM dNTP mix. The mix was added to the beads and incubated at 42°C for 50 minutes with agitation. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Second Strand Synthesis (RNA End)

100 uLs of first strand reaction was mixed per sample containing 20 U DNA Polymerase I, 1 U RNase H (NEB, M0297S), 1x NEBuffer 2, 2.4 mM DTT, and 0.25 mM dNTP mix. The mix was added to the beads and incubated at 37°C for 30 minutes with agitation. The beads were washed 2 times for 5 minutes with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100.

Sequencing Library Generation and Sequencing

The DNA was released from the beads with the NEBNext® Ultra™ II FS DNA Module (NEB, E7810S) using twice the reaction volume and a fragmentation time of 5 minutes. The end repair step was not performed. Libraries were then generated with the NxSeq® UltraLow DNA Library Kit (Lucigen, 15012-1) up to the final AMPure XP Bead purification before amplification. Each sample was eluted in 50 uLs Nuclease-free water, and added to 10 uLs of Dynabeads™ MyOne™ Streptavidin T1beads suspended in 50 uLs 1x PBS pH 7.4 with 0.1% Triton X-100. The selection was performed at room temperature for 1 hour. Beads were washed 2 times with 500 uLs Low Salt buffer (0.1% SDS, 0.1% Triton™ X-100, 2 mM EDTA, 20 mM Tris-HCI Buffer, pH 8, 150 mM NaCl), 2 times with 500 uLs 1x B&W Buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1M NaCl), and 2 times with 500 uLs 1x PBS pH 7.4 with 0.1% Triton™ X-100. Library amplification was then performed with the NxSeq® UltraLow DNA Library Kit as directed. Each library was paired end sequenced for 100 cycles on each end on an Illumina HiSeq 4000 or NovaSeq 6000.

RNA immunoprecipitation (RIP)

HEK293T cells were harvested from two 10 cm dishes and lysed in the lysis buffer (50 mM Tris-HCl, pH 7.5 (Invitrogen™, 15567027), 100 mM NaCl (Thermo Fisher Scientific, AM9759), 1% Triton X-100 (Sigma-Aldrich, T8787-50ML), 0.1% SDS (Invitrogen™, AM9820), 0.5% Sodium Deoxycholate (Sigma-Aldrich, 30970-25G), and a protease inhibitor cocktail (Roche, 4693159001)) together with 200 U of the RNase inhibitor, RNaseOUT™ (40 U/µL, Invitrogen™, 10777019) on ice for 30 minutes with occasional mixing. Cell lysates were centrifuged at 14k rpm for 20 minutes. Protein A Dynabeads (Invitrogen™, 10001D) were prepared by incubating 5 µg of Rabbit anti-PHGDH IgG (Proteintech, 14719-1-AP) or 5 µg of Rabbit IgG isotype control (Abcam, AB37415) with the pre-washed beads at 4°C for 2-3 hours. RNA immunoprecipitation (RIP) was conducted by an incubation of the supernatants of the cell lysates with pre-equilibrated Protein-A Dynabeads at 4˚C overnight. Beads were sequentially washed twice with high salt buffer (50 mM Tris-HCl, pH 7.5 (Invitrogen™, 15567027), 1 M NaCl (Thermo Fisher Scientific, AM9759), 1 mM EDTA (Research Products International, E14100-50.0), 1% Triton X-100 (Sigma-Aldrich, T8787-50ML), 0.1% SDS (Invitrogen™, AM9820), and 0.5% Sodium Deoxycholate (Sigma-Aldrich, 30970-25G)) and wash buffer (20 mM Tris-HCl, pH 7.5 (Invitrogen™, 15567027), 10 mM $MgCl_2$ (Invitrogen™, AM9530G), and 0.2% Tween-20 (Sigma-Aldrich, P9416-100ML)). Complexes were released from beads by incubation with 10% PK (Thermo Scientific™, EO0491) in PK buffer (50 mM Tris-HCl, pH 7.5 (Invitrogen™, 15567027) and

10 mM $MgCl_2$ (Invitrogen™, AM9530G)) at 50°C for 40 minutes. The supernatant was collected, and RNA was extracted by TRIzol Reagent (Invitrogen™, 15596026) followed by chloroform (Sigma-Aldrich, C0549-1QT). The mixtures were centrifuged at 14k rpm for 15 minutes at 4°C, the upper layer was collected. RNA was precipitated by the addition of 3 µL of glycogen (Thermo Scientific™, R0561), 50% of 2-propanol (Sigma-Aldrich, I9516-500ML), and 10% of 3 M sodium acetate, pH 5.5 (Invitrogen™, AM9740) with an incubation at -80°C overnight. The RNA was then pelleted by centrifugation at 14k rpm for 30 minutes at 4°C, washed with 1 mL of 75% ethanol (Sigma-Aldrich, 493546), and air-dried. The RNA was suspended in 20 µL of UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen™, 10977015).

Library prepareation & Sequencing

A library was prepared by cDNA synthesis, amplification, fragmentation, and adaptor ligation using NEBNext® Low Input RNA Library Prep Kit (NRB, E6420) and sequenced 150 base pairs from each end on an Illumina MiniSeq.

RT-qPCR

To perform a reverse transcription reaction, 9 µL of the eluted RNA, 50 ng random hexamers (50 ng/µL, Invitrogen™, 2039360), and 10 nmol dNTP mix (NEB, N0447S) were mixed. This reaction was brought to 65°C for 5 minutes to denature RNA and then quickly chilled on ice. The following reaction was performed by sequentially adding 2 µL of 5X First-Strand buffer, 0.2 µmol of DTT, 40 U of RNaseOUT™ (40 U/µL, Invitrogen™, 10777019). The reaction was incubated at 25 °C for 2 minutes for annealing. cDNA was synthesized from the eluted RNA by adding 200 U of SuperScript™ II Reverse Transcriptase (Thermo Scientific™, 18064014) into the mixture to a 20 µL final volume and incubating at 25 °C for

10 minutes, 42°C for 50 minutes, and 70 °C for 15 minutes to terminate the reaction. The reaction was chilled on ice.

Three replicates of 20 μL qPCR reaction containing 1x Power SYBR® Green PCR Master Mix (Thermo Fisher Scientific, 4367659) and 0.6 μM of each of the gene specific primers were prepared for the PHGDH IP sample and the IgG IP control to test 6 genes (1 housekeeping gene: GAPDH; 5 target genes: PTMA, HNRNPA2B1, BECN1, BCLAF1, and ATF4). The qPCR assay was run on a QuantStudio™ 3 Real-Time PCR System (Applied Biosystems™, A28567) with an initial denaturation of 95 °C for 5 minutes, 40 cycles of 95 °C for 10 seconds and 60 °C for 30 seconds. A melt curve was run to assess the purity of the qPCR products.

Quantification and Statistical Analysis

Processing PRIM-seq read pairs

The following data processing steps are implemented in the PRIMseqTools pipeline: https://github.com/Zhong-Lab-UCSD/PRIMseqTools. The sequencing reads were subjected to Cutadapt 2.5(Martin 2011) to remove the 3' linker sequence and the 5' adapter sequence. The remaining read pairs were subsequently subjected to Fastp 0.20.0(Huang et al. 2018) to remove low-quality reads (average quality per base < Q20) and short reads (<20 bp). The remaining read pairs were subsequently mapped to RefSeq transcripts (O'Leary et al. 2016) (based on GRCh38.p13, NCBI Homo sapiens Annotation Release 109.20211119) using BWA-MEM 0.7.12-r1039 (Li 2013) with the default parameters. A read was regarded as mapped to a gene if this read was mapped to any of the Refseq transcripts of this gene. The read pairs where one end was sensely mapped a gene and the other end was anti-sensely mapped a different protein-coding gene were identified. Any duplicated chimeric read pairs were subsequently removed to obtain non-duplicate valid chimeric read pairs.

Test of association between a gene pair and the chimeric read pairs

A Chi-square test was carried out on every RNA-protein pair. The null hypothesis is that the mapping of one end of a chimeric read pair to an RNA is independent of the mapping of the other end of this chimeric read pair to a protein. The contingency table of this association test is given in Figure 3.2A. FDR computed from the Benjamini-Hochberg procedure was used to control for family-wise errors.

Downloading RNAInter data and its subsets

RPIs were downloaded as a zipped TXT file from RNA Interactome Database (RNAInter) at http://www.rnainter.org/raidMedia/download/Download_data_RR.tar.gz. eCLIP, iCLIP and CLIP-seq RPIs were identified by the label of 'Homo sapiens' in both 'Species1' and 'Specie2' columns and by the corresponding method labels in the 'weak' column of the downloaded TXT file.

Downloading and processing RNA binding domains

Protein sequences of RBDs that were captured by either RBDmap (Castello et al. 2016) or pCLAP (Mullari et al. 2017) were obtained from the supplementary tables of their corresponding journal paper. Exonerate 2.4.0 (Slater and Birney 2005)Was applied to map the protein sequences back to the proteins and to get the genomic coordinates of the RBDs.

Identification of RNA motifs

The 'homer2 denovo' function of HOMER (Heinz et al. 2010) was applied to identify differential RNA motifs using all the RNA-end reads from PRIM-seq identified RPIs in PRIM v.1.0 as the background.

Downloading RNA consensus sequences

RNA consensus sequences that were known to bind specific RBDs were downloaded as a TXT file from a database of RNA binding proteins and associated motifs (ATtRACT) (Giudice et al. 2016) at https://attract.cnic.es/download with ‘Organism’ specified as ‘Homo sapiens’.

Prediction of RNA secondary structures

RNAstructure 1.0 (Reuter and Mathews 2010) was applied to PRIM-seq identified RNA motifs to detect potential RNA secondary structures.

GO term analysis

GO term enrichment analysis (Ashburner et al. 2000) was based on hypergeometric tests between the genes annotated by every GO term and the PROPER v1.0 nodes. FDR computed from the Benjamini-Hochberg procedure was used to control for family-wise errors. The entire PRIM v1.0 was plotted with Gephi 0.9.2 (Bastian, Heymann, and Jacomy 2009). All other network figures were plotted with Cytoscape (Shannon et al. 2003).

Calculating screening completeness, sampling sensitivity, assay sensitivity, precision, and protein interactome size for PRIM v1.0

Screening completeness, sampling sensitivity, assay sensitivity, precision, and protein interactome size were defined by Yu et al. (Yu et al. 2008) and Venkatesan et al. (Venkatesan et al. 2009). We calculated these metrics for PROPER v1.0 based on the methods described by Venkatesan et al. (Venkatesan et al. 2009) and the following positive reference set (PRS), random reference set (RRS), and orthogonal validation sets.

Positive reference set (PRS)

The RPIs from RNAInter that have both RNA and protein under the search space of

PRIM v.1.0 and were detected by strong detection methods ('strong' column not equal to 'N/A' in the downloaded file). These 7,207 RPIs are used as our PRS.

Random reference set (RRS)

Following Venkatesan et al. (Venkatesan et al. 2009), RRS was randomly sampled from PRIM v.1.0's search space outside the PRS to contain the same number of pseudo-RPIs as PRS.

Orthogonal validation assay

PAR-CLIP (Danan, Manickavel, and Hafner 2016) is used as the orthogonal validation assay. The PAR-CLIP data were retrieved from RNAInter based on the label of 'Homo sapiens' in both 'Species1' and 'Specie2' columns and on the corresponding method label in the 'weak' column of the downloaded TXT file.

Identifying PHGDH's binding RNAs from RIP-seq

fastp 0.20.0 was used to remove adaptors and raw reads with sequencing quality less than 15 from the sequencing library. bbmap 38.18 (Bushnell 2014) was used to remove human rRNA sequences. Clean reads were then mapped to human reference genome (based on GRCh38.p13, NCBI Homo sapiens Annotation Release 109.20211119) using STAR 2.7.9a (Dobin et al. 2013). Unaligned and secondary alignments were removed using samtools 1.8. Gene expression levels were counted using featureCounts 1.6.4 (Liao, Smyth, and Shi 2014) and CPM (counts per million) was calculated with customized code in R. PHGDH's binding RNAs were identified if the CPM of the gene is larger than 100, and if the CPM fold change of the gene over the IgG background is larger than 2.

3.8 Acknowledgements

Chapter 3, in full, is currently being prepared for submission for publication of the material. Qi. Zhijie; Johnson, Kara L.; Chen, Junchen; Zhao, Wenxin; Wen, Xingzhao; Xue, Shuanghong; Zhong, Sheng. The dissertation/thesis author was the primary investigator and author of this paper.

3.9 Reference

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium', *Nature genetics*, 25: 25-29.

Balcerak, A., A. Trebinska-Stryjewska, R. Konopinski, M. Wakula, and E. A. Grzybowska. 2019. 'RNA-protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity', *Open Biol*, 9: 190096.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." In *Third international AAAI conference on weblogs and social media*.

Benjamini, Yoav, and Yosef Hochberg. 1995. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, 57: 289-300.

Bushnell, Brian. 2014. 'BBMap: A Fast, Accurate, Splice-Aware Aligner'.

Castello, A., B. Fischer, C. K. Frese, R. Horos, A. M. Alleaume, S. Foehr, T. Curk, J. Krijgsveld, and M. W. Hentze. 2016. 'Comprehensive Identification of RNA-Binding Domains in Human Cells', *Mol Cell*, 63: 696-710.

Caudron-Herger, M., R. E. Jansen, E. Wassmer, and S. Diederichs. 2021. 'RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions', *Nucleic Acids Res*, 49: D425-D36.

Chu, C., K. Qu, F. L. Zhong, S. E. Artandi, and H. Y. Chang. 2011. 'Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions', *Mol Cell*, 44: 667-78.

Cook, K. B., H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes. 2011. 'RBPDB: a database of RNA-binding specificities', *Nucleic Acids Res*, 39: D301-8.

Danan, C., S. Manickavel, and M. Hafner. 2016. 'PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites', *Methods Mol Biol*, 1358: 153-73.

De Zio, Daniela, Valentina Cianfanelli, and Francesco Cecconi. 2013. 'New insights into the link between DNA damage and apoptosis', *Antioxidants & redox signaling*, 19: 559-71.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29: 15-21.

Ghosh, Pritha, Pavalam Murugavel, and Ramanathan Sowdhamini. 2018. 'hRBPome: a central repository of all known human RNA-binding proteins', *bioRxiv*: 269043.

Giudice, G., F. Sanchez-Cabo, C. Torroja, and E. Lara-Pezzi. 2016. 'ATtRACT-a database of RNA-binding proteins and associated motifs', *Database (Oxford)*, 2016.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. 2010. 'Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities', *Mol Cell*, 38: 576-89.

Huang, H. Y., Y. C. Lin, S. Cui, Y. Huang, Y. Tang, J. Xu, J. Bao, Y. Li, J. Wen, H. Zuo, W. Wang, J. Li, J. Ni, Y. Ruan, L. Li, Y. Chen, Y. Xie, Z. Zhu, X. Cai, X. Chen, L. Yao, Y. Chen, Y. Luo, S. LuXu, M. Luo, C. M. Chiu, K. Ma, L. Zhu, G. J. Cheng, C. Bai, Y. C. Chiang, L. Wang, F. Wei, T. Y. Lee, and H. D. Huang. 2022. 'miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions', *Nucleic Acids Res*, 50: D222-D30.

Huang, R., M. Han, L. Meng, and X. Chen. 2018. 'Capture and Identification of RNA-binding Proteins by Using Click Chemistry-assisted RNA-interactome Capture (CARIC) Strategy', *J Vis Exp*.

Huppertz, I., J. Attig, A. D'Ambrogio, L. E. Easton, C. R. Sibley, Y. Sugimoto, M. Tajnik, J. Konig, and J. Ule. 2014. 'iCLIP: protein-RNA interactions at nucleotide resolution', *Methods*, 65: 274-87.

Johnson, K. L., Z. Qi, Z. Yan, X. Wen, T. C. Nguyen, K. Zaleta-Rivera, C. J. Chen, X. Fan, K. Sriram, X. Wan, Z. B. Chen, and S. Zhong. 2021. 'Revealing protein-protein interactions at the transcriptome scale by sequencing', *Mol Cell*, 81: 3877.

Kang, J., Q. Tang, J. He, L. Li, N. Yang, S. Yu, M. Wang, Y. Zhang, J. Lin, T. Cui, Y. Hu, P. Tan, J. Cheng, H. Zheng, D. Wang, X. Su, W. Chen, and Y. Huang. 2022. 'RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility', *Nucleic Acids Res*, 50: D326-D32.

Li, Heng. 2013. 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', *arXiv preprint arXiv:1303.3997*.

Li, J. H., S. Liu, H. Zhou, L. H. Qu, and J. H. Yang. 2014. 'starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data', *Nucleic Acids Res*, 42: D92-7.

Liao, Y., G. K. Smyth, and W. Shi. 2014. 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30: 923-30.

Marmisolle, F. E., M. L. García, and C. A. Reyes. 2018. 'RNA-binding protein immunoprecipitation as a tool to investigate plant miRNA processing interference by regulatory proteins of diverse origin', *Plant Methods*, 14: 9.

Martin, Marcel. 2011. 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *2011*, 17: 3.

McHugh, C. A., and M. Guttman. 2018. 'RAP-MS: A Method to Identify Proteins that Interact Directly with a Specific RNA Molecule in Cells', *Methods Mol Biol*, 1649: 473-88.

McMahon, A. C., R. Rahman, H. Jin, J. L. Shen, A. Fieldsend, W. Luo, and M. Rosbash. 2016. 'TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins', *Cell*, 165: 742-53.

Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. 2021. 'Pfam: The protein families database in 2021', *Nucleic Acids Res*, 49: D412-D19.

Mullari, M., D. Lyon, L. J. Jensen, and M. L. Nielsen. 2017. 'Specifying RNA-Binding Regions in Proteins by Peptide Cross-Linking and Affinity Purification', *J Proteome Res*, 16: 2762-72.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F.

Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. 2016. 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Res*, 44: D733-45.

Ramanathan, M., K. Majzoub, D. S. Rao, P. H. Neela, B. J. Zarnegar, S. Mondal, J. G. Roth, H. Gai, J. R. Kovalski, Z. Siprashvili, T. D. Palmer, J. E. Carette, and P. A. Khavari. 2018. 'RNA-protein interaction detection in living cells', *Nat Methods*, 15: 207-12.

Reuter, J. S., and D. H. Mathews. 2010. 'RNAstructure: software for RNA secondary structure prediction and analysis', *BMC Bioinformatics*, 11: 129.

Saito, Takaya, and Marc Rehmsmeier. 2015. 'The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets', *PLoS One*, 10: e0118432.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Res*, 13: 2498-504.

Simon, M. D., C. I. Wang, P. V. Kharchenko, J. A. West, B. A. Chapman, A. A. Alekseyenko, M. L. Borowsky, M. I. Kuroda, and R. E. Kingston. 2011. 'The genomic binding sites of a noncoding RNA', *Proc Natl Acad Sci U S A*, 108: 20497-502.

Slater, G. S., and E. Birney. 2005. 'Automated generation of heuristics for biological sequence comparison', *BMC Bioinformatics*, 6: 31.

Stedman, A., S. Beck-Cormier, M. Le Bouteiller, A. Raveux, S. Vandormael-Pournin, S. Coqueran, V. Lejour, L. Jarzebowski, F. Toledo, S. Robine, and M. Cohen-Tannoudji. 2015. 'Ribosome biogenesis dysfunction leads to p53-mediated apoptosis and goblet cell differentiation of mouse intestinal stem/progenitor cells', *Cell Death Differ*, 22: 1865-76.

Stork, C., and S. Zheng. 2016. 'Genome-Wide Profiling of RNA-Protein Interactions Using CLIP-Seq', *Methods Mol Biol*, 1421: 137-51.

Syu, G. D., J. Dunn, and H. Zhu. 2020. 'Developments and Applications of Functional Protein Microarrays', *Mol Cell Proteomics*, 19: 916-27.

Tong, Z., Q. Cui, J. Wang, and Y. Zhou. 2019. 'TransmiR v2.0: an updated transcription factor-microRNA regulation database', *Nucleic Acids Res*, 47: D253-D58.

Tsai, B. P., X. Wang, L. Huang, and M. L. Waterman. 2011. 'Quantitative profiling of in vivo-assembled RNA-protein complexes using a novel integrated proteomic approach', *Mol Cell Proteomics*, 10: M110 007385.

Van Nostrand, E. L., G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo. 2016. 'Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)', *Nat Methods*, 13: 508-14.

Venkatesan, K., J. F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A. S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A. L. Barabasi, and M. Vidal. 2009. 'An empirical framework for binary interactome mapping', *Nat Methods*, 6: 83-90.

Wheeler, E. C., E. L. Van Nostrand, and G. W. Yeo. 2018. 'Advances and challenges in the detection of transcriptome-wide protein-RNA interactions', *Wiley Interdiscip Rev RNA*, 9.

Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. 2008. 'High-quality binary protein interaction map of the yeast interactome network', *Science*, 322: 104-10.

Zeng, F., T. Peritz, T. J. Kannanayakal, K. Kilk, E. Eiriksdottir, U. Langel, and J. Eberwine. 2006. 'A protocol for PAIR: PNA-assisted identification of RNA binding proteins in living cells', *Nat Protoc*, 1: 920-7.

Zhao, J., T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, K. Sarma, J. J. Song, R. E. Kingston, M. Borowsky, and J. T. Lee. 2010. 'Genome-wide identification of polycomb-associated RNAs by RIP-seq', *Mol Cell*, 40: 939-53.

Zhou, K. R., S. Liu, W. J. Sun, L. L. Zheng, H. Zhou, J. H. Yang, and L. H. Qu. 2017. 'ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data', *Nucleic Acids Res*, 45: D43-D50.

Zhu, Y., G. Xu, Y. T. Yang, Z. Xu, X. Chen, B. Shi, D. Xie, Z. J. Lu, and P. Wang. 2019. 'POSTAR2: deciphering the post-transcriptional regulatory logics', *Nucleic Acids Res*, 47: D203-D11.