UNIVERSITY OF CALIFORNIA SAN DIEGO

High Throughput Methods for Cell-Type Specific Elucidation of Protein Interactions

A dissertation submitted in partial satisfaction of the requirement for the degree
Doctor of Philosophy

in

Bioengineering with a Specialization in Multiscale Biology

by

Kara Lynn Johnson

Committee in charge:

Professor Sheng Zhong, Chair
Professor Ju Chen
Professor Stephanie Fraley
Professor Yingxiao Wang
Professor Kun Zhang

2020

The dissertation of Kara Lynn Johnson is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

This work is dedicated to the educators, mentors, and allies
who enriched and enabled the journey.

EPIGRAPH

Think and Wonder,

Wonder and Think.

Dr. Seuss

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

HEK PROPER-Seq Interaction Data: johnson_hek_network.csv

Jurkat PROPER-Seq Interaction Data: johnson_jurkat_network.csv

HUVEC PROPER-Seq Interaction Data: johnson_huvec_network.csv

Union PROPER-Seq Interaction Data: johnson_proper_network.csv

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Sheng Zhong for his support and guidance as the chair of my committee. I would also like to thank the members of the Zhong lab for their kind words and actions during difficult moments.

Particularly important in this journey were all of my family members and friends, and I would like to recognize them all for their unending support.

Chapters 4, 5, and 6, in part, are currently being prepared for submission for publication of the material. Johnson, Kara; Qi, Zhijie; Wen, Xingzhao; Chen, Chien-ju. The dissertation author was the primary investigator and author of this material.

VITA

2013 Bachelor of Science in Biological Systems Engineering,
University of California Davis

2013 Bachelor of Science in Genetics, University of California Davis

2017 Master of Science in Bioengineering, University of California San Diego

2019 Certificate of Project Management, UC San Diego Extension

2020 Doctor of Philosophy in Bioengineering with a Specialization in Multiscale Biology,
University of California San Diego

ABSTRACT OF THE DISSERTATION

High Throughput Methods for Cell-Type Specific Elucidation of Protein Interactions

by

Kara Lynn Johnson

Doctor of Philosophy in Bioengineering with a Specialization in Multiscale Biology

University of California San Diego, 2020

Professor Sheng Zhong, Chair

While it is widely understood that proteins are the functional tools of a cell, there are still no methods that efficiently illuminate the cell-wide networks of protein-protein and protein-RNA interactions through which proteins can act. Existing methods of detecting protein interactions primarily take one of two forms, they are either highly parallelized ‘one-by-one’ assays or several ‘one-by-many’

assays. However, these efforts are expensive and time intensive when attempting to address a cell-wide network.

This dissertation describes the development and validation of three technologies, a protein barcoding technology which generates a library of proteins labeled with specific nucleic acid sequences (SMART-Display), a high throughput proximity ligation technology that elucidates protein-protein interactions (PROPER-Seq), and a high throughput proximity ligation technology that elucidates protein-RNA interactions (PRIM). Leveraging these technologies, protein-protein and protein-RNA interactions can be assessed for a given cell type with an "all-vs-all" approach. The PRIM and PROPER-Seq workflows are characterized by a drastic increase in the number of interactions assayed in a single experiment relative to existing techniques, minimal labor, cost, and time, suitability for automation, and accessibility to any benchtop scientist, as they are devoid of the need for specialized technology or equipment.

SMART-Display produces highly complex protein libraries that closely reflect the mRNA population of the cell-type they are derived from, and which contain protein-specific DNA barcodes that can be used for proxy identification of the proteins themselves. This enables the application of PRIM and PROPER-Seq with libraries containing tens of thousands of uniquely labeled protein interactors. Precision and sensitivity analyses indicate that PROPER-Seq technique demonstrates a detection of interactions that is similar to, or better than, the levels demonstrated by gold-standard techniques such as yeast-2-hybrid. Several novel PROPER-Seq interactions were validated *in vivo*. Preliminary PRIM data indicates sufficient quality for precision and sensitivity analysis and both literature and *in vivo* validation.

1 Introduction

1.1 Protein Interactions in Living Systems

DNA is the basis of biological determination; however we know that the system by which the genetic code is executed is far more complex than a simple ordering of bases. In order to completely understand the impact of this expansive sequence on cells and cellular systems, we must consider the manner in which it is translated functionally. While it is widely understood that proteins are the tools of a cell, there is still no method that efficiently illuminates the cell-wide network of interactions through which proteins act.

Many protein anomalies result in erroneous protein interactions; interactions can have altered affinities, involve abnormal partners, or be lacking all together. As proteins are responsible for all the essential functions of the cell, such as replication, growth, and signaling, almost all cellular irregularities can be traced to a protein irregularity. Many of the health disorders and diseases that society currently faces can be characterized by such interaction variances. The significance of work focusing on increasing the ability to assay protein interactions is made clear just by enumerating the vast number of techniques developed since the advent of the gold standard protein interaction method, Yeast Two-Hybrid (Y2H), in 1989.

The seemingly infinite applications of high throughput interaction technologies has led to numerous approaches to their development. The majority of the resulting techniques take the form of a ‘one-by-many’ assay, where a single protein of interest is probed against a library of ‘prey’ molecules. Given the number of unique biomolecules found in a cell, it will take significant amounts of time to utilize these techniques to generate total network data. Only a handful of methods can successfully address library vs. library assays.

Within the *in vitro* protein-protein interaction (PPI) space, Dr. George Church's SMI-seq technique has demonstrated the greatest scale at 55 proteins by 200 peptide fragments[1]. His approach utilizes proteins barcoded with synthetic sequences, which are allowed to interact, are crosslinked, and then separated on a gel matrix for in situ sequencing. This approach, and the few others that have demonstrated a library-vs-library approach, are limited by the difficulty of preparing appropriate protein libraries for each system; and in some cases require specialized equipment, which increases their costs and restrains their application[2–5].

At the time of writing, we have not been able to identify any techniques that attempt to query protein-RNA interactions in a library-vs-library manner. However, there have been attempts to simultaneously identify all proteins bound to RNA, without consideration of the actual pairwise interactions occurring[6,7]. These studies determined a significantly larger portion of the proteome was binding RNA than previously thought, and indicated that RNA-binding proteins likely play a significant role in the cellular environment.

High-throughput protein interaction technologies offer insights for many cell-based fields. Developmental biology, stem-cell biology, and drug development are all noteworthy fields that would stand to benefit from such a technique. The seemingly infinite applications of a truly high-throughput technology has led to numerous approaches to its development. However, at this time, the field still lacks a method that is reliable and accessible to the average laboratory.

1.2 Statement of Purpose and Research Objectives

The purpose of the work described in this dissertation is to demonstrate a technology that allows for high throughput ranking of protein interactions with various molecular species. This work was divided into three primary objectives:

(1) Develop a technology that generates a distinguishable protein library reflecting a natural mRNA population.

(2) Develop a technology that leverages a distinguishable protein library to assay protein-protein interactions in a high-throughput manner.

(3) Develop a technology that leverages a distinguishable protein library to assay protein-RNA interactions in a high-throughput manner.

These objectives are pursued with the following major considerations mind:

(1) Maximize number of interactions assayed in a single experiment.

(2) Minimize labor and cost.

(3) Avoid specialized technology and equipment.

1.3 Justification of Approach

Because of the importance of proteins in living systems, there is a vast number of ways in which their interactions can be characterized. One of the original protein based methods, yeast two-hybrid (Y2H), remains the gold standard of the protein-protein interaction (PPI) community, and has also been used to study protein-DNA interactions. Yet, even high throughput Y2H variations are limited to binary testing, are labor intensive, expensive, and do not reflect the natural proteome of a cell type and state. Other techniques that can be performed in vivo, such as crosslinking and affinity purification followed by either mass spectrometry for protein interactors, or sequencing for nucleic acids, face similar cost and time limitations, demand high amounts of input sample, and often rely on complex gene constructs and specialized equipment.

A principal barrier in developing techniques that identify protein interactions in a high throughput manner is identifying the proteins themselves. Many common approaches, like western blotting, require individual antibodies that efficiently target the protein(s) of interest. As above, higher

throughput techniques like mass spectrometry are expensive, can lack precision, require relatively high input protein quantities, are not suitable for every protein, and often are incompatible with determining specific interacting partners. As an alternative to these classic experimental methods, where the protein itself is identified, several groups have already begun to leverage 'barcoding' technologies which utilize proxy identification to simplify the task. Much like license plate numbers are assigned to vehicles to make them easily discernable, in these technologies proteins are tagged with distinguishable barcodes. These barcodes can be used in turn to identify the protein, its interactions, or its binding properties via sequencing or hybridization.

The intricate networks in which proteins interact are excellent candidates for integration with the rapidly developing tools we have to manipulate nucleic acids and the powerful scale at which bioinformatics techniques allow us to work. There are significant and obvious benefit to the use of nucleic acids for the creation of these barcodes. Nucleic acids are readily synthesized, stable in cell-like environments, easily manipulated, and, with the advent of next generation sequencing, can be quickly queried in a high throughput manner. One technique of creating a protein with a nucleic acid barcode is called RNA or DNA display, depending on the molecule leveraged, and is a relatively new, but proven technique. In this method, the mRNA used to generate a protein is covalently bound to the protein during the translation process, via a puromycin molecule bound to the 3' end of mRNA. The puromycin molecule enters the A-site of the ribosome and is added to the growing poly-peptide chain. The mRNA is then often reverse transcribed to add a cDNA strand, as the cDNA/RNA complex is more stable than RNA alone, and therefor easier to handle.

The primary drawback to current cDNA display methods is their relatively low protein yields. However, because the reactions are performed *in vitro*, and current sequencing machines require relatively small amounts of sample, this limitation can be overcome by a careful titration of input materials. In this way the cost of larger reactions can be balanced with the need for a minimum display

protein quantity. A second concern is the influence of the DNA barcode on the binding properties of the protein, either by steric inhibition or DNA-protein interactions. A research group that works extensively with display proteins, under the direction of Dr. Naoto Nemoto, has shown that these effects are present but minimal[8]. Finally, any systems that requires protein production *in vitro* is limited by the population of protein that can successfully be translated. There are proteins that simply cannot be produced in this manner, and not always for known reasons. This tends to affect larger proteins more often than smaller ones.

Other methods of barcoding proteins are time and labor intensive, as they require each protein to be individually produced and labeled. One approach is to produce a protein that contains a tag, such as the HaloTag, and allow it to interact with a ligand-bound DNA sequence[1]. While the HaloTag is a covalent binding system, not all receptor-ligand interactions used for this purpose are, and in these systems there is a propensity for the tags on various proteins to be exchanged when mixed. This results in a population of protein which are incorrectly barcoded. A concern with all tags that require amino acid additions to the protein of interest is that these amino acids will either sterically inhibit protein interactions or will contribute to non-native interactions with other proteins or molecules in the system. With a large, enzymatically active receptor, such as the HaloTag, the likelihood of these effects being significant increases. The cDNA display method offers covalent binding of a unique barcode to a protein with a small, relative inert molecule (puromycin), and allows all proteins in the systems to be processed in a single interaction. For this reason, cDNA display is the method of choice in this work for generating nucleic-acid barcoded protein libraries. In this dissertation, we build and improve upon the existing cDNA display protein approaches in order to develop a quick, inexpensive, and high throughput method of constructing PPI networks.

While the barcodes contained on display proteins allow for identification of those proteins, the problem of identifying interactions is yet unaddressed. Other groups have approached this challenge

from a spatial perspective, as proteins that are interacting are in close proximity. These groups have separated protein libraries on a gel matrix, and sequenced *in situ* to identify binding partners. Our work also considers a spatial perspective, but avoids the necessity of specialized sequencing equipment by mirroring the pairwise binding of proteins as a representative pairwise ligation of the respective nucleic acid barcodes. Much like the previous published technique of proximity ligation[9], nucleic acid barcodes that are brought into proximity by the binding of their associated protein are ligated to create a chimeric fragment containing information from both protein binding partners. A population of these chimeric fragments can be generated and sequenced simultaneously to generate a network of interactions characterized by both a proteins binding partners and the number of binding events.

The principal that spatial proximity can be used as in indicator for interaction underlies many protein-interaction techniques, such as those that rely on cross-linking. Nucleic acid barcode dependent proximity ligation techniques have been demonstrated to accurately reflect protein interactions in small libraries[10], and RNA-RNA interactions in a cell wide manner[11]. However, physical proximity of proteins does not always indicate interactions or binding, particularly in high-density systems. By manipulating the environment that the protein interact within, the number of spatially proximate but non-interacting protein pairs can be reduced.

In order to execute ligation interaction technologies based on cDNA display methods, a specialized gene library must be generated. In this library, every protein coding gene must contain specific sequences that are required for the cDNA display process. One approach is to individually introduce the sequences to the gene for each desired protein product via the polymerase chain reaction (PCR) or other molecular technique. This approach allows for the generation of small, selective libraries for use in PPI assays.

Alternatively, a large library that represents the protein expression for a given cell type and state can be synthesized by utilizing mRNA extracted from a homogenous cell population. To replicate the

diversity of such a library using the method detailed above, thousands of individual PCR reactions would be required. Existing library generation methods designed to introduce the necessary sequences in a gene independent manner, such that mRNA libraries can be processed in a single reaction, take several days to complete and rely on several inefficient ligation steps[12]. In order to reduce the time required and to maximize the efficiency of library generation from mRNA, a method was sought that allowed our desired sequences to be introduced within the cDNA generation process. By leveraging template switching oligonucleotide (TSO) technology popularized in the SMART-Seq sequencing library preparation method, a process was developed that allows for a two-step synthesis of a DNA library from cellular mRNA. This technique is completed in under a day and results in library content largely comparable to the original mRNA library and which is appropriate for use in protein display.

When the representative power of the TSO library generation is combined with high throughput capability of the ligation and sequencing based protein interaction assay, the result is a unique and illuminating manner of examining cell-wide protein interactions. This approach is a relatively inexpensive, rapid, and accessible method for generating a characteristic protein interaction network for the library generated from a given cell type and state.

1.4 Contribution of Research

The novel methods described in this text fill two primary spaces in the proteomics field: the need for an accessible high-throughput protein-interaction profiling technology, and the need for a technique that quickly and efficiently generates custom cDNA libraries from a cell's mRNA. In doing so, these techniques facilitate work in wide range of research and medical fields.

2 Literature Review

2.1 Nucleic Acid Labeling of Proteins

The methods for labeling a protein with a nucleic acid fall largely into two categories, those that use some variant of covalent or affinity tagging to attach a synthetic oligonucleotide, and those rely on the precursor genetic material for the labeling. Typically, proteins in the first class need to be in isolated systems for the labeling process (each protein species needs to be labeled separately), while those that fall into the second class can be individually labeled even within a mixture of proteins. Due to the desired scale of our technologies, the discrete labeling of proteins would be inhibitory with respect to time and costs. Therefore, only techniques that could be applied to populations are reviewed here.

2.1.1 Display Technologies

Display technologies allow the protein phenotype to be linked to its genotype. There are three primary methods of display, phage display, ribosome mediated display, and cis-display. All of the display methods presented here require manipulation of the native gene sequence to perform display.

In phage display, phage DNA is modified to allow the protein of interest to be displayed within the phage coat. Selection for the protein displayed on the phage coat also results in the selection of the phage and its genetic material – which can be sequenced to identify the protein in question. The scope and size of proteins expressed in the phage display system are limited by the bacterial transformation process and the coat expression capacity[13].

Ribosome mediated display does not require the entire virus structure to accomplish this linkage, but rather uses the ribosomes to facilitate the connection between precursor RNA and subsequent protein. This results in the ability to produce libraries of greater scale containing larger proteins. Ribosome mediated display can be performed by creating conditions in which the ribosome complex is stabilized while both the amino acid chain and the mRNA are still attached[14]. Additionally, it can be

accomplished by the use of an amino acid analogue, puromycin, attached to the 3' end of the mRNA. As the ribosome approaches the 3' end of the template, the puromycin diffuses into the A site and the ribosome attaches the puromycin to the end of the amino acid chain via a peptide bond, forming an mRNA-puromycin-protein complex[15]. This second version of ribosome mediated display, also called mRNA display, offers a more stable bond and less steric hindrance than standard ribosome display. However, the mRNA itself is prone to degradation; stabilization is often accomplished with the generation of a cDNA strand[16].

Cis-display is a variation of the display technology that leverages DNA binding domains. The protein of interest is expressed as a fusion protein, coupled to a DNA binding domain. This DNA binding domain has a specific, known DNA sequence it will bind to that is also engineered into the gene. During coupled transcription and translation, the DNA binding domain on the protein causes the protein to bind to the gene sequence from which it was expressed. There are several difference binding sequences that have been used to perform cis-display, including RepA, P2A, and AGT[17–19]. This method offers increased library and protein sizes with respect to phage display, and a direct DNA to protein linkage, but the labeling method is slightly more promiscuous than the afore mentioned methods.

2.2 Protein Interaction Methodologies

The significance of work focusing on increasing the ability to assay protein interactions is made clear just by enumerating the vast number of techniques developed since the advent of the gold standard protein interaction method, Yeast Two-Hybrid (Y2H), in 1989.

The seemingly infinite applications of high throughput interaction technologies has led to numerous approaches to their development. The majority of the resulting techniques take the form of a 'one-by-many' assay, where a single protein of interest is probed against a library of 'prey' molecules. Only a handful of methods successfully address library vs. library assays.

2.2.1 High Throughput Protein-Protein Interaction Technologies

High throughput protein interaction technologies can be classified according to their method of identifying the interacting pairs. There are currently three primary classes of high throughput protein-protein interaction technologies, those that utilize the Y2H system, those that rely on mass spectrometry as a readout, and those that detect proteins using a nucleic acid barcode. Outside of these three main classes, there are just a handful technologies with alternative readouts.

2.2.1.1 Yeast Two-Hybrid Technologies

Y2H[20] has been an influential and successful method of querying protein and protein-DNA interactions for decades, and many efforts have focused on optimizing this system for high-throughput applications. Y2H has inherent disadvantages, however, including the need for the protein to be expressed, folded, and to interact in the yeast environment, the form of the protein as a fusion, and the possibility of interference from the native yeast biomolecules. In addition, the preparation of the plasmids, the transformation of the yeast, and the maintenance of the resulting cell lines can be challenging and time consuming.

The strategies to increase the throughput of this fundamentally binary system are either to massively parallelize the binary interactions[21], or to "stitch" the gene sequencing containing plasmids in positively interacting pairs and sequence the stitched plasmids (RLL-Y2H, Stitch-Seq)[3,22]. One technique does not attempt to query a library-by-library space; it uses a single "bait" gene against many "preys" and uses sequencing to identify the positively interacting partners (QIS-seq)[2]. While these systems do increase the number of interactions that can be surveyed simultaneously, they do not avoid the primary disadvantages of the Y2H system, and still require extensive preparation of gene materials.

There is one technique that utilizes a system highly similar to the Y2H reporter system, but in a mammalian environment, MAPPIT[23]. While this technique does allow the examination of mammalian

proteins in a more native environment, it still is subject to all other drawbacks of the Y2H system, and requires individual screenings of the bait-prey pairs.

2.2.1.2 Mass Spectrometry Technologies

Mass Spectrometry (MS) is not used independently to identify protein interactions, but rather alongside another technique which selects for or marks a population of interactors. MS is subsequently applied to identify the proteins within this population. Techniques commonly used alongside mass spectrometry are affinity purification[24] and proximity biotinylation (BioID)[25]. Both of these approaches use a bait protein, and select for a subset of the prey population as interactors. This population of interactors is then decoded via mass spectrometry. These techniques are limited to identifying the partners of the single bait protein used in the initial experiments.

Although MS technology has advanced significantly in the last decade, improving the sensitivity and versatility of the systems while simultaneously reducing the per sample costs, there are still several challenges to be aware of when using MS as a protein readout. Not all proteins are suited for MS. Various buffers and reagents can be utilized to circumvent compatibility issues, but it can be challenging to find one condition that is ideal for all the constituents of the system. Additionally, expensive reagents can be required to determine background protein levels in control samples, and MS system often are insensitive to the strength of an interactions.

2.2.1.3 Nucleic Acid Barcode Technologies

As sequencing techniques become more accessible and cost effective, techniques that utilize nucleic acid barcodes as a proxy for protein identity have become more prevalent in high throughput approaches; as the scale of the information that can be obtained with sequencing is able to match the scale of the potential protein interactome. Most nucleic acid barcode based protein-protein interaction technologies are limited to probing one bait protein against a population of prey – the bait is used to

perform a select step, and the barcodes used to identify the selected partners. Proteins can be labeled directly with either a display technique (phage[26,27], ribosome[1,28] or mRNA[29]) or a molecular attachment[1], or they can be labeled indirectly by using an oligo conjugated antibody which binds to the protein of interest[30].

Within the *in vitro* protein-protein interaction (PPI) space, Dr. George Church's SMI-seq technique has demonstrated the greatest scale at 55 proteins by 200 peptide fragments[1]. His approach utilizes proteins barcoded with synthetic sequences, which are allowed to interact, are crosslinked, and then separated on a gel matrix for in situ sequencing. The greatest challenge facing barcoding based systems is the difficulty of preparing appropriate protein libraries for each system. In some cases, these techniques require specialized equipment, which increases their costs and restrains their application.

2.2.1.4 Other High-Throughput Technologies

There are a handful of other approaches to identifying protein-protein interactions being explored that do not fall into one of the above categories. One significant area of exploration is the use of protein microarrays[31]. The challenges facing this technology are the difficulty of fixing proteins and allowing them to interact on a solid substrate, the creation of the bait proteins to be used as the bait on the microarray, and the detection of the bound proteins to the microarray. These techniques are currently limited to one prey protein against the arrayed bait proteins.

2.2.2 High Throughput Protein-RNA Interaction Technologies

At the time of writing, we have not been able to identify any techniques that attempt to query protein-RNA interactions in a library-vs-library manner. However, there have been attempts to simultaneously identify all proteins bound to RNA, without consideration of the actual pairwise interactions occurring[6]. These studies determined a significantly larger portion of the proteome was

binding RNA than previously thought, and indicated that RNA-binding proteins may play a significant role in the cellular environment.

Existing protein-RNA interaction techniques can be divided into four categories based on the method of identifying the binding partner, those that pull down a protein of interest and examine its RNA partners using sequencing (CLIP-seq, RIP-seq, PAR-seq)[32–34], those that pull-down an RNA of interest and examine the protein partners using MS (RAP, PAIR, MS2-BioTRAP, ChiRP, CHART)[35–39], those that use enzymatic labeling to determine the partners of a given RNA or protein of interest (TRIBE, RaPID)[40,41], and those utilizing microarrays (ProtoArray)[42]. As for protein-protein interactions, current methods to identify specific interaction partners require massive parallelization of one-by-many approaches, and therefore are extremely costly and demanding.

3 Overview of Methodology

This chapter provides and overview of the various steps contained within the library generation and the protein-protein interaction technologies. For a detailed discussion of the methodology and its validation, please see the subsequent chapters.

3.1 Protein Library Generation

A library of proteins must be generated with the necessary nucleic acid barcodes to allow for efficient identification of proteins during the implementation of the high-throughput protein interaction assay. As it is unique to each protein, and already present in the system, the precursor mRNA of each protein is covalently linked to its respective protein in the cDNA display process. This process has been described numerous times in the literature.

3.1.1 DNA Libraries from Constructs

To create gene templates that contain the necessary sequences to undergo the cDNA display process, PCR amplification is applied to either the plasmid or linear DNA fragment containing the gene of interest. The primers used are generally are gene specific, except in the case of a plasmid library lacking a stop codon. In such a library the general sequences flanking the gene of interest can be used, allowing for a single set of universal primers. Primers are designed to overlap the 5’ and 3’ region of the gene (or plasmid), excluding the stop codon at the 3’ end, and to contain non-homologous regions with the necessary sequences for cDNA display. The required sequences are: a translational promotor, transcriptional regulatory sequences, a protein affinity tag, and a GC-rich 3’ ligation site.

3.1.2 DNA Libraries from Natural Systems

Natural libraries are more complex, as the stop codon needs to be excluded from the gene sequence, all sequences need to be processed simultaneous (a gene specific primer cannot be used), and

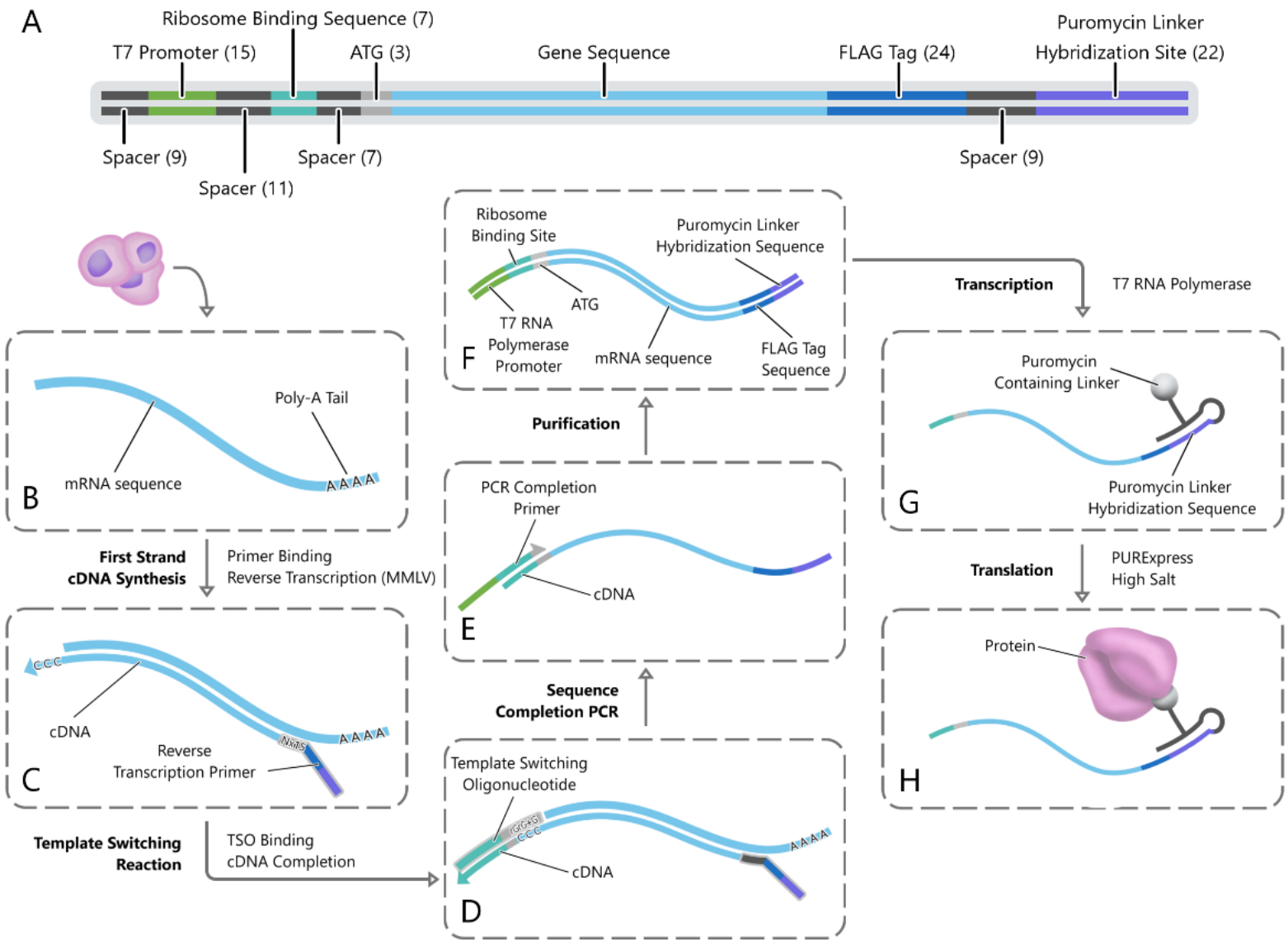
the starting material is mRNA rather than DNA. In these libraries, we use a variation of the SMART-Seq sequencing library generation technique; a template switching oligonucleotide in concert with random priming and cDNA synthesis. Both the random primer and the TSO sequences contain non-homologous regions containing the necessary synthetic sequences for cDNA display (Figure 3-1, A–F). The mRNA population used in this library generation are extracted from cultured cells.

3.1.3 Protein Libraries

The DNA libraries generated above are transcribed to RNA with a T7 RNA polymerase and subsequently ligated to a short hairpin linker containing a puromycin and a biotin molecule (puromycin linker). The ligation occurs in the 3' GC rich region of the gene template and results in a single stranded loop containing two RNA bases and a short double stranded region. This puromycin linker-RNA complex is pulled down on streptavidin beads and subject to *in vitro* translation (IVT) in the PURExpress system. When the ribosomes in the system encounter the double stranded region of DNA at the 3' end of the template, they stall, and the puromycin contained in the puromycin linker enters the A site of the ribosome and is added to the peptide chain. This results in a peptide-puromycin linker-RNA complex that is bound by the ribosome (Figure 3-1, G–H). Immediately following IVT, salt is added to increase the efficiency of the puromycin reaction and to release the ribosomes from their templates. Display reactions are then frozen overnight, which also improves the efficiency of the display. The complexes are then immobilized on streptavidin beads via a biotin contained within the puromycin linker. At this point in the procedure, display proteins are fully formed and are bound to the surface of the streptavidin beads.

Figure 3-1 SMART-Display Method for Generating Display Proteins from mRNA

A) A sequence diagram illustrating the final structure of gene templates resulting from the SMART-Display process. B) Poly-A selected and rRNA depleted mRNA is used as the input. C) A 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 is performed with the SuperScript II enzyme, and the TSO allowed to bind to the untemplated bases it adds to the 3' end of the cDNA. The cDNA is then extended by SuperScript II to incorporate the TSO sequences. 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) The final double stranded DNA structure. G) The RNA the results from the transcription of this DNA product can be ligated to a puromycin linker and used to generate a display protein complex. H) The structure of the final display protein complex, in which protein is attached to the puromycin linker which is attached to the precursor RNA strand.


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

3.2 Protein Protein Interactions: One Against Many

A single protein can be brought through the display independently to act as a ‘bait’ protein in a small-scale protein interaction assay. This bait protein can be assayed against 10’s of proteins and their affinities qualified by the ratio of their quantities before and after selection against the bait protein. Their quantities can be assayed with quantitative PCR (qPCR) for their mRNA barcodes.

In the process, the ‘bait’ protein remains on the streptavidin beads, while the ‘prey’ protein population, which can be processed in the same reaction, is released from the beads via digestion of the deoxyinosine nucleotides found in the loop region of the puromycin linker. The free prey proteins are mixed with the bound bait proteins in phosphate buffered saline, and a sample of the mixture is retained for comparative analysis. The bait and prey proteins are optionally cross-linked, and the bait proteins pulled down and washed to remove non-specifically bound proteins. The bait proteins are then released from the beads and the solutions retained for qPCR analysis.

The before and after selection samples are analyzed via qPCR and a ratio calculated as the Ct after interaction over the Ct before interaction. These ratios are ranked highest to lowest and correspond to the prey protein with the highest affinity for the bait protein to the protein with the lowest affinity respectively.

3.3 Protein Protein Interactions: Libraries

In the library interaction assay, a great number of proteins (thousands) can be tested against all other proteins in the system simultaneously. Interaction partners are determined by sequencing and subsequent bioinformatic identification and counting of chimeric barcode fragments.

The RNA library is split and processed in two cDNA display reactions. One reaction is ligated with a non-cleavable puromycin linker and will become the ‘bait’ library, the other is ligated with a cleavable puromycin linker and becomes the ‘prey’ library. Free biotin is added to the bait population

to block remaining binding sites on the streptavidin beads. The mRNA tags converted to double stranded DNA by a template switching reaction followed by second strand synthesis. The TSO used in this process introduces a non-palindromic restriction enzyme site the ends of the DNA barcodes, which is subsequently used for digestion. The prey population of display proteins are ligated to a biotin-conjugated interaction linker which contains the complementary restriction enzyme site on either end, and are then released from the streptavidin beads via digestion of the deoxyinosine nucleotides found in the loop region of the puromycin linker. The free and bound libraries are then mixed and incubated in an interaction buffer, and optionally cross-linked to stabilize interactions. The bait proteins are pulled down with the streptavidin bound prey and washed to remove non-specifically bound proteins. The biotin conjugated interaction linker is then proximity ligated to adjacent DNA barcodes.

After proximity ligation, the proteins are digested, and the DNA is fragmented with the NEB FS Fragmentation kit and prepared for sequencing with the NxSeq library preparation platform. DNA fragments containing the interaction linker are pulled down on streptavidin beads via their biotinylated nucleotide, and the library is amplified (Figure 3-2).

The paired sequencing reads are aligned and categorized as either chimeric if they align to two separate genes, or non-chimeric if they align concurrently. The chimeric calls are statically analyzed and tested for significance against either the background in the positive sample or the background in a set of experimental controls. For each gene pair extrapolated from a chimeric read, a chi-square test, an odds ratio threshold and a positive read count threshold is applied to identify significant protein-protein interactions. The chi-square test and the odds ratio cutoff are used to identify protein pairs for which the two participating proteins are more likely to be found interacting together than interacting with other proteins. The positive read count cutoff compensates for the bias of a high odds ratio as a consequence of low read counts for the involved genes. The chimeric read pairs that pass all tests of significance and signal strength are identified as representative of candidate protein-protein interactions.

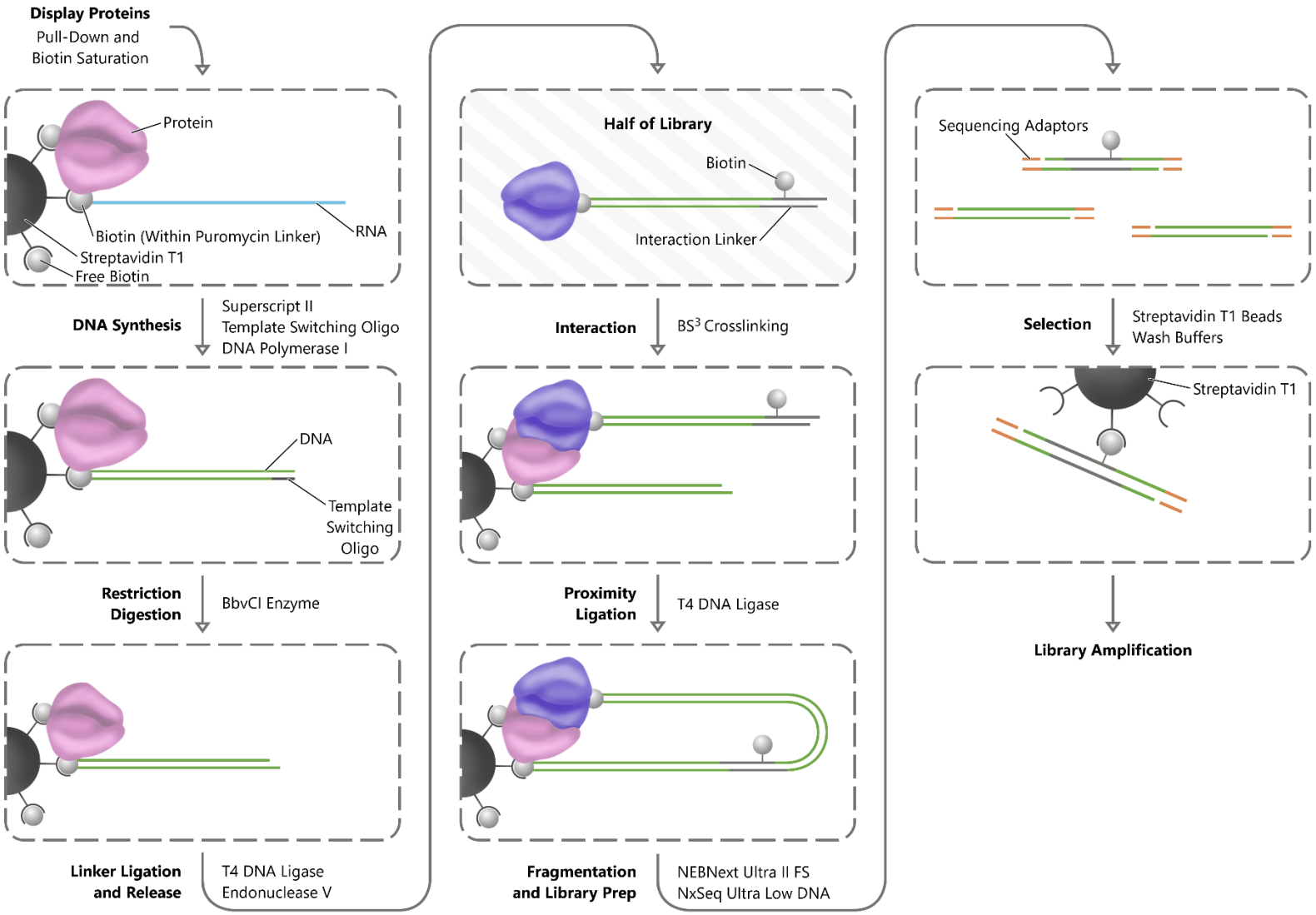


Figure 3-2 Workflow for PROPER-Seq Protocol
The protein-protein interaction workflow begins with prepared display proteins immobilized on streptavidin beads, and ends with a prepared sequencing library..

3.4 Protein RNA Interactions: Libraries

In the Protein RNA library interaction assay, a great number of proteins (thousands) can be tested against total RNA in the system simultaneously. Interaction partners are determined by sequencing and subsequent bioinformatic identification and counting of chimeric barcode fragments.

The process of performing the protein-RNA interaction assay for libraries is very similar to the process described in section 3.3. The SMART-RNA library is converted to a SMART-Display library in one reaction with a non-cleavable puromycin linker. This becomes the bait library. Free biotin is added to the bait population to block remaining binding sites on the streptavidin beads. The mRNA tags are stabilized with a cDNA strand in a template switching reaction. The prey population is a total RNA sample extracted from cells. The TSO used in this process introduces a non-palindromic restriction enzyme site to the ends of the DNA barcodes. Separately, the prey RNA are ligated to a biotin-conjugated DNA linker via single end RNA to DNA ligation. This linker is single stranded on one end, but contains the complimentary non-palindromic restriction enzyme site on the other. The bait and prey libraries are then mixed and incubated in an interaction buffer, and optionally cross-linked to stabilize interactions. The bait proteins are pulled down with the streptavidin bound prey and washed to remove non-specifically bound molecules. After second strand DNA synthesis and restriction digestion are performed on the tags for the display proteins, the biotin conjugated interaction linker is then proximity ligated to adjacent barcodes via the complementary restriction enzyme site in the linker.

After proximity ligation, the proteins are digested, the RNA is converted to DNA, and the DNA is fragmented with the NEB FS Fragmentation kit and prepared for sequencing with the NxSeq library preparation platform. DNA fragments containing the interaction linker are pulled down on streptavidin beads via their biotinylated nucleotide, and the library is amplified.

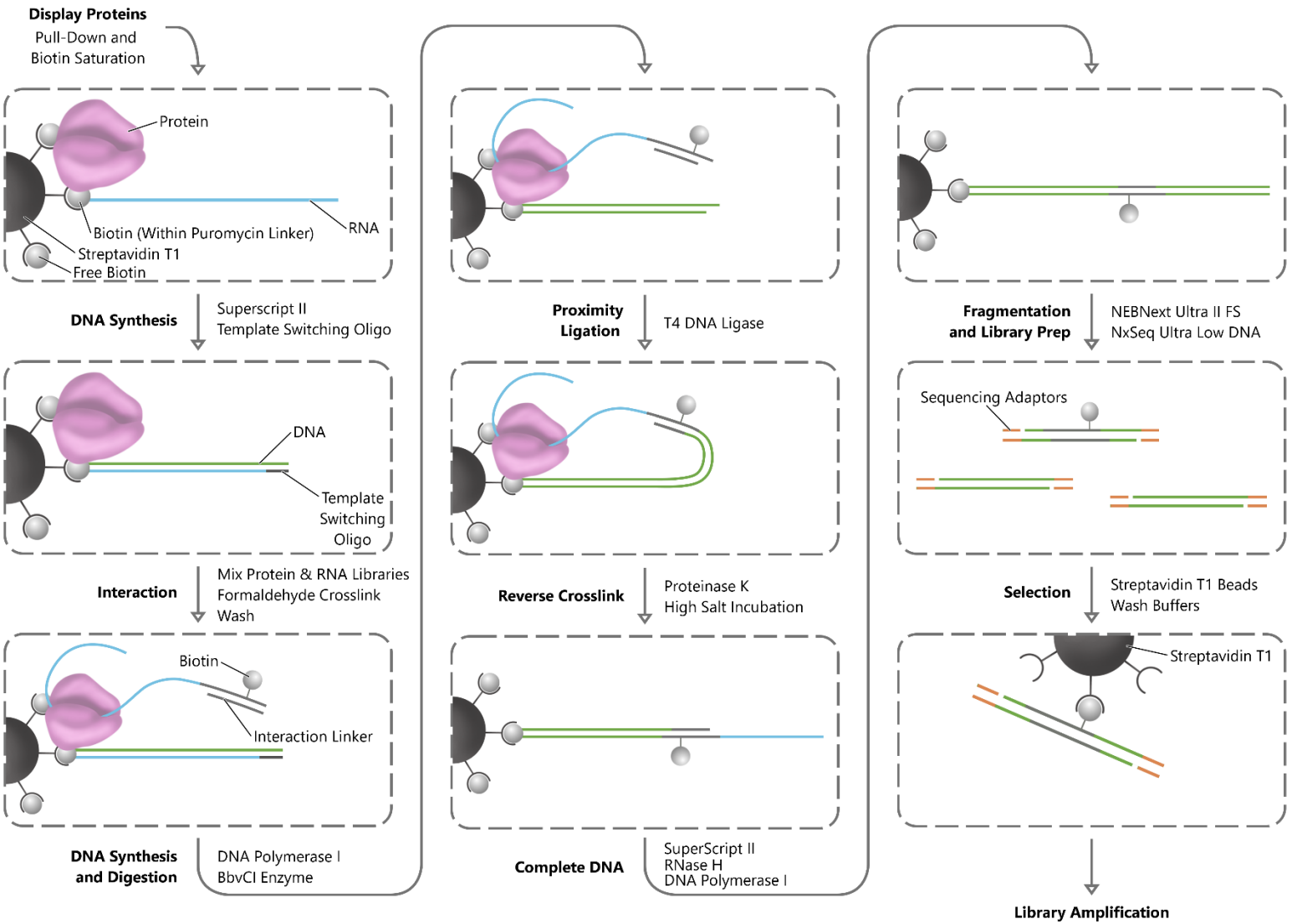


Figure 3-3 Workflow for Protein-RNA Interaction Protocol
The protein-RNA interaction workflow begins with prepared display proteins immobilized on streptavidin beads, and ends with a prepared sequencing library.

The paired sequencing reads are aligned and categorized as either chimeric if they align to two separate loci, or non-chimeric if they align concurrently. The orientation of the reads within the fragment indicate which end corresponds to RNA and which to protein. The chimeric calls are statically analyzed and tested for significance against either the background in the positive sample or the background in a set of experimental controls. The chimeric read pairs that pass all tests of significance and signal strength are identified as representative of candidate protein-RNA interactions.

4 SMART-Display: Display Protein Library Generation

4.1 Nucleic Acid Libraries from Constructs

4.1.1 Aim

This step of the protocol aims to generate DNA copies of genes that contain the sequences necessary for transcription, translation, and puromycin linker ligation such that they are amenable to the generation of a cDNA display protein.

4.1.2 Requirements

The cDNA display technologies used in this dissertation require additions to the protein coding DNA sequences for successful execution. These sequences for expression in the PURExpress system are detailed in Table 4.1. Other sequences may be optionally included, such as the protein affinity tag found in the constructs described in this text, but are not required. There is one sequence that cannot be included within the template; the presence of a stop codon causes the release of the ribosome during translation and a failure of the puromycin linker to be attached to the peptide chain. Therefore, no display proteins will be created during the process.

Table 4.1 Required Sequences for cDNA Display
This table includes a list of the sequences that must be included with a protein coding gene to successfully create a cDNA display protein. The sequence name, location, purpose, and sequence are given.

Sequence Name	Location	Purpose	Nucleotide Sequence
T7 Polymerase Promoter	5' most, before the gene sequence	Allows for transcription of the DNA template by T7 RNA Polymerase.	5' TACGACTC ACTATAG 3'
Shine Delgarno and Translational Start Site	5', before the gene sequence and following the T7 Polymerase Promoter	The Shine Delgarno sequence is required for ribosome recruitment and binding to RNA template in bacterial IVT systems, the start site is required for the initiation of translation.	5' AAGGAGN NNNNATG 3'
Puromycin Ligation Site	3' terminus, after the gene sequence, replaces stop codon.	This GC-rich region allows the RNA template to be annealed and ligated to a hairpin DNA linker containing a puromycin (the puromycin linker).	5' AGGACGGGGGG CGGCGGGGAAA 3'

4.1.3 Approach

In our synthetic systems, gene sequences originate from synthesized sequences (GeneBlocks from IDT) or from plasmid constructs. Using sequences in these formats allows specifically selected genes to be used in the technology development and validation process. Gene sequences originating from a GeneBlock or plasmid are quick and easy to handle, facilitating the development phase of the project.

4.1.3.1 Primers

Primers synthesized by IDT are used to introduce the sequences outlined in Section 4.1.2 above. When amplifying genes from GeneBlocks, the primers are gene specific, as the gene blocks contain no homologous regions from sequence to sequences. However, when amplifying genes with their stop codons removed from a plasmid, the conserved regions flanking the 5' and 3' end of each gene can used to create a single set of universal primers for the plasmid family.

The initial version of the universal primers designed for the plasmid library used in this text contained the most simplistic ordering of the required sequence elements, simply placed one after another as shown in Figure 4-1 and Figure 4-2.

Figure 4-1 Version One of the Forward Universal Plasmid Primer
The forward universal plasmid primer version one introduces the necessary 3' sequences for transcription and translation in a bacterial system. The plasmid sequence from the test genes is directly amended to the PURExpress recommended primer containing the required T7 Promoter and the Shine Delgarno sequence. This primer was used with the reverse universal plasmid primer presented in Figure 4-2.

Figure 4-2 Reverse Universal Plasmid Primer
The reverse universal plasmid primer contains the 5' sequences for binding of the puromycin linker and a FLAG affinity purification tag following the plasmid sequence. This primer was used with the primers presented in Figure 4-1 and Figure 4-3.

This version of the forward primer yielded no translation of kinase genes when used in the PURExpress IVT system (Supplementary Figure 1). However, the positive control provided with the PURExpress system was successful, indicating that the lack of translation was due to either poor template quality

and/or suitability, or protein incompatibility with the system. The sequence of the template was confirmed with Sanger sequencing. Conventionally, the distance between the Shine-Delgarno sequence and the translational start site is five nucleotides or less. Because of the included plasmid sequence, the Shine Delgarno sequence in version one is fifteen bases from the start site. This increased distance is likely the cause of the absence of translation.

To avoid the challenges with the original primer, versions two and three of the forward primer use short regions of non-homology within the plasmid sequence to introduce a Shine Delgarno sequence five bases from the translational start site, as shown in Figure 4-3. The universal plasmid primers were tested with the DHFR gene, as it is the positive control used in the PURExpress system. This controlled for the effectiveness of the primers without concern for the compatibility of the gene with the translation system. Figure 4-4 illustrates the differences in expression levels of the three primer sequences. It is clear that translation levels are significantly impacted by the distance of the Shine-Delgarno sequence from the start site. Primer version three was ultimately selected for use in the system.

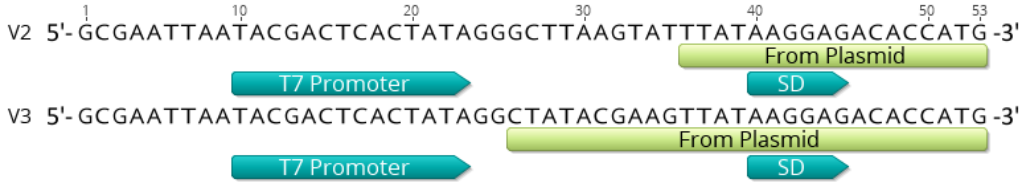


Figure 4-3 Versions Two and Three of the Forward Universal Plasmid Primer
Versions two and three of the Forward Universal Plasmid Primer introduce the necessary 3' sequences for transcription and translation in a bacterial system. The region containing the Shine Delgarno (SD) sequences do not have 100% homology with the plasmid sequence, allowing the Shine Delgarno sequence to be introduced five base pairs upstream of the translational start site. These primers are used with the reverse universal plasmid primer presented in Figure 4-2.

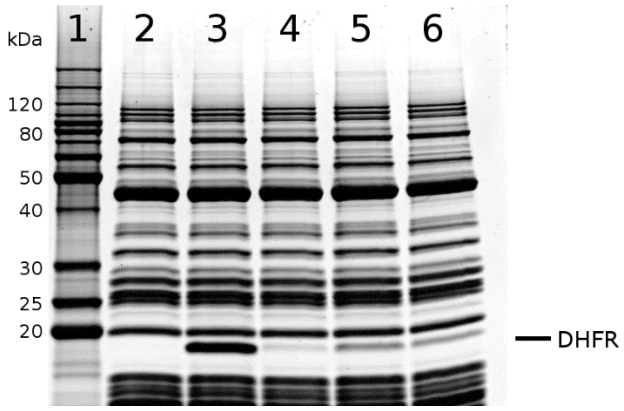


Figure 4-4 PURExpress IVT Products for Versions of the Universal Plasmid Primer
Lane (1) in this polyacrylamide gel contains Benchmark Unstained Protein Ladder. Lanes 2-6 six contain PURExpress products for reactions containing (2) No Template (3) DHFR Template (PURExpress Positive Control), (4) DHFR template with primer version one, (5) DHFR template with primer version two, and (6) DHFR template with primer version three. Gel was stained with Sypro Ruby and imaged in a UV lightbox.

In comparison to the positive control in Figure 4-4 (lane 3), the DHFR expressed with the universal plasmid primers is significantly less concentrated. As the proteins in lanes 4, 5 and 6 result from templates with 3' sequences that lack a stop codon, the ribosomes in the system tend to stick to the templates and do not get recycled. The manufacturers of the PURExpress system estimate that this will result in five-fold less expression than in a system with a typical stop codon. This estimate is supported by our own data.

4.1.3.2 Amplification Conditions

The primers described above were used to generate DNA templates appropriate for cDNA synthesis during a fairly standard PCR reaction. The amplification enzyme predominately used was GoTaq Green Master Mix; this regent is inexpensive and tolerant to most templates and conditions. While this is a not a high fidelity enzyme, sequences in the bulk of the PCR product will have a conserved sequence. If extensive amplification is to be performed, or a template is particularly long, a high fidelity DNA polymerase enzyme would be preferred.

The PCR primers used in the system only partially overlap the target sequence during the initial stages of PCR, but then have long binding regions in the subsequent cycles (up to 80 bp). To achieve specific binding in these PCR reactions, the initial two cycles of PCR are performed with an annealing temperature appropriate for the shorter binding region. All other cycles are performed as a two-step PCR, with a combined annealing and extension phase at 72 °C.

4.1.3.3 Transcriptional System

Before proteins can be translated, the DNA libraries must be transcribed into RNA, and the RNA ligated to the puromycin linker. Transcription in our system has been performed with a variety T7 polymerase based kits with very little variation in quality and yield. Before use in the next step, the RNA is checked with a High Sensitivity Bioanalyzer run to ensure full-length templates have been produced. With few templates, this task is simple, the size of each template can be independently confirmed. In libraries, individual templates cannot be identified, and the overall size distribution must be compared to the precursor DNA library to ensure quality.

4.1.4 Validation

Each sequence generated in this manner is validated in two ways, sizing and sequencing. Each gene construct is be subject to either an agarose gel or an Agilent DNA Bioanalyzer analysis to establish that the template is of the expected size. If a template is appropriately size, it is sent for Sanger sequencing with primers internal to the gene. Sanger sequencing will determine both the inclusion of regulatory sequences and puromycin binding site at the end of the DNA fragment and the gene sequence and content within the template. Due to the limitation in sequencing length to just under 1000 base pairs, longer templates cannot be verified with a single set of primers in this manner, but can be with the use of several sets of Sanger sequencing primers.

4.2 Nucleic Acid Libraries from Natural Systems

4.2.1 Aim

This methodology aims to generate a DNA library that is representative of the mRNA population of the cell. The genes present in the mRNA library should be present in the DNA library as well. Each gene should contain the sequences necessary for transcription, translation, and puromycin linker ligation such that they are amenable to the generation of a cDNA display protein library. As mRNA libraries contain thousands of genes, we wish to create the DNA library in a single reaction, to avoid the time and cost associated with individual preparations.

4.2.2 Requirements

The sequence requirements for a DNA library generated from mRNA are the same as described in section 4.1.2 above, and are detailed in Table 4.1. The primary difference between libraries generated from synthetic sequences and/or plasmids and from mRNA is the method used to introduce the required sequences, and not the content of the sequences themselves.

4.2.3 Approach

Libraries have been successfully generated for this project from mouse E14, HEK 293T, Jurkat, K562, and HUVEC cell. These cell lines are all well-studied systems and familiar to many labs.

Total RNA is purified from a homogenous population of cultured cells with Trizol treatment followed by phenol-chloroform extraction; typically, on the order of millions of cells are purified in a single reaction. In theory, fewer cells are required to generate the necessary amount of mRNA for the subsequent library generation, but this lower limit has not been explored. Typical mammalian cells contain 10-30 pg of total RNA, with 1-5% of that comprised of mRNA.

The mRNA is selected for from the total RNA with magnetic oligo-dT beads. The purity and distribution of the sample is checked with the RNA Pico Bioanalyzer kit, and the oligo-dT purification repeated as necessary. Size distributions vary for different cell types, but generally have a bell like shape, with the peak shifted to the left of center. Significant contamination of the sample with rRNA or other non-coding RNAs will reduce the library yield in future steps. A case where rRNA contamination was evident, and additional purification was required is depicted in Figure 4-5. Removal of the rRNA with an rRNA depletion kit was sometime necessary to ensure a high level of purity. The Illumina RiboZero kit, the NEBNext rRNA Depletion kit, and the Invitrogen RiboMinus kit have all been used with success.

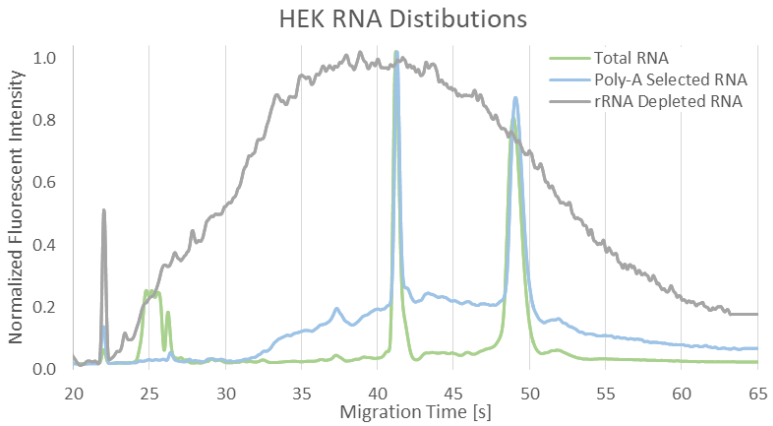


Figure 4-5 Oligo-dT Purification of HEK Total RNA
The distribution of HEK RNA at various point during mRNA purification. RNA was poly-A selected with oligo-dT beads and depleted of rRNA with Illumina's RiboZero Gold Human Kit.

The purified mRNA is then appropriate for use in the library generation protocol. Populations of mRNA present special challenges to library generation. The three most prominent are:

(1) The sequences in our mRNA library are unknown.

(2) A method is required that can simultaneously introduce sequences to all genes present in the system. As each gene sequence is unique, PCR with universal primers is not feasible.

(3) The stop codon must be removed from the gene while preserving as much of the remaining coding sequence as possible.

To accomplish this, existing technologies that introduce sequences were explored. Many of these technologies are represented in sequencing library preparation kits, as sequencing also requires sequence amendments to the 5' and 3' ends of linear DNA fragments. Three common approaches are ligation (either blunt ended or restriction enzyme based), overhanging PCR primers (as used with our plasmid and synthetically based genes), and SMART-Seq cDNA synthesis[43]. PCR based techniques were excluded due to the restrictions of the system, and ligation based approaches demand significant time and are inefficient. A SMART-Seq based library generation technique was developed for this project and is described in section 4.2.3.1.

4.2.3.1 SMART-Display Library Generation

The SMART-Seq based library generation method relies on a specialized oligonucleotide, called a template switching oligonucleotide, or TSO, and the nonspecific tailing activity of a family of reverse transcriptase enzymes. A TSO-based approach was developed to generate nucleic acid libraries that are suitable for the display process. This technique, through the subsequent transcription and translation, has been named SMART-Display.

4.2.3.1.1 Template Switching Oligonucleotide (TSO) Design

A TSO typically contains approximately 40 bases, with the three 3' most bases consisting of three riboguanines or two riboguanines and a locking guanine. These three bases have greater than average binding strengths to deoxycytosine bases. The TSO is used in concert with a reverse transcriptase derived from the Moloney Murine Leukemia Virus, which have terminal transferase activity that causes an average of two to five non-templated bases to be added to the 3' end of the generated cDNA, with a heavy preference for cytosine. In this work, the SuperScript II enzyme was used. The overhanging bases added by the reverse transcriptase are available for binding, and the TSO with its modified 3' end will bind to the short overhang quite effectively. After the TSO is bound to the cytosine bases, the reverse transcriptase will continue to transcribe, adding the sequence of the TSO to

the existing cDNA strand. This process allows us to introduce sequences to the 5' end of a DNA template independent of its sequence, and during the reverse transcription process. To prevent the introduction of multiple TSOs on the 5' end of a single template, the TSO is blocked on the 5' end with a biotin or chemical spacer.

The amount of sequence that can be introduced with the TSO is limited, as the TSO has an optimal length at which the template switching reaction is most effective. This limit is at approximately 40 bases. Because the sequences required for the cDNA display reaction exceed this 40 base allowance, the remaining 5' sequences are introduced during a subsequent PCR step. The design of the TSO in use in the protocol is shown in Figure 4-8.

Figure 4-6 TSO Design in SMART-Display
Structure of the TSO used in the SMART-Display Protocol, including regulatory sequences.

4.2.3.1.2 Reverse Primer Design

The obvious place to introduce sequences to the 3' end of the template is within the cDNA synthesis primer, which must bind to the 3' end of the template to initiate reverse transcription. These primers typically take one of two forms, a poly-T primer, designed to hybridize to the poly-A region of an mRNA template, or a random primer, which can hybridize anywhere within the template. The poly-T primer has the advantage of binding on the 3' side of the coding sequence, guaranteeing that the entire coding region is captured. However, this also ensures that the stop codon is captured, which will cause the cDNA display reaction to fail. For this reason, a random primer is used in this methodology (Figure 4-7). This primer will bind randomly within the template, causing loss of coding sequences in some cases, and inclusion of the stop codon in others. While the loss is undesirable, random primers allow a significant portion of the DNA library to be viable in the cDNA display process. The random primer is used to introduce sequences by the inclusion of a non-complementary region on the 5' end of the primer.

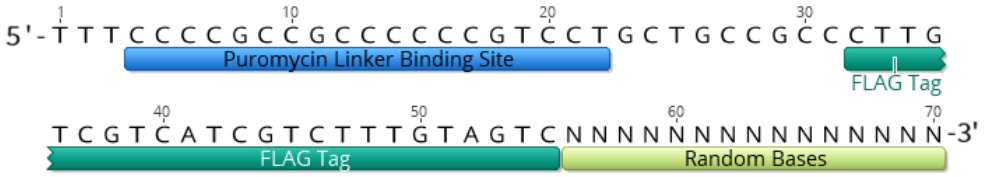


Figure 4-7 Reverse Random Primer Design in SMART-Display
Structure of the reverse random primer used in the SMART-Display Protocol to introduce 3' sequences to mRNA templates.

Following the completion of the template switching reaction, a PCR amplification step is performed. This serves the purpose of both completing the necessary sequences for cDNA display and selecting for and enriching sequences that were successfully modified with the reaction. The entire library preparation process is described in detail in Figure 4-8. In less than eight hours, a DNA library suitable for use in the cDNA display protocol can be generated.

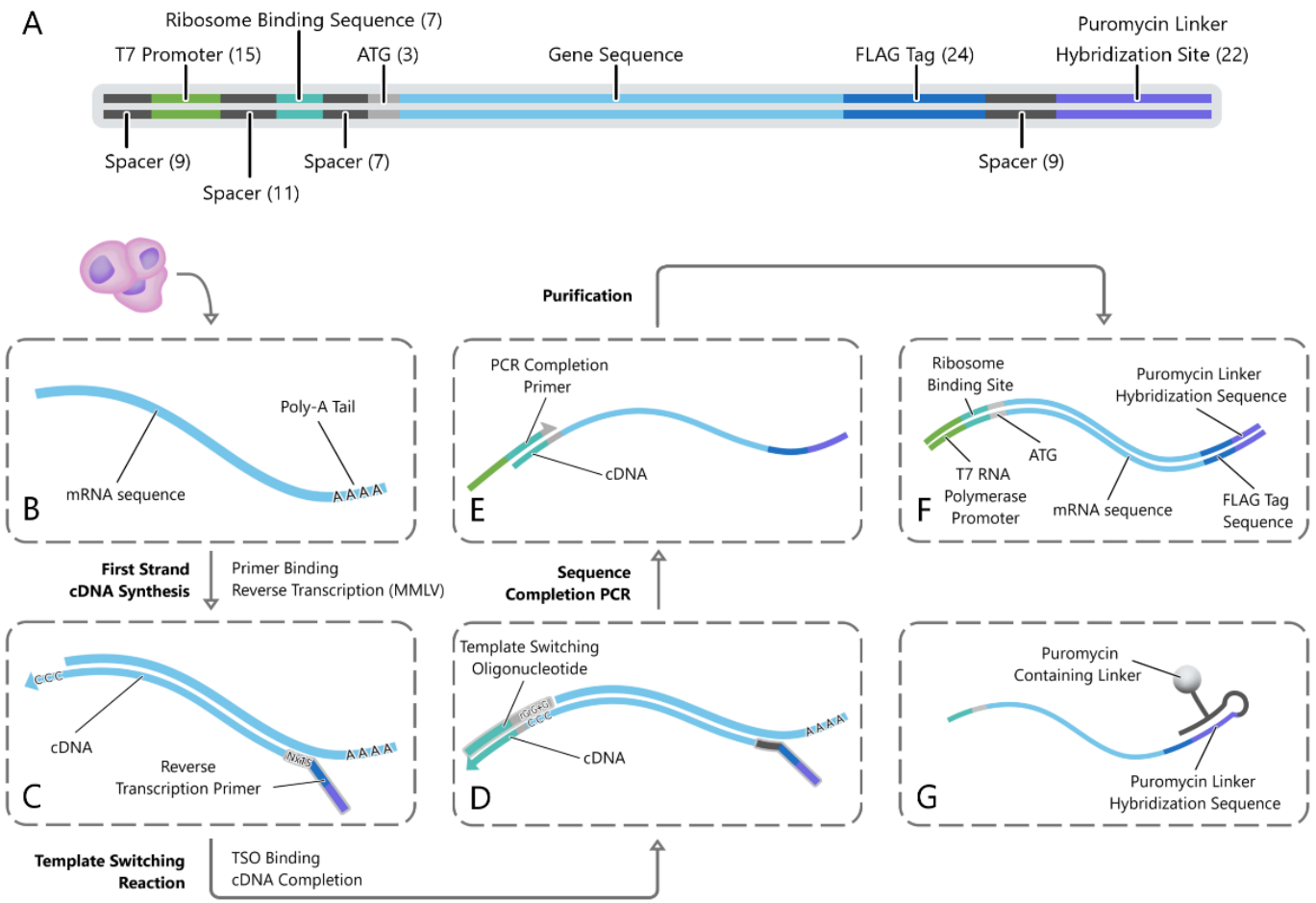


Figure 4-8 SMART-Display Method for Generating DNA Libraries from mRNA

A) A sequence diagram illustrating the final structure of gene templates resulting from the SMART-Display process. B) Poly-A selected and rRNA depleted mRNA is used as the input. C) A 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 is performed with the SuperScript II enzyme, and the TSO allowed to bind to the untemplated bases it adds to the 3' end of the cDNA. The cDNA is then extended by SuperScript II to incorporate the TSO sequences. 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) The final double stranded DNA structure. G) The RNA the results from the transcription of this DNA product can be ligated to a puromycin linker and used to generate a display protein complex.

4.2.3.2 Transcriptional System

As with the plasmid and synthetic sequence based systems, the DNA libraries must be transcribed into RNA, and the RNA ligated to the puromycin linker prior to protein synthesis. Transcribed RNA libraries are examined with a High Sensitivity Bioanalyzer run to ensure the distribution is as desired. A sample RNA distribution is illustrated in Figure 4-9.

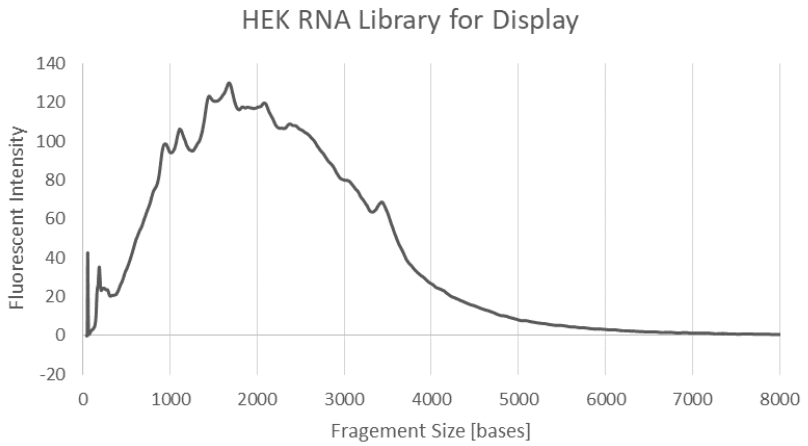


Figure 4-9 RNA Library distributions for Display Libraries
RNA Pico Bioanalyzer trace for HEK RNA libraries for display generated by NEB's HiScribe™ T7 High Yield RNA Synthesis Kit from the HEK DNA libraries for display.

4.2.4 Validation and Results

The DNA and RNA libraries generated from mRNA populations must be both representative of the gene population found in the original mRNA, and must contain the introduced sequences necessary for the display protein process. Libraries are evaluated for all these attributes.

A MiniSeq is utilized to analyze the gene and introduced sequence composition of a complete library. This measure will allow for a comparison of the gene composition in the DNA and subsequent RNA libraries to that of the original mRNA library. The initial data set used to evaluate the effectiveness of the method was generated from a cultured population of HEK 293T cells. The resulting cDNA library was prepared with the NEB Ultra DNA Library Prep Kit for Illumina Sequencing, with an average insert

size of 200 base pairs, and the RNA library with NEB Ultra RNA Library Prep Kit. Due to the large quantity of introduced sequence (just over 100 base pairs total), the libraries were sequenced with 5% PhiX DNA to balance the base distribution.

For the DNA library, the MiniSeq provided approximately 30 million 75 base long paired end reads, which exceeds the estimated 20 million reads required for a RNA expression analysis – which this DNA library reflects. The alignment of this sequencing run is low, due to two a large adaptor population. The RNA library resulted in just under 42 million 75 base long paired end reads, which far exceeds 20 million reads. The alignment of this sequencing run is more typical, 83% of reads aligning. This alignment should be slightly lower than a typical mRNA alignment, as a portion of the reads with be comprised of introduce sequences and will contain no transcriptomic matches.

This data was compared to the RNAseq library generated from HEK mRNA that was unmodified, prepared with the NEB Ultra RNA Library Prep Kit for Illumina Sequencing, and sequenced as described above. The mRNA library yielded 44 million reads, with an alignment rate of 92.88%. The sequencing statistics for the two libraries are described in Table 4.2, and the sequencing work flow is illustrated in Figure 4-10.

Table 4.2 Sequencing Statistics for HEK mRNA and associated DNA and RNA Libraries
This tables gives the library preparation kit used, the total number of reads obtained, and the alignment rate to the Human19 transcriptome for the three HEK samples, mRNA, DNA and RNA. The RNA and DNA libraries result from the processing of the mRNA with the SMARTSeq based library generation method for cDNA display and subsequent transcription.

Sample	Sequencing Library Prep Kit	Number of Paired Reads	Read Alignment (%)
HEK mRNA	NEB Ultra RNA	43,937,453	90.81
HEK DNA Library	NEB Ultra DNA	29,366,021	60.20
HEK RNA Library	NEB Ultra RNA	41,918,076	83.40

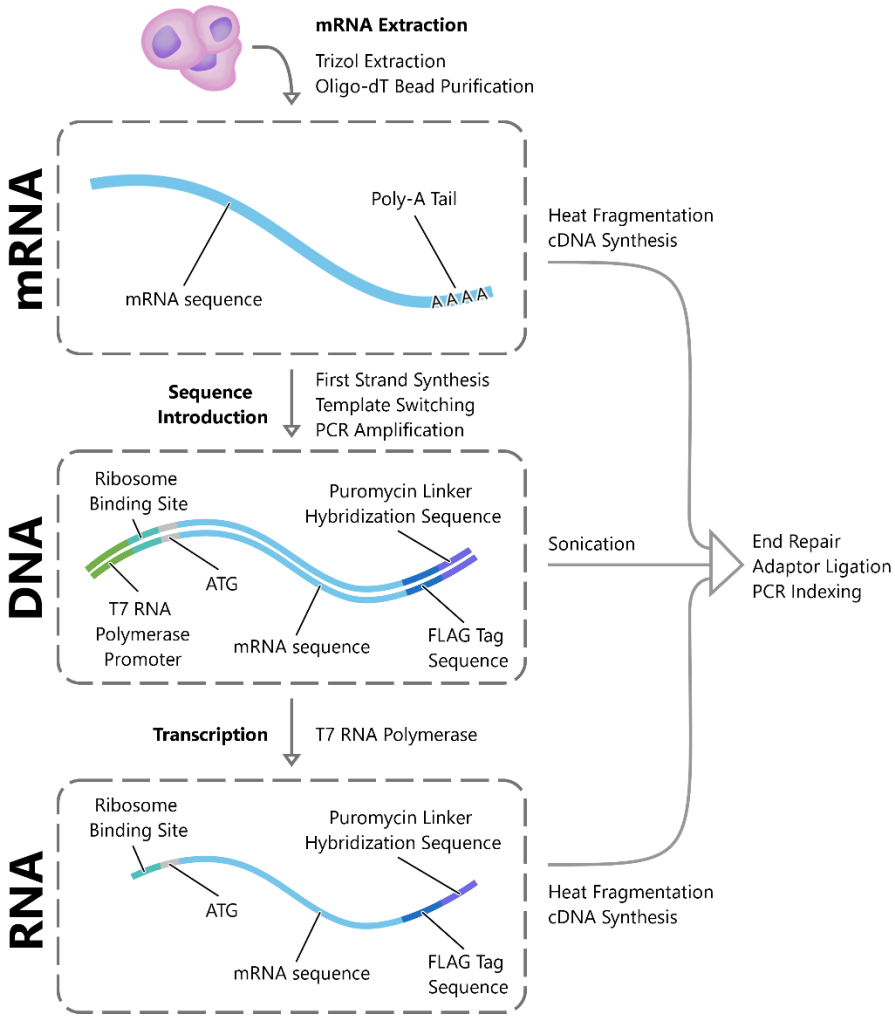


Figure 4-10 Library Generation Workflow for Display Libraries
This diagram illustrates the experimental workflow applied in the generation of each of the three libraries. Each of the major steps for generating both the libraries themselves and their corresponding sequencing libraries are listed.

In the DNA sequencing data, the introduced sequences were identified, counted, and trimmed from the reads with the package 'Bio Pieces'. The longest introduced sequence is 55 bases, leaving 20 bases of mRNA sequence for alignment, which, along with its paired 75 base read, should be sufficient for mapping. Subsequently, the trimmed reads were aligned to the Human19 transcriptome with HISAT2, sorted and duplicates removed with Samtools, and the FPKM for represented genes calculated with Cufflinks against the RefSeq database. The mRNA library underwent the same computational pipeline, including the processing with Bio Pieces. Although no cassettes are expected in the mRNA library, Bio

Pieces identified and trimmed approximately ten thousand instances of introduced sequences from the mRNA data. This is an insignificant number compared to the millions of instances found in the DNA library, and is due to randomly matching sequences in the mRNA. As the minimum required match to the introduced sequence is only ten bases, several random matches are expected, and should not influence the alignment of the mRNA library.

In the DNA data, each of the reads containing either a 5' or 3' introduced sequence was tallied and reported. This data is presented in Table 4.3, along with the percentage of total reads that contained each sequence.

Table 4.3 Introduced Sequence Representation in Sequenced Libraries
This table gives the number of reads from the sequencing data that each of the 5' and 3' sequences were identified in, along with the percentage of total reads this number represents for each library.

Sequence	Location in DNA	Library	Number of Reads Identified In	Percent of Total Reads Identified In
5' GCGAATTAATACG ACTCACTATAGGGCT TAAGTATAAGGAGGA AAAAATATGGG 3'	5'	mRNA	7,141	0.01
		DNA	5,561,037	9.47
		RNA	5,924,675	7.07
5' TTTCCCCGCCGCCCC CCGTCCTGCTGCCGCC CTTGTCGTCATCGTCT TTGTAGTC 3'	3'	mRNA	3,665	4E-3
		DNA	7,196,203	12.25
		RNA	6,165,681	7.35

As the sequences are almost identical in length and both are found at the extreme ends of each DNA/RNA fragment, it is expected that the 5' and 3' sequence would be found in approximately even numbers. This is re-enforced by the PCR amplification step that occurs prior to the library preparation, as templates that do not contain both the introduced 5' and 3' sequences should not amplify and therefore

should be underrepresented. While the data indicates a small deviation in sequence representation of the 5' and 3' sequences in the DNA library, it is acceptable.

The DNA and RNA libraries will be considered representative if they capture 80% of the genes with RPKMs greater than zero identified in the mRNA reference library. Due to the low alignment rate in the DNA library, the expectation of co-occurring genes is less, as the DNA library is likely not completely represented in the 15 million aligned reads. The alignment to the human19 genome with HISAT2 was stringent, and only unique alignments were kept. Cufflinks was used to identify and calculate the FPKMs of transcripts and genes against the human19 RefSeq database. At the time of this writing, this database contained 55,957 genes. The gene counts for each library were tabulated in R. The data is given in Figure 4-11.

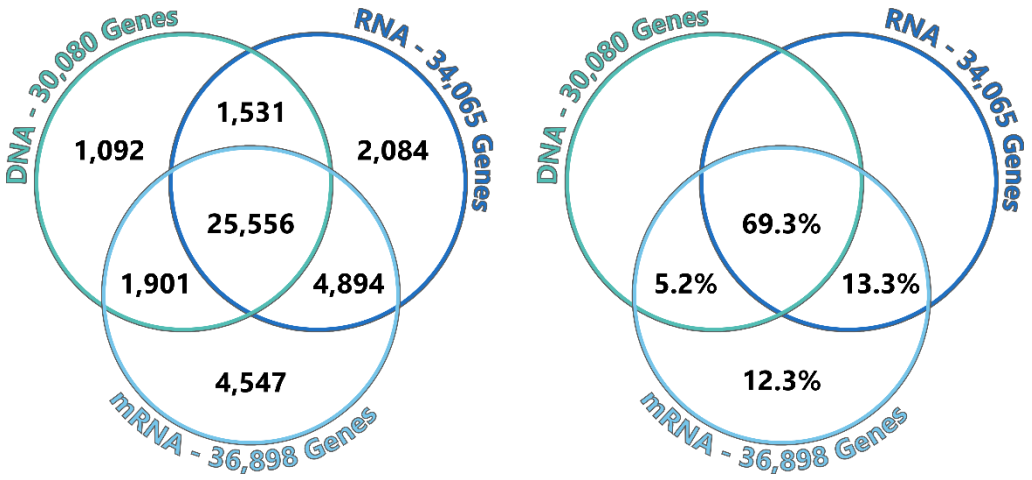


Figure 4-11 Genes Detected in HEK Sequencing Libraries
The Venn diagrams above illustrate the number of genes identified in each sequencing library with an FPKM greater than zero, and the overlap of those gene with the other libraries. The total number of gene detected is given along the outside edges of each library's circle. On the left, absolute gene counts are given. On the right, the percent of the total mRNA library represented by a selection of these gene counts are given.

The RNA library successfully at captured the genes contained within the mRNA library, with just over 82% of genes represented. As the FPKM limit is increased 0 to 0.1 and 1, this percentage stays at 80%. The data for the DNA library, however, indicates that approximately 75% of the genes found in the mRNA library were also found in the DNA library. As we adjust the FPKM upwards from 0 to 0.1 and

1, this percentage hovers between 70% - 80%, which falls short of the 80% mark. However, this is still a representation of almost thirty thousand human genes, which is quite a significant number. Interestingly, if we restrict the genes identified in the mRNA library only to those with FPKM of 0.1 or higher, while still only requiring an FPKM of 0 or greater in the DNA and RNA libraries, the represented proportion in the DNA and RNA libraries rises to 83% and 90% respectively. At an FPKM of 1, these numbers again increase to 93% and 96%. This indicates that the genes that are not being captured by our libraries are those that are poorly represented in the original sample.

Additionally, the final library of display prepared RNAs reflects the relative concentration of those RNAs in the original mRNA. When the FPKMs of each gene are plotted for the mRNA library and the display RNA library, we see a positive linear correlation with a coefficient of 0.69 (Figure 4-12).

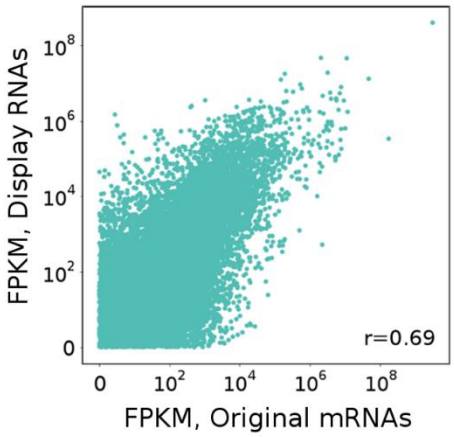


Figure 4-12 Correlation between mRNA FPKMs and display RNA FPKMs
For each gene identified in both the mRNA and display RNA libraries, the respective FPKMs were plotted. A linear correlation was taken and the coefficient 'r' was calculated to be 0.69.

Therefore, the display RNA libraries are representative of, to a reasonable degree, the expression levels of the RNA found in the original cells.

For each library, there are a small number of genes not represented in any other library. For the mRNA library, we expect there to be uncaptured genes. However, in the other two libraries, the genes

represented should be restricted to those found in the mRNA library, and the genes should therefore always co-occur. The library preparation and sequencing process however are imperfect, and there will be genes that are contained in the biological mRNA sample that will not represented in the library, and genes detected in the RNA and DNA libraries via sequencing and data analysis that may be false positives. Whatever the cause for these 'orphan' genes, they make up only 3% and 6% of the genes detected in the DNA and RNA libraries respectively. This representation should not have any significant impact on the interaction data gathered from these libraries.

4.3 SMART-Display Protein Libraries

The generation of display proteins is a critical step in the overall goal of assaying high throughput protein interactions via nucleic acid barcodes. However, proteins are among the most challenging of biomolecules to express and manipulate *in vitro*. There are two factors that influence the display process, the design and ligation of the puromycin molecule (which forms the covalent bridge between nucleic acid and protein) and the translation reaction (which forms the proteins themselves).

When designing a puromycin linker, the considerations regard the attachment of the linker to the nucleic acid, incorporation of functional molecules such as biotin, and the flexibility and length of the puromycin arm which allows the puromycin free access to the A site of the ribosome.

The primary consideration in translating proteins *in vitro* is the selection of the translation system. There are four primary translation systems commercially available; bacterial, insectoid, mammalian (rabbit), and human. Bacterial systems typically have the greatest versatility and the lowest cost, but result in no protein modification and often do not translate eukaryotic proteins effectively. Rabbit based systems handle eukaryotic proteins better than the bacterial systems, and offer some modification, but generally have less overall protein yield and breadth. Finally, human systems offer the most accurate protein modifications for human genes, but are significantly less broad in application and

are much higher in cost[44]. None of these systems is ideal, and each system has unique applications for which it is best suited.

All three of these systems are generated from cellular lysate that is processed to reduce RNase activity, proteinase activity, and background levels of endogenous mRNA. Residues of these three components reduce the efficiency and breadth of translation. In order to address these shortfalls, and to improve the ease of target protein purification, NEB offers an alternative product to traditional cell lysate, PURExpress. PURExpress is a synthetic bacterial translation system, containing only the essential machinery for translation. Additionally, all of the components contain a His-Tag to allow for reverse purification of protein products.

We have successfully and repeatedly utilized the PURExpress system to generate full-sized, active proteins. The ability of this system to generate proteins is not inhibited by the ligation of a puromycin linker to the mRNA, and therefore is suitable for the expression of display proteins. We similarly demonstrated successful creation of the display complexes with several puromycin linker designs, each with their own function and application.

4.3.1 Aim

This phase of the protocol aims to produce a library of proteins bonded to their respective RNA sequences, such that the proteins can be identified by their unique RNA barcode. This protein library should reflect the genes found in the precursor DNA library.

4.3.2 Requirements

Unlike the DNA libraries, the protein library does not have strict requirements in terms of the sequence or structure of the proteins. Rather, there are a set of properties that the library as a whole must meet, and a separate set of desired properties for individual protein structures.

It is a necessity that the protein population reflect the genes contained within the precursor DNA library. However, the population of successfully produced proteins will be some fraction of the genes contained in the DNA library, as there are known limitations of the *in vitro* translation technology that results in a population of genes that will not translate. It is also a necessity that these proteins are bound to their respective nucleic acid barcodes. Without the barcodes, there is no way in which the protein interactions can be identified and decoded.

Beyond these two requirements, it is beneficial if the proteins are full length and appropriately modified. Proteins that are not full length or modified can still yield useful interaction information, as protein binding does not always require the entirety of a protein, simply a binding domain, and not all interactions require modifications. Many existing protein interaction technologies that leverage in vitro expression, or expression in non-native systems, face similar challenges.

4.3.3 Approach

4.3.3.1 Puromycin Linker Design

Complete and validated RNA libraries can be hybridized to and ligated with the puromycin linker. While our initial puromycin linker designs were not successful or efficient in producing display molecules, designs were eventually developed that routinely resulted in detectable amounts of display protein. The successful designs are illustrated in Figure 4-15 and Figure 4-17.

The first linker design used was based off of a puromycin linker design presented and successfully implemented by Shingo Ueno from the Nemoto group[45], and is illustrated in Figure 4-13.

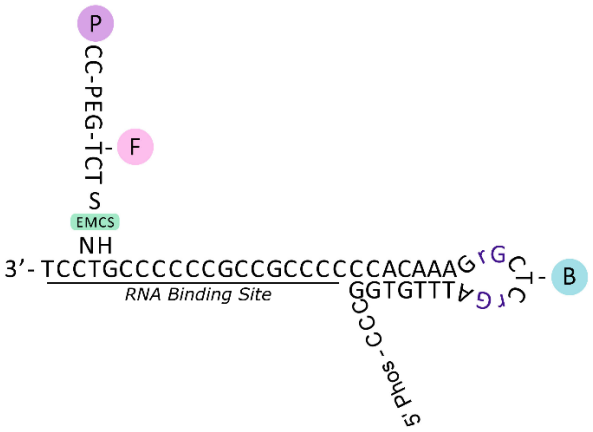


Figure 4-13 Initial Puromycin Linker Design
The puromycin linker is composed of two components, the DNA hairpin and the puromycin arm. The DNA hairpin has a single-stranded region that hybridizes with a specific ligation sequence in the mRNA. Hybridization results in a recessed 3' DNA end which serves as a primer in the reverse transcription of the mRNA. Several adenine bases on the 3' end of the bound mRNA template are un-hybridized and available for single-stranded ligation with the free cytosine bases on the 5' end of the DNA hairpin. The loop region contains a biotin (blue circled B) flanked by two single stranded ribo-guanines. These ribo-guanines can be selectively cleaved to release the biotin. The puromycin arm is a flexible linker containing three hexa-ethyleneglycol repeats (PEG), a fluorescein (pink circled F), and a puromycin (purple circled P).

This design, however, was never successful in generating a display protein library. Due to the unstable nature of ribonucleotides, a second design was tested; similarly based off a design first proposed by Shingo Ueno[45] which utilizes a guanine analogue, inosine, that can be recognized by the endonuclease Endo V to create a specifically cleavable loop. This design is depicted in Figure 4-14, and demonstrated a very minimal level of display.

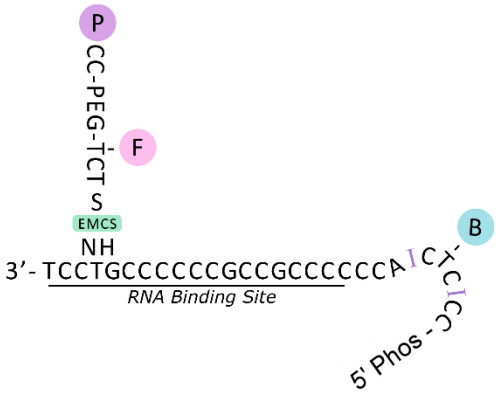


Figure 4-14 Inosine Puromycin Linker Design
The puromycin linker is composed of two components, the DNA hairpin and the puromycin arm. The DNA hairpin has a single-stranded region that hybridizes with a specific ligation sequence in the mRNA. Hybridization results in a recessed 3' DNA end which serves as a primer in the reverse transcription of the mRNA. Several adenine bases on the 3' end of the bound mRNA template are un-hybridized and available for single-stranded ligation with the free cytosine bases on the 5' end of the DNA hairpin. The loop region contains a biotin (blue circled B) flanked by two inosine nucleotides. These inosine bases can be selectively cleaved to release the biotin. The puromycin arm is a flexible linker containing three hexa-ethyleneglycol repeats (PEG), a fluorescein (pink circled F), and a puromycin (purple circled P).

To troubleshoot the display efficiencies, a linker was designed with a single fragment of nucleotides comprising both the biotin and the puromycin arm. This linker was inexpensive relative to the two arm designs, and did not require additional crosslinking and purification steps to synthesize. The fluorescein was also removed to simplify synthesis. As the length and flexibility of the puromycin arm can affect its ability to diffuse into the "A" site of the ribosomes, and therefore influence the rate of display formation, designs were obtained that contained one to five "Spacer 18" molecules, which are 18 atom configurations of hexa-ethylenelycol, or a string of adenine bases. The structure of a four spacer version is shown in Figure 4-15, and the yield from the various designs show in Figure 4-16.

Figure 4-15 Single Arm Puromycin Linker Design
The single arm puromycin linker is composed of a single DNA stand. It contains a region that hybridizes with a specific ligation sequence in the mRNA. Several adenine bases on the 3' end of the bound mRNA template are un-hybridized and available for single-stranded ligation with the free cytosine bases on the 5' end of the DNA. The loop region contains a biotin (blue circled B) flanked by two inosine nucleotides. These inosine bases can be selectively cleaved to release the biotin. The puromycin arm is a flexible linker containing four 18 atom hexa-ethyleneglycol repeats (Sp18), and a puromycin (purple circled P).

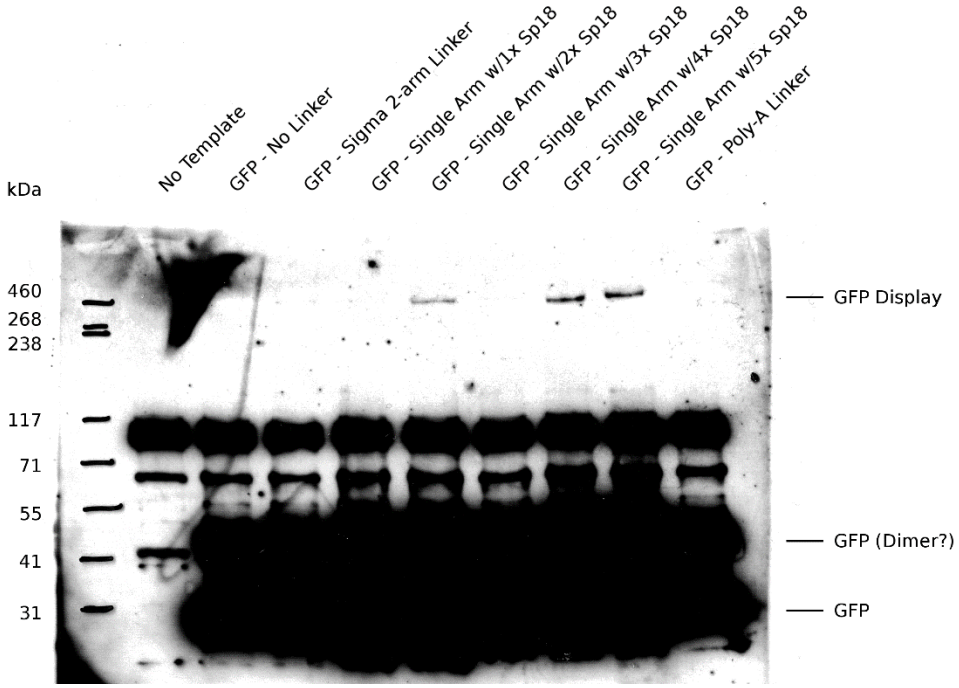


Figure 4-16 Display Complex Yields for Single-Arm Puromycin Linker Variants
Single-arm synthesized puromycin linkers with varying arm lengths were tested for their display yields. The structure of the single arm linker is given in Figure 4-15, the structure of the "Sigma" linker (used as a reference) is given in Figure 4-13. Linkers were ligated to an RNA which codes for GFP and translated in the PURExpress system. One fifth of the reaction was loaded on a Tris-Glycine PAGE gel. The gel was transferred to a membrane and blotted with an anti-FLAG antibody. The lane labels indicate the linker type and/or the number of Spacer 18 (Sp18) molecules included in the puromycin arm of the linker. The expected size of GFP protein alone is approximately 27 kDa; the expected size of a GFP display complex (protein, linker, and RNA) is approximately 350 kDa.

Because the single-arm puromycin does not contain a free 3’ DNA end that can be used as a primer for reverse transcription, a two arm design similar to the one seen in Figure 4-14 needed to be developed. Instead of using a bifunctional chemical crosslinker, which requires several steps to in order to join the two nucleic acid arms, the new two arm designs leveraged copper free click chemistry. The puromycin segment contains a 5’ DBCO, and the hairpin segment contains an internal azide. The segments were ordered independently, then conjugated in PBS and PAGE purified in house. Versions with one to five Spacer 18 repeats were ordered for testing and optimization; the four repeat version is illustrated in Figure 4-17, and the yields from translation reactions with the four different designs are shown in Figure 4-18. The four repeat version of the linker was selected for use in the protocol due to its yields.

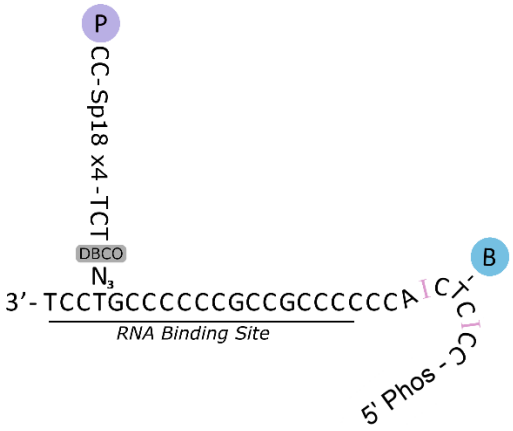


Figure 4-17 Two Arm Click Puromycin Linker Design
The puromycin linker is composed of two components, the DNA hairpin and the puromycin arm. The DNA hairpin has a single-stranded region that hybridizes with a specific ligation sequence in the mRNA. Hybridization results in a recessed 3’ DNA end which serves as a primer in the reverse transcription of the mRNA. Several adenine bases on the 3’ end of the bound mRNA template are un-hybridized and available for single-stranded ligation with the free cytosine bases on the 5’ end of the DNA hairpin. The loop region contains a biotin (blue circled B) flanked by two inosine nucleotides. These inosine bases can be selectively cleaved to release the biotin. The puromycin arm is a flexible linker containing three hexa-ethyleneglycol repeats (PEG), a fluorescein (pink circled F), and a puromycin (purple circled P).

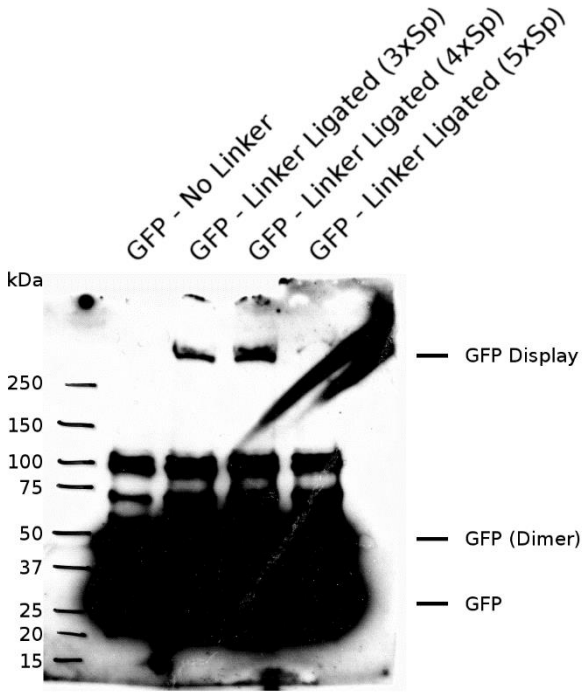


Figure 4-18 Display Complex Yields for Two-Arm Puromycin Linker Variants
Two-arm click synthesized puromycin linkers with varying arm lengths were tested for their display yields. Linkers were ligated to an RNA which codes for GFP and translated in the PURExpress system. One fifth of the reaction was loaded on a Tris-Glycine PAGE gel. The gel was transferred to a membrane and blotted with an anti-FLAG antibody. The number of spacer-18 units is indicated in the parenthesis following each sample ID. The structure of the linker is given in Figure 4-17. The expected size of GFP protein alone is approximately 27 kDa; the expected size of a GFP display complex (protein, linker, and RNA) is approximately 350 kDa.

Alternate versions of these linkers were produced with two primary variations. “No Biotin” designs lacked a biotin molecule, and “Non-Cleavable” versions contain guanine bases in place of the inosine bases; all other aspects of the puromycin linker remained the same.

The puromycin linker ligation process is the same for all the designs presented above, and occurs in two steps. The RNA and the puromycin linker are first annealed in hybridization buffer in a reaction that is initially heated to 75 ^{o}C, and slowly brought down to room temperature. This process ensures specific hybridization. Ligation buffer is then added, and T4 RNA Ligase 1 enzyme is used to ligate the overhanging single stranded mRNA to the single stranded DNA in the linker.

4.3.3.2 Puromycin Linker Ligation

Ligation of the puromycin linker to an RNA template is performed in a buffer with ideal salt conditions for nucleic acid binding. The linker is mixed with RNA template, the mixture heated, and then slowly cooled to achieve specific ligation. The two molecules are covalently bonded with a single strand RNA ligase, which joins the 3' end of the RNA to the 5' end of the puromycin linkers.

Ligation of the templates to the puromycin linker is confirmed with another high sensitivity RNA bioanalyzer, comparing the template migration before and after the ligation reaction. Because of the non-nucleic acid groups contained within the linker, it is difficult to predict how it will behave in the bioanalyzer environment. However, our ligation results are consistent, and have allowed us to experimentally determine that successful ligation of a single template results in an approximately 100 base shift in the bioanlayzer data. With a two to one ratio of puromycin linker molecules to template molecules, we typically observe ligation rates in the 40-60% range, although this seems to vary considerably with the template and the purity of the puromycin linker. A highly efficient ligation reaction is shown in Figure 4-19, to highlight the typical shift. While this shift is also present in libraries, it can be much more difficult to visualize due to the population distribution.

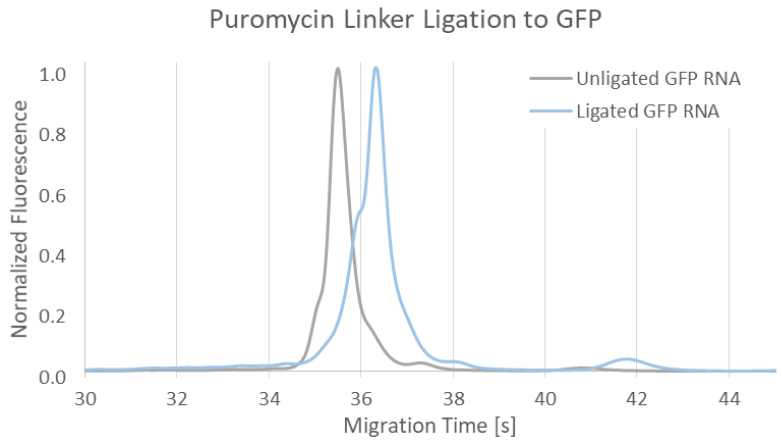


Figure 4-19 Ligation Shift in Bioanalyzer Data
Bioanalyzer RNA Pico traces for GFP RNA before and after ligation to the puromycin linker. Migration time is equivalent to fragment size, the increase in fragment size between the unligated sample and the ligated sample is about 100 bases.

To remove any linker that is unligated to RNA, T5 Exonuclease is added to the system. This enzyme digests DNA with a free 5' end. The ligated RNA is subsequently purified with a RNA purification column.

4.3.3.3 Translational System

The puromycin-RNA complexes are translated in a standard PURExpress reaction with the addition of a broad-range RNase inhibitor.

We explored the possibility of capturing puromycin-RNA complexes on a solid substrate prior to translation to improve the ratio of puromycin ligated RNA templated to unligated RNA templates. While the PURExpress system is very robust, it did not tolerate the addition of Dynabeads MyOne Streptavidin C1 beads. Several experiments were performed with and without the previously mentioned beads in a PURExpress translation reaction containing GFP template. Reactions were subject to the inclusion of 50 ugs of streptavidin C1 beads and orbital rotation at 1200 rpm with otherwise normal translation conditions. GFP fluorescence was measured in the completed reactions to assay the efficiency of each condition. The results are given in Table 4.4.

Table 4.4 Influence of Streptavidin C1 Beads on PURExpress Translation
Four standard PURExpress reactions were set up with 250 ngs of GFP DNA template. Four reactions were incubated for two hours at 37 °C; with Streptavidin C1 beads and shaking, beads and no shaking, no beads and shaking, or with no beads or shaking. The numbers indicated in table are the GFP protein yields (in ngs) detected in each condition via fluorescence.

	With Reaction Agitation	**No Reaction Agitation**
Reaction **With** Beads	14	39
Reaction **Without** Beads	417	219

The data demonstrates that the streptavidin C1 beads inhibit the translation of GFP in the PURExpress system. There is still a possibility that beads with other chemistry may be used successfully. When no

beads were present, shaking positively influenced the amount of GFP produced. Some PURExpress reactions may benefit from agitation during the translation reaction.

4.3.3.4 Post-IVT Treatment

Post-*in vitro* translation treatments have been demonstrated to be critical for improving the yield of display proteins. Salts are immediately added following the translation reaction, and the reaction is subsequently frozen. Both of these treatment have been shown to empirically improve display rates, and may influence the movement of the puromycin into the A site. Typical display yields for s single template are approximately 1% of the input RNA, although yields as high at 25% have been reported for specific systems[46].

4.3.4 Validation and Results

4.3.4.1 Display Complex Formation

Display complex formation has been successfully demonstrated for both single template and library inputs. Western blotting was used to establish that our protocol is effective in generating display complexes for single templates. Western blotting is a well-established method of protein identification and typically replies on an anti-body specific to the protein of interest. The need for such an anti-body can pose challenges, with respect to cost, specificity, and availability. Because a 3' FLAG tag has been incorporated into the protein templates, these potential pitfalls can be entirely avoided. The anti-FLAG antibody is relatively inexpensive, highly specific, and widely available. Additionally, the same anti-body can be used against all proteins in the system, reducing costs. Because the antibody is not protein specific, the size of the protein is used to confirm its identity.

GFP templates were ligated to a puromycin linker, translated in PURExpress, underwent post-translational treatment, and then were run on a PAGE gel for detection with an anti-FLAG antibody. When RNA is successfully conjugated to its precursor RNA, the complex is significantly larger in size

than the protein alone, as the nucleic acid dominates the composition of the molecular weight. Display complexes are visualized in western blotting as extremely high molecular weight bands relative to the unconjugated protein population. A sample western blot using the single-arm four spacer linker design is presented in Figure 4-20, for blots showing the optimization of the linker used in the display process see Figure 4-16 and Figure 4-18.

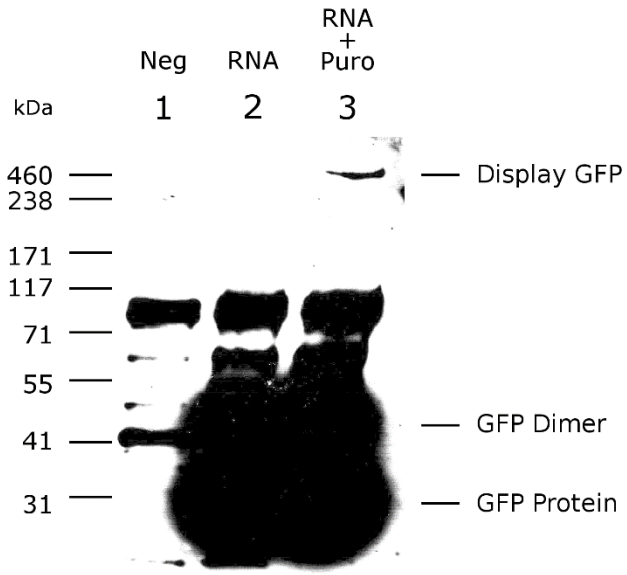


Figure 4-20 Western Blot for GFP Display Validation
PURExpress translation reactions were performed with either no template (Lane 1: Neg), RNA coding for GFP (Lane 2: RNA), or RNA coding for GFP that had been ligated to a puromycin linker (Lane 3: RNA + Puro). After post-translational processing, one fifth of the reaction was loaded on a Tris-Glycine PAGE gel. The gel was transferred to a membrane and blotted with an anti-FLAG antibody. The expected size of GFP protein alone is approximately 27 kDa; the expected size of a GFP display complex (protein, linker, and RNA) is approximately 350 kDa.

To further demonstrate the presence of the display complex, a second assay was performed that required both the biotin-containing puromycin linker and the protein to be present for detection. A post-translation display reaction with GFP RNA ligated to a puromycin linker was subject to pull-down with two different type of streptavidin magnetic beads. The complexes were removed by boiling in SDS loading buffer and then run on a PAGE gel and blotted as in Figure 4-20. The pull-down with streptavidin beads requires the presence of the puromycin linker, and the western blot requires the

presence of protein. Illustrates that this process selects for, to a different degree for different bead types, the high molecular weight product that is presumed to be the GFP display complex.

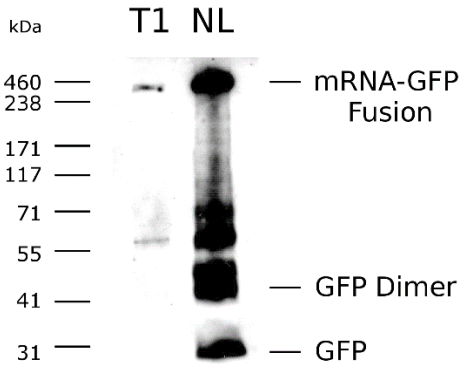


Figure 4-21 Bead Selection and Western Blot for GFP Display Validation
PURExpress translation reactions were performed with RNA coding for GFP that had been ligated to a puromycin linker. These reactions were selected with either MyOne Streptavidin T1 beads (T1) or with NanoLink streptavidin beads (NL). After release from the beads by boiling in SDS loading buffer, the reaction was loaded on a Tris-Glycine PAGE gel. The gel was transferred to a membrane and blotted with an anti-FLAG antibody. The expected size of GFP protein alone is approximately 27 kDa; the expected size of a GFP display complex (protein, linker, and RNA) is approximately 350 kDa.

4.3.4.2 Use of Display Barcode in Proxy Identification

To test whether a specific interaction can be detected by using the mRNA "barcode" on the display protein, the GFP antibody to GFP protein interaction was assayed. A small SMART-Display library was constructed as follows. Four full-length mRNAs, GFP, CKMT2, MAPKAPK2, and DHFR, were brought through the SMART-Display process. The resulting mRNA-protein fusions were mixed equimolarly to create a small SMART-Display library. To quantify each mRNA, qPCR reactions were performed on this un-selected library (pre-selection value). GFP antibody was then used to pulldown the library on magnetic beads, applying a stringent wash to remove non-specific RNA-bead attachments. Each mRNA was also quantified in the post-pulldown mixture (post-selection value) by qPCR. 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 genes (CKMT2, MAPKAPK2, and DHFR) were much smaller than that of the GFP.

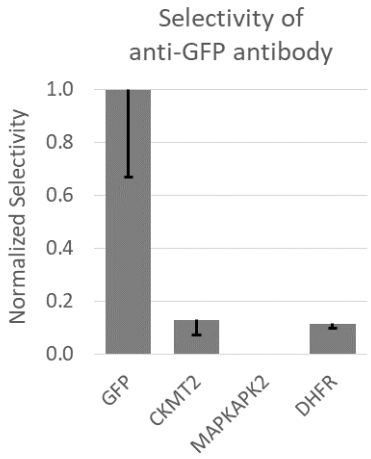


Figure 4-22 Use of the Display Nucleic Acid as a Proxy Identifier
Specificity of antibody-antigen interaction, measured by the ratio of post- and pre-selection qPCR-based mRNA quantification (y axis) for each mRNA (column). The ratios were normalized against GFP's ratio (post-/pre-selection = 1). Post- and pre-selection refer to after and before anti-GFP antibody pulldown, respectively. The ratio for MAPKAPK2 is reported as 0 because MAPKAPK2 was not detected post-selection. Error bar: standard error.

Table 4.5 qPCR Data for Use of the Display Nucleic Acid as a Proxy Identifier
Specificity of antibody-antigen interaction, measured by the ratio of post- and pre-selection qPCR-based mRNA quantification for each gene. The ratios were normalized against GFP's ratio (post-/pre-selection = 1). Post- and pre-selection refer to after and before anti-GFP antibody pulldown, respectively. "ND" indicates no detection of a gene in the qPCR assay.

	GFP	**CKMT2**	**MAPKAPK2**	**DHFR**
Pre-interaction Ct				
Average	5.51	8.10	22.21	6.10
Standard Deviation	0.14	0.12	0.05	0.09
Post-interaction Ct				
Average	27.04	32.60	ND	0.77
Standard Deviation	0.81	1.07		0.38
Selectivity Ratio	3.29E-7	4.21E-7		3.75E-8
Normalized Selectivity	1.00	0.13		0.11
Standard Error of Normalized Selectivity	3.30E-1	5.51E-2		1.79E-2

This test indicates that the specificity of a protein interaction is reflected by the quantitative changes in the mRNA “barcodes” displayed on their surface.

4.3.4.3 Display Library Diversity

Library samples contain thousands of proteins each at low concentration, making the detection of a display library using a western blot challenging. No blot has been successfully produced for these samples.

To circumvent this challenge and look at the proteins successfully displayed from a library RNA input, a pull-down strategy was employed. A puromycin linker that did not contain biotin was used in a display reaction along with Transcend tRNA. Transcend tRNA is a tRNA charged with biotinylated lysine. The biotin-lysine competes with the native lysine for incorporation into peptides, and results in biotinylated proteins. The products of this display reactions were pulled down onto streptavidin beads. The beads should capture all the biotinylated proteins in the reaction, and will only carryover RNA if it is joined to a protein via the display mechanism. To ensure no non-specific carryover occurs, the beads are washed stringently with 8M urea. The beads are then subject to library preparation for sequencing. As only the RNA joined to display proteins is carried over in the pull-down, by sequencing the RNA we recover the identity of the proteins which successfully formed display complexes.

Two controls were used to ensure that the libraries generated were specific to displayed proteins. The first control used unligated RNA as the template in the display reaction, and in the second proteins were digested during the pull-down step. This controls for any non-specific carry-over of RNA in the reaction either due to the properties of the RNA or of the puromycin linker. Libraries were generated for the positive sample along with the two controls, but the quantity of library generated by the positive sample was significantly larger than in either of the controls, as seen in Figure 4-23, so we can conclude that there is very little non-specific signal in our experiment, and that a significant number of display complexes are being formed from the library input.

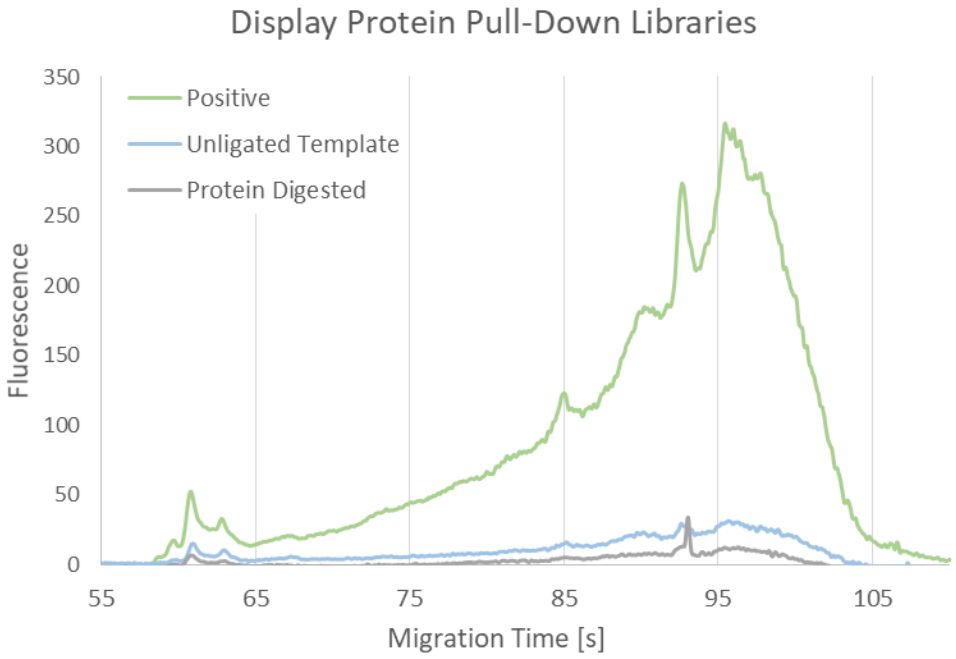


Figure 4-23 Bioanalyzer Traces for Display Protein Pull-Down Libraries
Sequencing library distribution for the SMART-Display pull-down assay. The distributions of the libraries were obtained with the High Sensitivity DNA Bioanalyzer kit. Libraries were generated for the positive sample along with the two controls, one performed without ligating the puromycin linker to the RNA library, and the other by digesting the translated library with proteinase K.

This experiment was replicated, and the libraries compared to determine repeatability. Our analysis of these sequencing reads revealed that 18,860 unique genes were detected in the SMART-Display protein library; which is approximately 1.5 times the number of genes that can be obtained from the largest ORFeome, the human ORFeome version 8.1 containing 11,149 human genes. The average fragment in the SMART-Display pull-down library contained 810 base pairs of gene content, corresponding to a protein length of 270 amino acids. The distribution of the quantity of gene content in the pull-down library is similar to the gene content in the input SMART-Display RNA library. When applying the same library preparation method directly to the input RNA library, without any selection, a slightly larger distribution is observed (Figure 4-24).

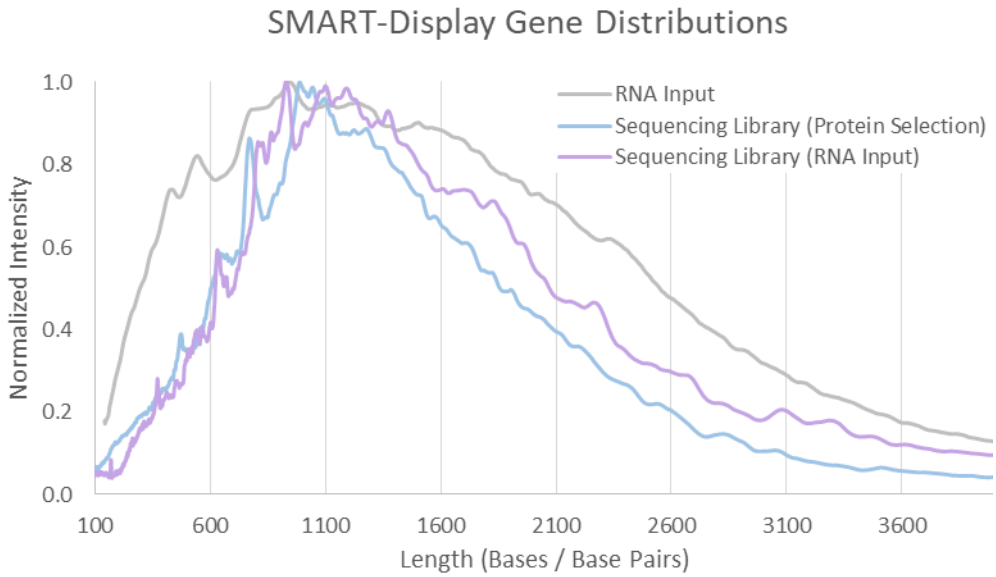


Figure 4-24 Bioanalyzer Traces for SMART-Display Libraries
The distributions of the libraries were obtained with either the RNA Pico 6000 or High Sensitivity DNA Bioanalyzer kits as appropriate. The lengths of the introduced sequences were subtracted from the observed lengths of the libraries to obtain the length of the gene content only.

This indicates that some bias for smaller fragments may be introduced by the translation and selection processes, but that the library preparation method itself is imperfect in its ability to recapitulate the RNA distribution.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Johnson, Kara; Qi, Zhijie; Wen, Xingzhao; Chen, Chien-ju. The dissertation author was the primary investigator and author of this material.

5 PROPER-Seq: High Throughput Identification of Protein Protein Interactions

The PROPER-Seq technique described here allows for the high-throughput evaluation of protein-protein interactions in an *in vitro* population of display proteins. A robust computational pipeline has been developed to assess the significance of the experimental readout, and the process has been validated in HEK, HUVEC, and Jurkat cell lines. The data was assessed for repeatability, precision, and sensitivity. The PROPER-Seq datasets from the three cell types were compared to look for cell-specific interactions, and evaluated as a whole to identify human protein interaction networks.

5.1 Aim

The aim of this technique is to identify protein-protein interactions in a population of SMART-Display proteins.

5.2 Requirements

In a library by library interaction analysis, pairwise interaction data is necessary, as there is not a single 'bait' protein, but potentially thousands. This methodology requires that the barcodes on interacting proteins be joined, and that we select the chimeric fragments for sequencing, as they contain information from both interaction partners. Informatically, both interaction partners need to be identified, and the number and frequency of interactions for each protein enumerated and tested for significance.

5.3 Approach

Protein interactions are detected as chimeric DNA fragments that are formed from the proximity ligation of the nucleic acids in the display complex of two interacting proteins. For an overview in methodology and an illustrative figure, see Section 3 and Figure 3-2.

5.3.1 Design of the Proximity Ligation Method

The first consideration in the design of the interaction protocol was how the nucleic acid barcodes from the two display proteins would be joined. Because efficiency is a significant concern given the number of steps in the protocol and the already inefficient process of SMART-Display, it was preferable to use a DNA to DNA ligation strategy over RNA to RNA or RNA to DNA which have significantly less efficiency. Of all of the DNA to DNA ligation strategies available, sticky end ligation (from a restriction enzyme digestion) is known to be the most efficient.

In selecting a restriction enzyme to be used in this application, several features were considered. One concern in a proximity based interaction protocol where molecules are stabilized on a solid surface is the discrimination between pairs that are formed due to true interaction and pairs formed due to proximity on the solid surface. Using a non-palindromic restriction enzyme site on the ends of the display nucleic acids enables reduction of ligation event between display nucleic acids without an intermediate linker containing the complementary restriction sequencing. Once the possible restriction enzymes were narrowed down to those that are non-palindromic, BbvCI was selected based on its cost, digestion rate, availability and efficiency.

An intermediate linker that joins the nucleic acid strands from interacting display proteins was implemented for two reasons. First, it enabled the use of the non-palindromic restriction enzyme which reduces the background in the experiment. Second, by adding a selectable marker to the linker, in this case biotin, the interaction linker becomes a way to select for chimeric nucleic acids in the system. This linker is ligated to the ends of the prey library before the bait and prey libraries are mixed. This encourages chimera formation only between bait and prey (not bait to bait or prey to prey), and allows any unligated biotinylated interaction linker to be washed from the system.

5.3.2 Conversion of Display RNA to DNA

The restriction enzyme site was originally introduced with the other SMART-Display sequences on the 5' end of the gene template. However, when implementing the restriction enzyme digestion on the DNA generated from the RNA in the display complexes, very low rates of digestion and subsequent ligation were observed. It seemed that in the process of converting the DNA library into RNA, and then back into DNA, the restriction enzyme site was lost or damaged. To circumvent this challenge, the RNA was converted into DNA not by the standard method, using a primer to generate cDNA and RNase H and a polymerase to generate the second strand, but with another template switching reaction. This allows for the introduction of a complete and unaltered restriction enzyme site at the 5' end of the barcode during the RNA to DNA conversion.

Converting the RNA in the display complexes to DNA using the template switching reaction demonstrated much better rates of digestion and subsequent ligation (Figure 5-1).

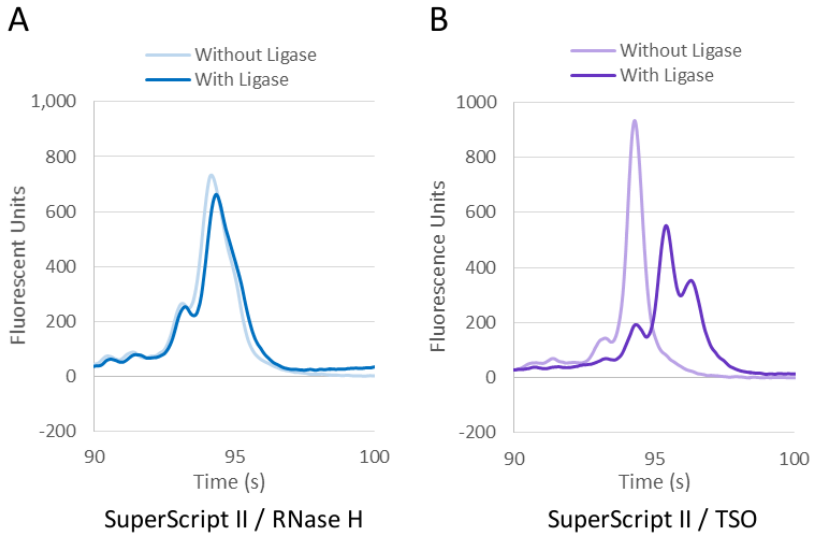


Figure 5-1 Ligation Rate with and without TSO Based Conversion of RNA to DNA
GFP RNA templates were converted to DNA and subsequently digested with BssSI and ligated to the BssSI ligation linker. A) RNA to DNA conversation using a standard cDNA synthesis followed by RNase H digestion and second strand synthesis. Ligation efficiency approximately 0%. B) RNA to DNA conversion with a TSO based approach. The ligation reaction with enzyme shows a size shift relative to the no enzyme reaction. The original GPF template and the TSO both contain digestion sites, yielding two peaks. Ligation efficiency is approximately 90%.

5.3.3 Interaction Conditions

The interaction conditions for this protocol were carefully considered. The goal would be to recapitulate the cellular environment to the greatest degree possible, while also considering ways to reduce background in the experiment. To these ends, the interaction steps are performed in dilute conditions, to reduce the chance that two proteins will be in proximity simply due to diffusion. They are also performed at physiological pH and with physiological salt concentration. Finally, interactions are crosslinked to allow for more rigorous washing and the removal of proteins that have bound non-specifically in the system. BS3 is a protein specific cross linker, and should not stabilize any nucleic acid – protein interactions or nucleic acid – nucleic acid interactions.

A future goal for the PROPER-Seq protocol would be to perform the interaction in a cell-type specific lysate – to best mimic the cellular environment.

5.3.4 Library Preparation

Once the proximity ligation is complete, all the nucleic acid material in the system is fragmented to release it from the streptavidin beads. The first half of the library prep is then performed, including end repair and adaptor ligation. This total population of nucleic acid is them subjected to a second streptavidin bead pull-down, which should enrich for fragments that contain the biotinylated interaction linker, and are therefore more likely to be chimeric and representative of an interaction event. The selected fragments are then amplified on the beads, and the resulted libraries are ready for sequencing.

5.3.5 Controls

Initial PROPER-Seq experiments were performed with two experimental controls, ‘No Linker’, and ‘No Bait’. The ‘No Linker’ control omits the biotinylated interaction linker, and therefore the formation of chimeric fragments should be prevented. The lack of biotin interaction linker in this sample should mean that any carry-over into the library generation represents non-specific background.

Similarly, the 'No Bait' sample omits the bait protein library. As both of these conditions prevent any protein interactions that should result in chimera formation, libraries generated from these samples represent non-specific background. The steps in which the controls varied from the positive are illustrated in (Figure 5-2).

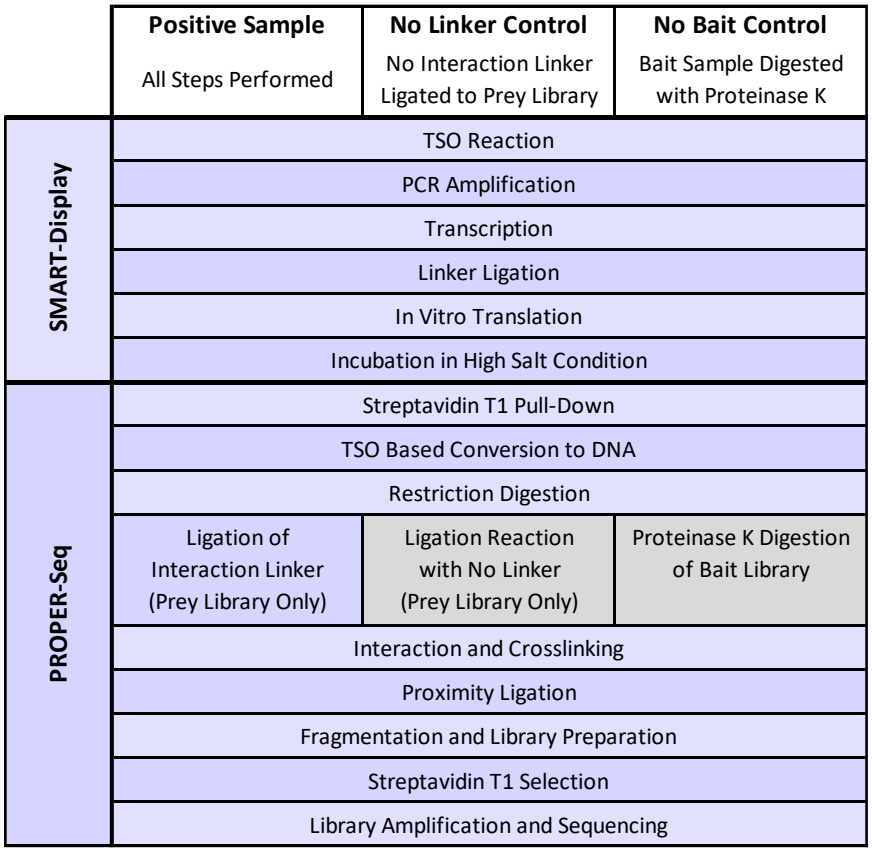


Figure 5-2 Workflow of PROPER-Seq Controls
A visual chart of the steps of PROPER-Seq that were adjusted to generate the technical controls.

5.3.6 Summary of Optimizations

There were many steps in the PROPER-Seq protocol where different experimental variations were considered and tested. A brief summary of these efforts is presented in Table 5.1.

Table 5.1 Summary of PROPER-Seq Optimizations
This table indicates the experimental variations that were tested for different PROPER-Seq steps.

PROPER-Seq Step	Variation	Observed Results
Streptavidin Pull-Down	The percent of linker ligated RNA pull-down on to the beads was measured for T1, C1 and NanoLink streptavidin magnetic beads.	T1 and NanoLink streptavidin magnetic beads demonstrated similar levels of pull-down, while the C1 beads demonstrated about 50% of the efficacy.
RNA to DNA Conversion	Conversion from RNA to DNA was performed either by 1) reverse transcribing with SuperScript II, digesting with RNAse H, and synthesizing the second strand with DNA Polymerase 1 or by 2) reverse transcribing with SuperScript II, performing template switching, and synthesizing the second strand with DNA.	Both methods demonstrated similar yields of double stranded DNA, but the template switching approach resulted in better capture of the 5' end of the RNA template.
RNA to DNA Conversion	The efficiency of the reverse transcription reaction was tested both with and without a heating an annealing step prior to the reverse transcription.	The additional heating and annealing step had no impact on DNA yields.
Restriction Digestion	The non-palindromic restriction enzymes BssSI and BbvCI were both tested for their digestion and subsequent ligation efficiencies.	When templates were no immobilized on beads, both restriction enzymes demonstrated similar digestion and ligation efficiencies. When templates were immobilized on with T1 or NanoLink streptavidin magnetic beads, the BbvCI demonstrated slightly better ligation rates.
Restriction Digestion	Both enzymes considered were tested on a cDNA/RNA hybrid template for digestion efficiency.	Neither enzyme demonstrated activity on the cDNA/RNA template.
Restriction Digestion	The BbvCI enzyme was tested for its digestion and subsequent ligation activity on T1 and NanoLink streptavidin magnetic beads.	The enzyme demonstrated slightly higher ligation rates on the T1 streptavidin magnetic beads.

Table 5.1 Summary of PROPER-Seq Optimizations, Continued

PROPER-Seq Step	Variation	Observed Results
Restriction Digestion	Ligation rates were evaluated when templates contained a restriction enzyme site in both the RNA template and TSO, and when they contained a site in the TSO only.	No significant differences in ligation rates were observed between the presence of a single or two digestion sites.
Proximity Ligation	Strategies for ligation that were evaluated included A to T ligation, blunt ended ligation, and restriction enzyme/sticky end ligation.	Templates showed the highest ligation rates for restriction enzyme/sticky end ligation (at almost 100%), followed by blunt ligation (~50%) and A to T ligation (~33%).
Proximity Ligation	The volume, concentration of ligase, and concentration of interaction linker were varied.	The concentration of ligase and the concentration of linker seemed to have little impact on ligation efficiency. Increase the reaction volume (6.6 uLs reaction volume / uL of beads from 1.6 uLs reaction volume / uL of beads) seemed to have a slight, positive impact on the ligation yield.
Streptavidin Selection	A high concentration (2M) NaCl wash was compared to a high concentration (4M) urea wash.	While both wash conditions maintained the relative relationship between sample yields, the urea wash significantly reduced the total signal.
Sequencing Length	PROPER-Seq libraries were sequenced and analyzed using 100 bp paired reads and 150 bp paired reads.	There were no significant differences in the library statistics using the different reads lengths.

5.4 Results and Validation

PROPER-Seq libraries were generated for three different cells lines: HEK 293T, HUVEC and Jurkat. Two technical replicates were performed for each cell type, experimental controls were also

performed in duplicate for the HEK experiments. All informatics processing was designed in collaboration with and carried out by Zhijie Qi, network analysis was performed in collaboration with Zhijie Qi and Xingzhou Wen. To see detailed methodology, including the informatics processing and equations, please see Appendix A.

5.4.1 Experimental Features of PROPER-Seq Libraries

The HEK experiments were performed with experimental controls. These controls were design such that, when the experiment is successful, the amount of library after the final biotin selection should be less in the controls. Plots of library enrichment during this step confirm enrichment of the positive library relative to the controls, and Bioanalyzer traces illustrate that positive libraries are at higher concentrations than their respective controls, and have a slightly larger average size (Figure 5-3).

The resulting sequencing datasets were also checked for their features, including mapping rate and number of chimeric reads identified. The HEK replicate libraries are referred to here as HEK1 and HEK2. HEK1 resulted in 343,861,373 read pairs and HEK2 produced 248,657,713 read pairs; of these, millions were determined to be chimeric (

Table 5.2). In both replicates, significantly more chimeras, and therefore protein interactions, were detected in the positive libraries than in either of the controls.

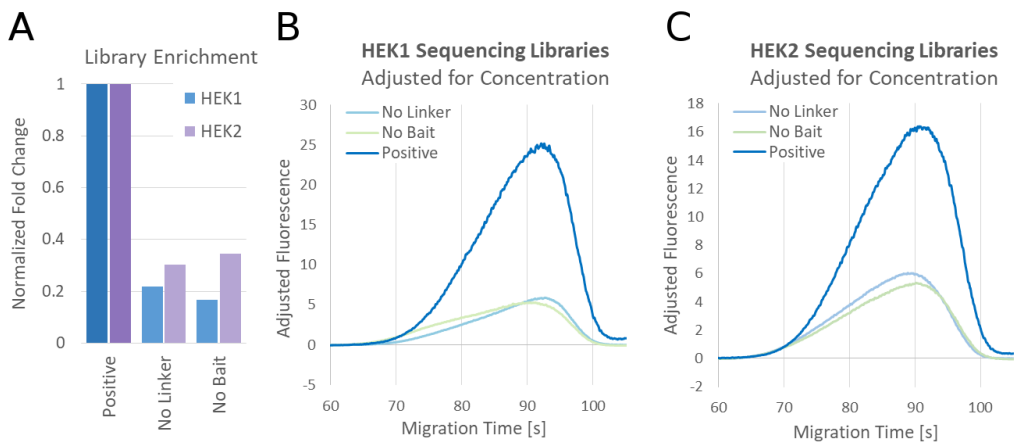


Figure 5-3 Distributions of HEK PROPER-Seq Sequencing Libraries
A) Bars show relative enrichment of the positive library over the controls; calculated by dividing the amount of library after biotin selection by the amount present before, and normalizing all those values to the positive library. B and C) The sequencing library distributions as measured by Bioanalyzer for HEK1 and HEK2.

Table 5.2 Read statistics for HEK PROPER-Seq Libraries
This table shows the type, total reads, protein coding mapped reads, mapping rate, and chimeric reads for each of the HEK PROPER-Seq sequenced libraries.

Library ID	Library Type	Total Reads	Protein Coding Mapped Reads	Mapping Rate	Chimeric Reads
HEK1	Positive	343,861,373	205,881,483	59.87%	12,581,208
HEK1_noLinker	No Linker Control	69,732,544	42,197,977	60.51%	2,152,085
HEK1_noBait	No Bait Control	87,444,917	41,828,629	47.83%	1,766,424
HEK2	Positive	248,657,713	173,300,648	69.69%	7,747,982
HEK2_noLinker	No Linker Control	97,353,678	64,671,472	66.43%	2,462,181
HEK2_noBait	No Bait Control	64,497,521	46,428,119	71.98%	2,237,573

Two PROPER-Seq experiments were carried out in duplicate on Jurkat cells (JKT1 and JKT2) and on HUVEC cells (HUVEC1 and HUVEC2). As with the HEK experiments, millions of chimeric reads were identified from 100s of millions of total reads (Table 5.3)

Table 5.3 Read statistics for Jurkat and HUVEC PROPER-Seq Libraries
This table shows the type, total reads, protein coding mapped reads, mapping rate, and chimeric reads for each of the HEK PROPER-Seq sequenced libraries.

Library ID	Library Type	Total Reads	Protein Coding Mapped Reads	Mapping Rate	Chimeric Reads
JKT1	Positive	444,413,111	262,211,890	59.00%	9,988,056
JKT2	Positive	390,643,931	236,283,970	60.49%	9,385,745
HUVEC1	Positive	359,807,741	194,690,153	54.11%	6,404,274
HUVEC2	Positive	483,597,124	283,434,465	58.61%	9,705,398

5.4.2 Reproducibility

HEK1 and HEK2 share 34,244 protein protein interactions, 66.35% of the total number of interactions. The odds ratio resulting from overlapping HEK1 with HEK2 is 14242.22. This significant overlap suggests that the protein-protein interactions generated by PROPER-seq are highly reproducible (P-value<1e-20, Fisher's exact test). As a reference, two yeast-two-hybrid dataset were compared, HuRI and HI-II-14, and have an odds ratio of 2230.89 (see section 5.4.3.1 for more information on these datasets). The odds ratio between HEK1 and HEK2 increases from ~10^4 to ~10^6 when increasing the positive read count threshold from 4[X] to 20[X]. This suggests that the reproducibility of PROPER-Seq improves at higher confidence levels (Figure 5-4).

In the Jurkat replicates, 25,306 interactions are shared with an odds ratio of 20561.09; in HUVEC, 12,765 interactions are shared with an odds ratio of 14317.28. This again suggests good scale and reproducibility of the PROPER-Seq data. For both the Jurkat and HUVEC libraries, the odds ratio increases from ~10^4 to ~10^7 with increased positive read count thresholds from 4[X] to 40[X] (Figure 5-5).

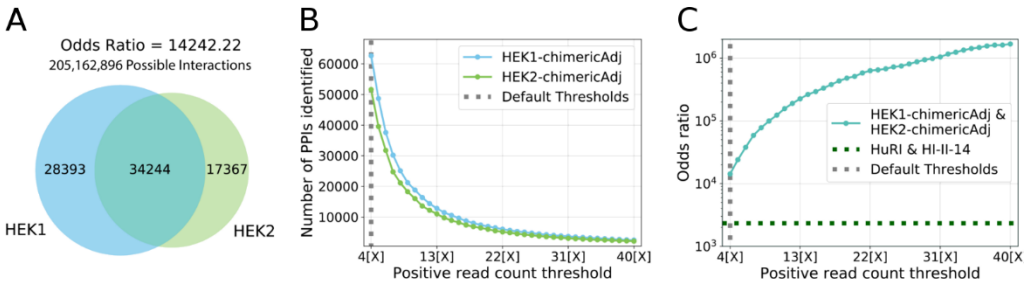


Figure 5-4 Reproducibility of the HEK PROPER-Seq Replicates
A) A venn diagram illustrating the overlap of identified protein-protein interactions between the two HEK PROPER-Seq replicates. B) The number of protein-protein interactions for each replicate as the positive read count threshold is varied. C) The odds ratio for replicate overlap as the positive read count threshold is varied. Odds ratio calculated for the overlap of HuRI and HI-II-14 is shown as a reference.

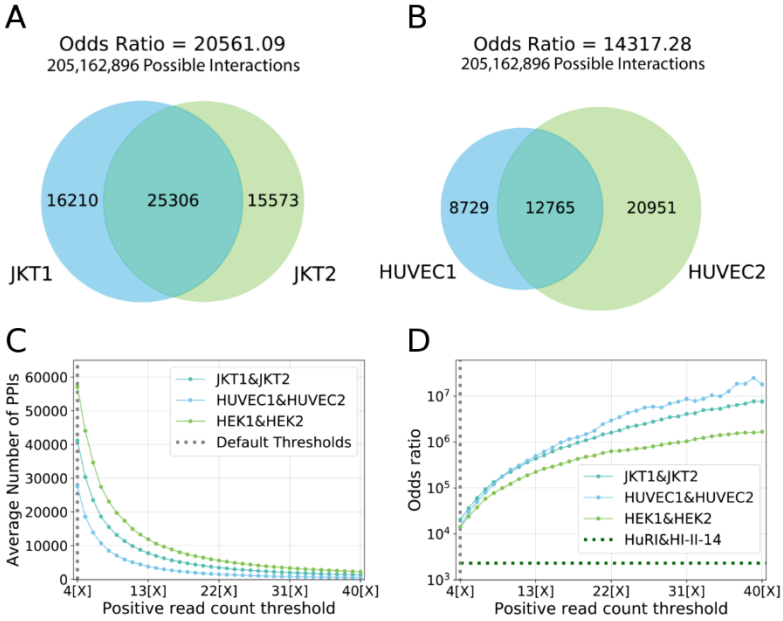


Figure 5-5 Reproducibility of the Jurkat and HUVCEC PROPER-Seq Replicates
A, B) A venn diagram illustrating the overlap of identified protein-protein interactions between the Jurkat and HUVEC PROPER-Seq replicates. C, D) The odds ratio for replicate overlap as the positive read count threshold is varied. Odds ratio calculated for the overlap of HuRI and HI-II-14 is shown as a reference.

5.4.3 Precision and Sensitivity

5.4.3.1 Overview of "Reference" Protein Interactions

While there are not gold standards for protein interactions, there is significant literature available detailing the various protein interactions that have been elucidated from numerous experimental approaches. In the various comparisons that have been made across these datasets, it is clear that there is considerable variability in the protein interactions identified by each experimental approach. To benchmark the PROPER-Seq technology and to assess its precision and sensitivity, the PROPER-Seq data sets were compared to both a databases of protein interactions (from multiple sources) and to a protein interaction network generated from a single experimental approach.

The Agile Protein Interactomes Data Server (APID)[47] is an integrated database of experimentally verified protein-protein interactions unified from BIND, BioGRID, DIP, HPRD, IntAct and MINT. A total of 367,739 interactions were available upon the writing of this document. To illustrate the variability in protein interaction detection, interactions identified by the four most prevalent techniques in our literature references were compared; affinity purification-mass spectrometry (AP-MS), two-hybrid, co-immunoprecipitation (co-IP) and liquid chromatography-mass spectrometry (LC-MS). As illustrated in Figure 5-6, interactions generated from different techniques have little overlap with each other, and only 328 PPIs are shared across all four techniques.

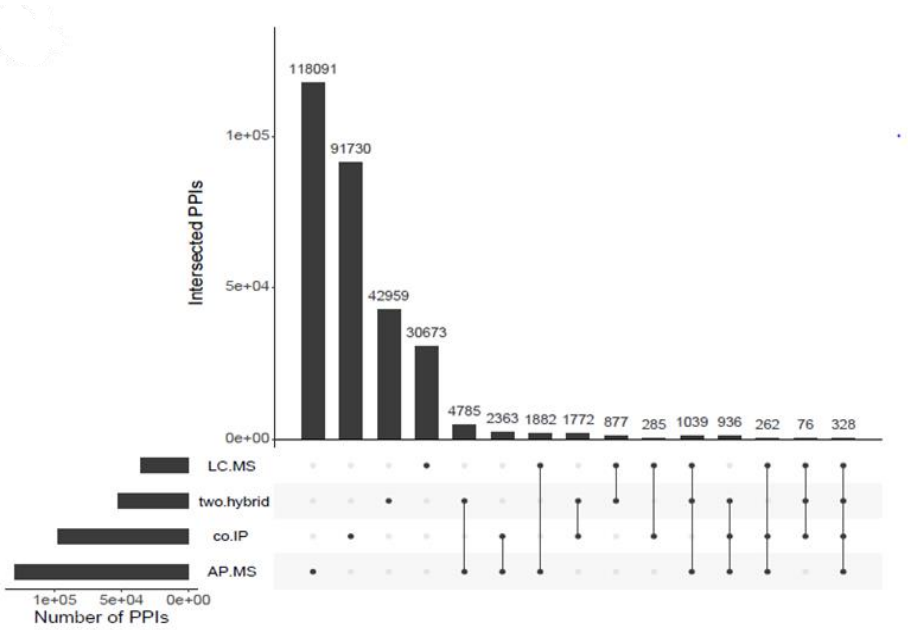


Figure 5-6 Protein Protein Interaction Overlap between Techniques
An UpSet plot illustrating the overlap in protein protein interactions from the four most represented techniques in the APID database.

The Human Reference Interactome (HuRI)[48] is large protein interactions dataset generated by yeast-two-hybrid screening containing just over 76,000 interations. The protein interactions from the HuRI dataset were compared to the APID database to evaluate the ability of a single protein interaction dataset to capture interaction data. Interactions in the APID database that were established by yeast-two-hybrid were removed from this analysis to reduce bias. The resulting precision and sensitivity of the HuRI dataset against the APID database is relatively low, at 3.98% and 0.74% respectively. When HuRI is compared against the technical subsets of APID, precision is even lower and sensitivity is about equivalent (Figure 5-7).

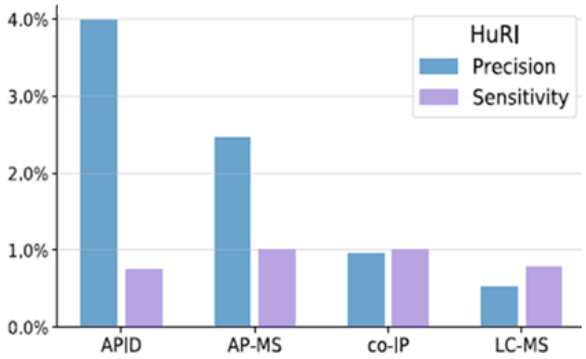


Figure 5-7 Precision and Sensitivity of HuRI against APID
The percent precision and sensitivity of the HuRI database when compared to the whole and to subsets determined by the technique used.

5.4.3.2 Adjusting for Background Using Experimental vs. Internal Controls

The two HEK libraries and their controls were used to determine the most appropriate manner in which to remove noise in PROPER-Seq libraries. For each gene pair that appeared in the positive library, a chi-squared test, an odds ratio threshold and positive read count threshold were applied to identify significant protein-protein interactions. The chi-squared test and odds ratio cutoff identify protein pairs for which the rate of co-identification is higher than expected by random protein paring. The chi-squared test was performed two ways to ensure the significant signal was robust, by considering only the background in the positive sample (chimericAdj) or using the background from the two experimental controls (controlAdj). For further discussion on the two methods, please see Appendix A. The HEK1 and HEK2 libraries were merged such that the two strategies could be applied and compared. HEK-chimericAdj yielded 109,539 significant interactions at default thresholds, while HEK-controlAdj yielded 86,338. These two populations heavily overlap; as the positive read count threshold is increased, the percentage of interactions shared between HEK-chimericAdj and HEK-controlAdj converges to 100% (Figure 5-8B).

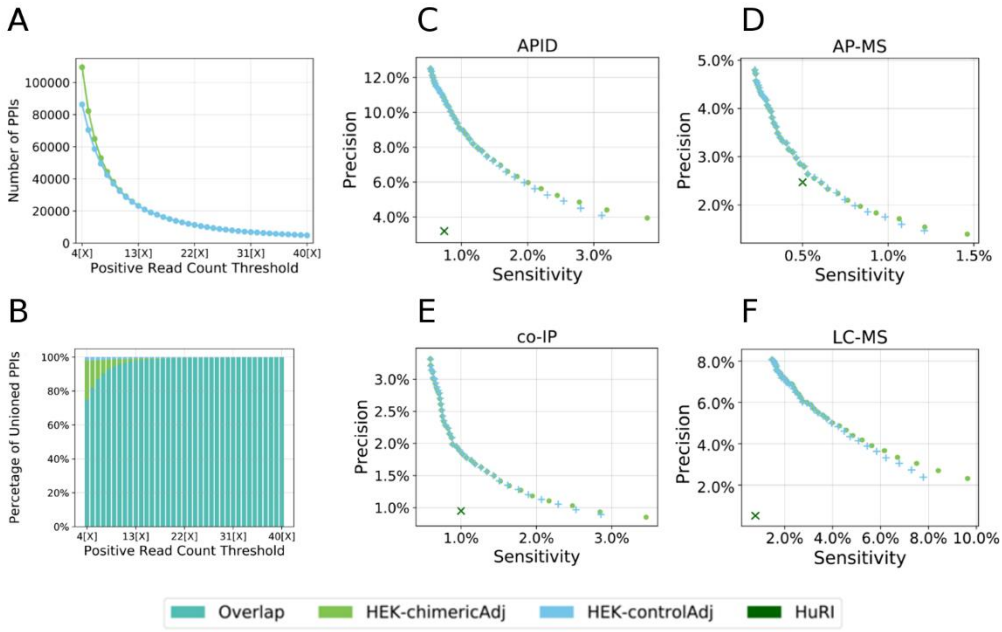


Figure 5-8 Comparison of Noise Reduction Methods in HEK libraries.
A) With maximum adjusted p-value kept at 0.05 and minimum odds ratio kept at 1, the number of protein protein interactions identified from merged HEK libraries (Y-axis) at different positive read count thresholds (X-axis) for the 'chimericAdj' and 'controlAdj' methods. B) Percentage of overlapped interactions (Y-axis) between HEK-chiemricAdj and HEK-controlAdj for different positive read count thresholds (X-axis). C-F) Precision-recall curves of HEK-chimericAdj (green) or controlAdj (blue) protein protein interactions against all APID interactions, APID AP-MS identified interactions, APID co-IP identified interactions, and APID LC-MS identified interactions. Precision-recall of HuRI against each of these dataset is illustrated as a dark green cross.

To determine if HEK-controlAdj performs better than HEK-chimericAdj in capturing known protein interactions, the positive read count threshold was varied from 4[X] to 40[X] and the associated precision-recall (PR) curves from analysis against the APID database and subsets plotted. For all four of the interaction groups, the PR curves for HEK-controlAdj are above the PR point of HuRI, but are slightly lower than the PR curves for HEK-chimericAdj. This suggests that HEK-chimericAdj better captures known protein protein interactions than HEK-controlAdj (Figure 5-8C-F). For all other library analysis, the chimericAdj method has been applied.

5.4.3.3 PROPER-Seq Precision and Sensitivity in Other Cell Lines

JKT1 and JKT2 were merged into one PROPER-Seq library (JKT) and 72,409 total significant protein protein interactions were identified from 835,057,042 read pairs using default thresholds. HUVEC1 and HUVEC2 were also merged into one PROPER-Seq library (HUVEC), and 51,125 interactions identified from 843,404,865 read pairs using default thresholds. PR curves were generated from both JKT and HUVEC by varying the positive read count threshold from 4[X] to 40[X] and compared with the APID database and its technical subsets. For both JKT and HUVEC, the PR curves lie above the reference PR point of HuRI in all four cases (Figure 11). This indicates that PROPER-Seq captures protein protein interactions in multiple cell lines as efficiently, or more efficiently, than the yeast-two-hybrid method used to generate HuRI.

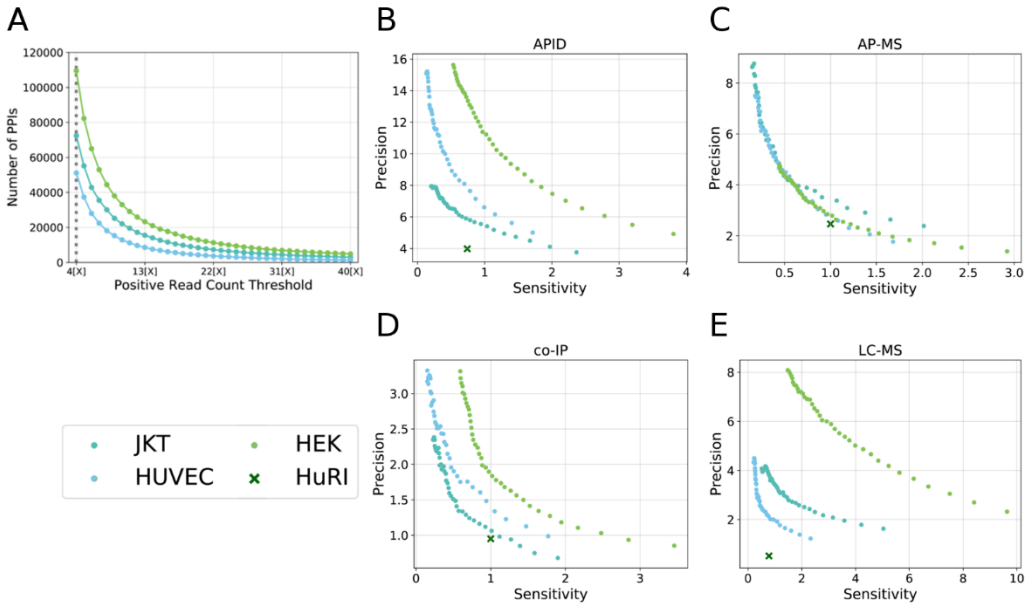


Figure 5-9 Precision and Sensitivity of PROPER-Seq Datasets
A) With maximum adjusted p-value kept at 0.05 and minimum odds ratio kept at 1, the number of protein protein interactions identified from PROPER-Seq libraries (Y-axis) at different positive read count thresholds (X-axis). B-E) Precision-recall curves of PROPER-Seq libraries against all APID interactions, APID AP-MS identified interactions, APID co-IP identified interactions, and APID LC-MS identified interactions. Precision-recall of HuRI against each of these dataset is illustrated as a dark green cross.

5.4.4 Validation of Novel PROPER-Seq Interactions *In Vivo*

To demonstrate that the PROPER-Seq *in vitro* results have implications for *in vivo*, interactions from the HEK PROPER-Seq data were validated by Proximity Ligation Assay (PLA). In a PLA assay, fluorescent signal is observed in fixed and immunostained cells when the target proteins are in very close proximity[49]. PLA has become a standard for detecting protein interactions in cells. PARP1 interactions detected in both HEK libraries above the 8[X] threshold were targeted, selecting one previously reported PARP1 interactor (SUMO1) and three novel interactors (XPO1, MATR3, and IPO5). GFP was selected as an experimental negative control. HEK cells were fixed, and the pairwise interactions queried with the Duolink PLA kit (Table 5.4, Figure 5-10). The PLA images show signal in all the samples stained with both the PARP1 and the interactor antibody, except GFP, over the single and no antibody controls. The signal strength does vary from pair to pair. As expected based on the known nuclear localization of PARP1, almost all of the PLA signal is constrained to the DAPI-stained nuclear area. These results indicate an in vivo interaction in HEK cells for all four of the PROPER-Seq identified protein pairs, with a particularly evident interaction between PARP1 and XPO1.

Table 5.4 Positive Cell Counts in PLA Assay
This table quantifies the positive cells in the PLA assay. A cell was considered positive if 5 or more fluorescent foci were identified. Antibody pairs are given by the row and columns, and the table entry indicates: the number of positive cells / total cells imaged (percent positive cells).

PLA Antibodies	No Antibody	mPARP1	rPARP1
No Antibody	0 / 23 (0%)	0 / 10 (0%)	0 / 11 (0%)
SUMO1	0 / 19 (0%)	9 / 11 (82%)	
XPO1	1 / 10 (10%)	15 / 15 (100%)	
MATR3	0 / 18 (0%)	22/23 (96%)	
IPO5	0 / 18 (0%)		9 / 11 (82%)
GFP	0 / 7 (0%)	0 / 8 (0%)	

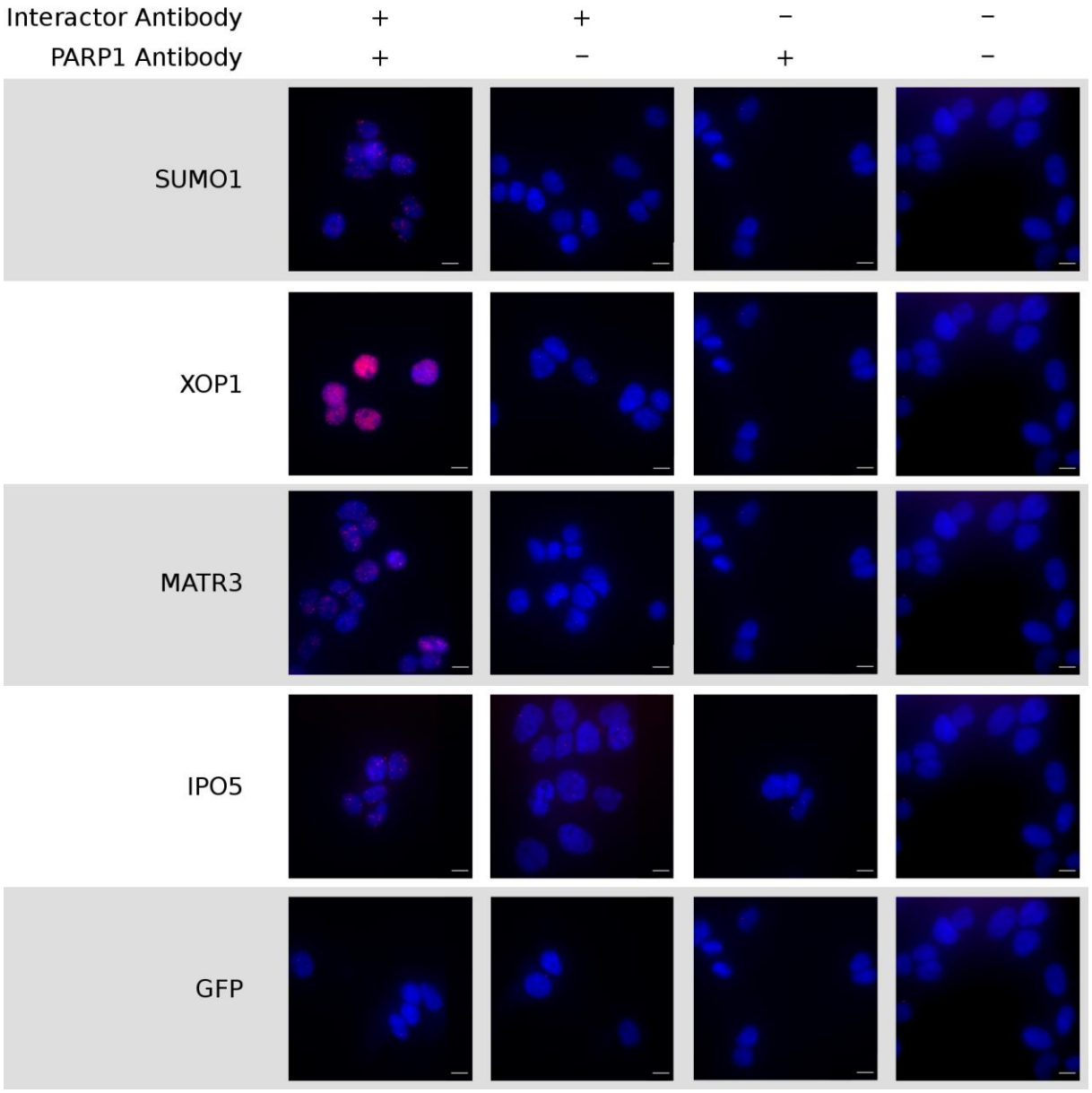


Figure 5-10 PLA Assay of Novel PROPER-Seq Identified Interactions in HEK cells
Fixed HEK cells were stained with the Duolink ® PLA Red kit with antibody combinations as shown. Fluorescent signal is illustrated in red, DAPI staining of the nucleus is shown in blue.

Co-immunoprecipitation (co-IP) validation was attempted for many of the interactions validated by PLA. PARP1 was immunoprecipitated from reversibly crosslinked HEK cell lysate, and the eluate probed for a given interactor using the Simple WES capillary detection system. However, as these co-IP experiments were performed in cells expressing native quantities of protein (not in over expression systems), obtaining a co-IP signal was challenging. A weak signal was obtained only for the PARP1-XPO1 protein pair, which is in line with the observation from the PLA experiment, in which the PARP1-XPO1 pair demonstrated the strongest signal. The synthetic western blot generated by the WES system is given in Figure 5-11, the WES generated quantitative distribution of the signal between targets is shown in Table 5.5. Because this pair is already at the threshold for detection in the WES system, it is not necessarily expected that positive co-IP results would be able to be obtained from the pairs with fewer interactions per cell.

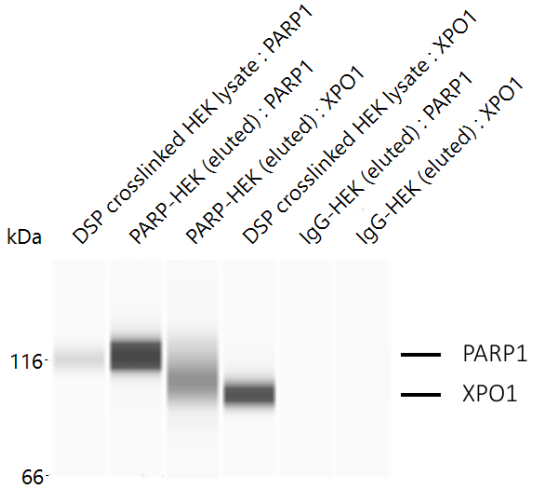


Figure 5-11 WES Detection of PARP1 and XPO1 in co-IP
The WES capillary based western blotting system was used to detected and quantify the presence of targets in a co-IP experiment. Lane labels are structured as sample : antibody target. The DSP crosslinked HEK lysate represented the unselected HEK protein lysate. Target-HEK (eluted) indicates the product of a co-IP experiment where the anti-target antibody was used for protein selection. PARP1 has an expected size of 116 kDa, and is detected at that size in the HEK lysate. XPO1 has an expected size of 123 kDa, but the apparent size on the WES is approximately 110 kDa.

Table 5.5 WES Quantitative Detection of PARP1 and XPO1 in co-IP
This table indicates the quantitative WES assignment of signal in the co-IP experiments to the two potential targets. The DSP crosslinked HEK lysate represented the unselected HEK protein lysate. PARP1-HEK (eluted) indicates the product of a co-IP experiment where the anti-PARP1 antibody was used for protein selection. Shading indicates a protein peak that is consistent with the primary antibody used against that sample.

Sample	Antibody	Absolute XPO1 Signal	Absolute PARP1 Signal	% XPO1 Signal	% PARP1 Signal
DSP crosslinked HEK lysate	PARP1	0	412,964	0	100
PARP-HEK (eluted)	PARP1	133,926	1,629,881	7.59	92.41
PARP-HEK (eluted)	XPO1	221,815	0	100	0
DSP crosslinked HEK lysate	XPO1	1,137,801	0	100	0

5.4.5 Biological Protein Interaction Subnetworks

To demonstrate that PROPER-Seq data can recapitulate well studied cellular interactions, two example subnetworks are presented below. These subnetworks represent just a small fraction of those currently identified in the data. These analyses were carried out with the dataset that results from the union of the HEK, Jurkat, and HUVEC PROPER-Seq datasets. The union dataset is referred to as "PROPER".

5.4.5.1 Ribosome Complex

Translation (GO:0006412) related proteins are enriched in the whole network with a corrected p-value of 1.22E-50 (hypergeometric test, Benjamini-Hochberg correction). A total of 2520 interactions and 135 proteins are present in this subnetwork. The literature verification ratio is 47%, which is significantly above the whole network verification ratio 3.8%. This can be potentially be attributed to the stability of the ribosome complex, a high representation in the protein libraries, and abundant

literature regarding the complex. The PROPER-Seq data captures not only the macrostructure of the ribosomes complex, but also smaller units within the complex, such as the elongation factor-1 complex (Figure 5-12).

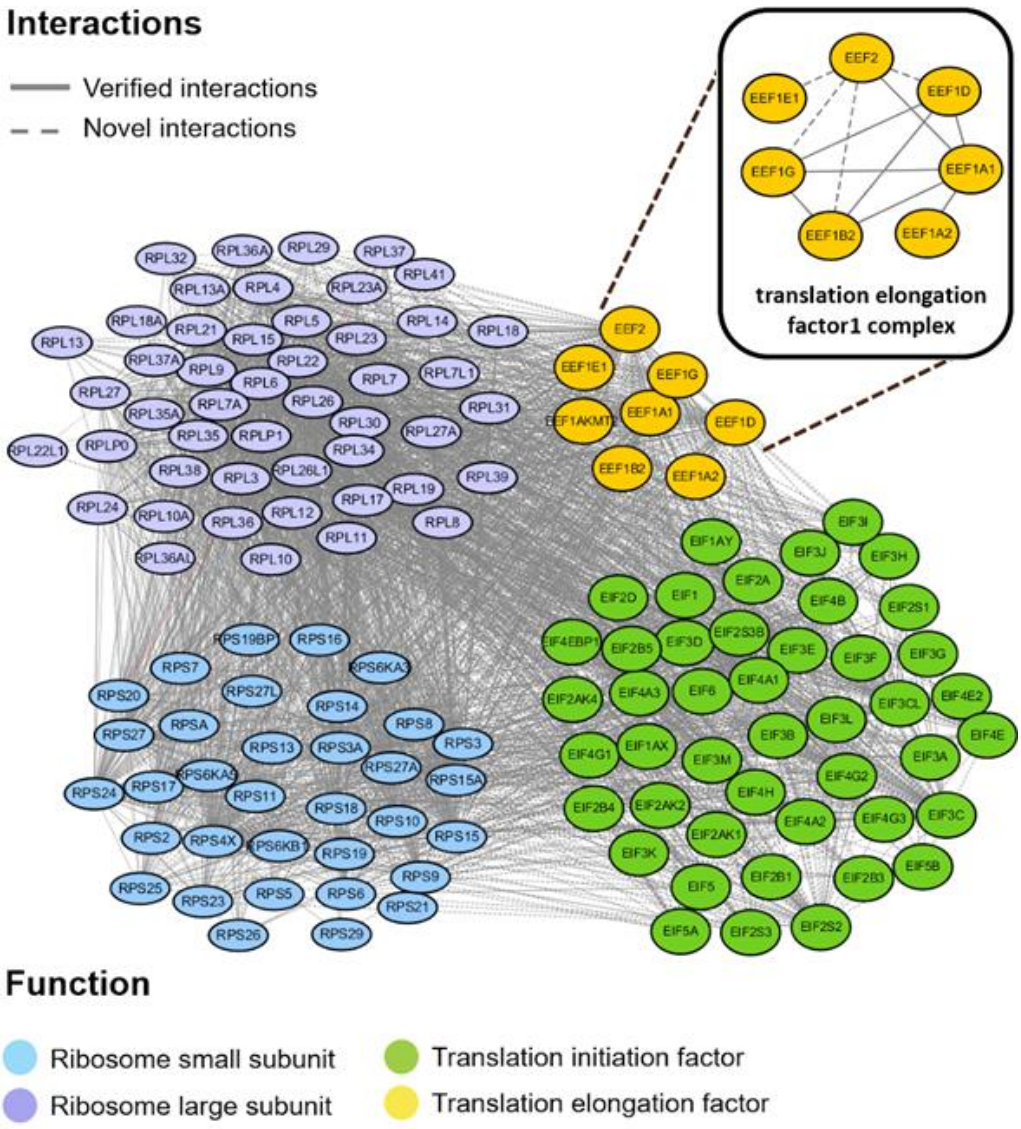


Figure 5-12 Translation Related Interaction Subnetwork
All the translation associated interactions detected above the default thresholds in the union PROPER-Seq dataset including all three cell types.

5.4.5.2 T-Complex Protein

The T-Complex Protein is a chaperonin complex containing two identical stacked rings of eight proteins. The complex assists in the folding of several proteins, including actin and tubulin. All eight of the subunits were identified in the union dataset from HEK, Jurkat, and HUVEC cells, and many of their interactions recapitulated (Table 5.6, Figure 5-13). This data highlights that interactions were better captured in some cell types than others, and that confidence in the PROPER-Seq data can be strengthened by performing multiple experiments.

Table 5.6 T-Complex Protein Interactions
This table shows the protein interactions identified in the three PROPER-Seq data sets. A "TRUE" in the cell type column indicates that the interaction was detected in that cell type above the default thresholds.

Protein 1	Protein 2	HEK	HUVEC	Jurkat
CCT3	CCT2	TRUE		
CCT5	CCT2	TRUE		TRUE
CCT8	CCT2			TRUE
CCT4	CCT3	TRUE		
CCT5	CCT3	TRUE	TRUE	TRUE
CCT6A	CCT3	TRUE		
CCT8	CCT3	TRUE		
CCT5	CCT4	TRUE		
CCT6A	CCT5	TRUE		TRUE
CCT7	CCT5	TRUE		TRUE
CCT8	CCT5	TRUE		TRUE
TCP1	CCT5	TRUE		TRUE
TCP1	CCT7			TRUE
TCP1	CCT8			TRUE

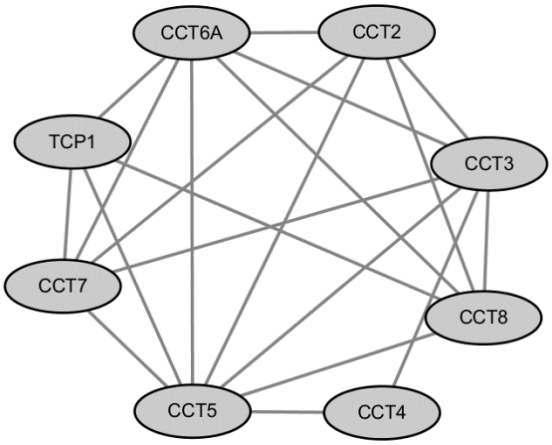


Figure 5-13 T-Complex Protein Interaction Subnetwork
All the T-Complex Protein interactions detected above the default thresholds in the union PROPER-Seq dataset including all three cell types.

5.4.5.3 CD3 Complex

The CD3 (cluster of differentiation 3) complex is a T-cell receptor involved in the TCR signaling pathway. All three subunits of this complex were identified in the PROPER-Seq data from Jurkat cells. Although none of the subunits show an interaction with each other, it is expected that if they form a complex, they would each show similar proximity interaction data with other partners. In fact, two of four identified interactions for the CD3G subunit are also are partners identified for the CD3E and CD3D subunits. 21 of the 29 and 67 interactions found for CD3E and CD3D, respectively, overlap (Figure 5-14). One of the two proteins that demonstrated interactions with all three of the CD3 subunits, HMGB1, has been previously demonstrated to have a cooperative effect with CD3 stimulation on the proliferation of T-cells[50]. It has been postulated that this effect may be mediated by the AGER receptor, but the true underlying mechanism is not known. PROPER-Seq data also shows HMGB1 interacting with both CD247 and TRAT1, other known components of the CD3/TCR complex. Further, PROPER-Seq reports an interaction between CD3 and TRAT1, which is corroborated in the literature and cited as important for stabilizing the CD3/TCR complex[51].

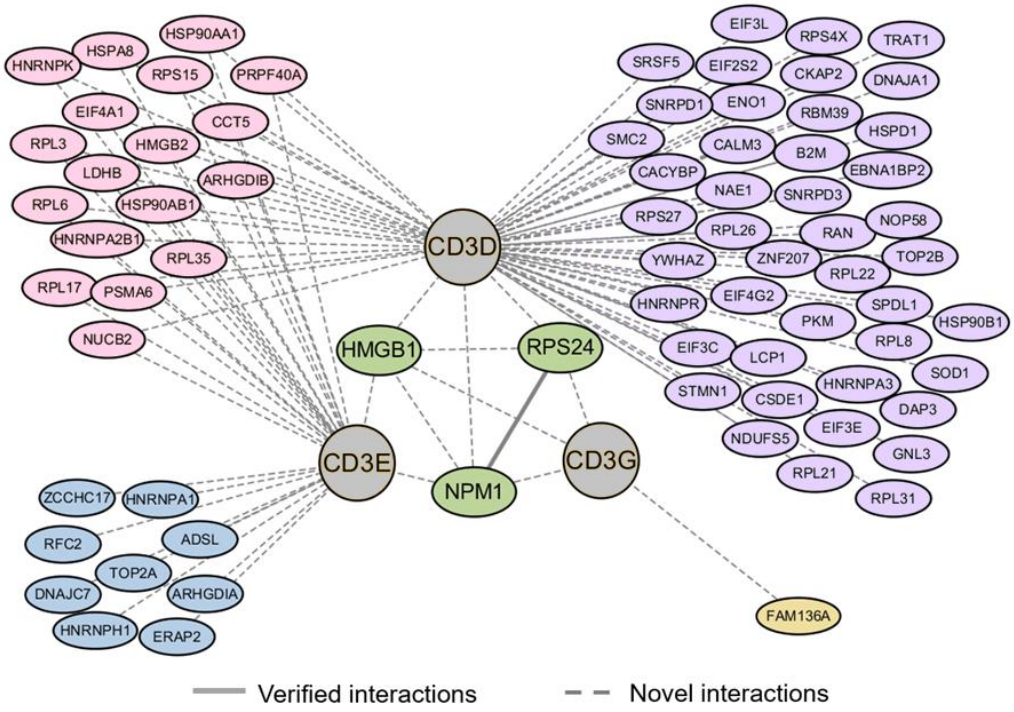


Figure 5-14 CD3 Complex Interaction Subnetwork
All the CD3 protein interactions detected above the default thresholds in the Jurkat PROPER-Seq dataset.

5.4.5.4 ESM1 Interactions

Endothelial Cell Specific Molecule 1 (ESM1) has almost 300 identified interactions in the HUVEC PROPER-Seq data, but only two identified partners in the APID database. This is somewhat surprising, as ESM1 has been found to be a highly valuable blood marker of sepsis and several types of cancer[52]. ESM1 has been reported to bind directly to ITGB2 and ITGAL, two integrin subunits that form a complete integrin heterodimer, and is implicated in the recruitment of leukocytes during the inflammatory response[53]. While neither of these two proteins was identified in the PROPER-Seq data, ITGB1 and ITGA6, which also create a complete integrin heterodimer, were found. In addition, an integrin subunit binding protein, ITGB1BP1, and a known integrin ligand, ICAM2, were identified in the ESM1 PROPER-Seq data. As the ITGA6/B1 complex is known to bind to laminins, the data was examined for laminin proteins, and LAMA4 was found to associate with ESM1 in the PROPER-Seq data. LAMA4 also demonstrated interactions with LAMB1 and LAMC1; together these laminin

subunits form the Laminin-8 complex. Interestingly, Laminin-8 has been demonstrated in the literature to be expressed in many blood cells, to be secreted upon activation of those cells, and to specifically bind to ITGA6/B1[54]. It is feasible that ESM1 is involved in a similar inflammation response with the ITGA6/B1 complex as with the ITGAL/B2 complex. This selected network of proteins is illustrated in Figure 5-15.

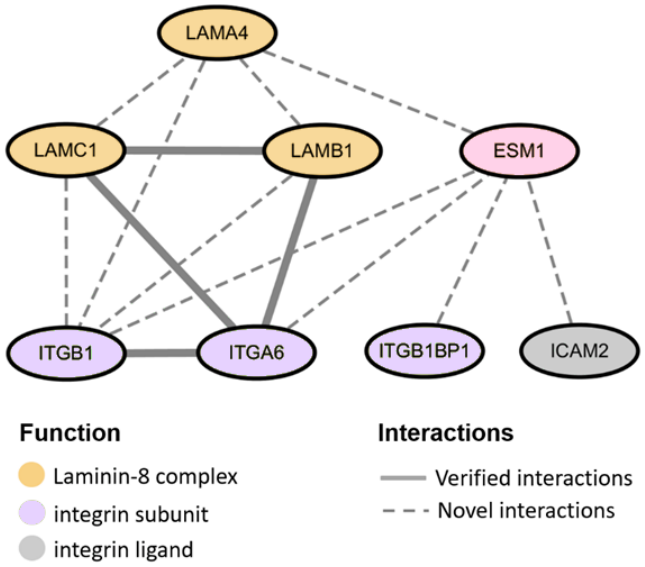


Figure 5-15 Selected ESM1 Interaction Subnetwork
Selected ESM1 protein interactions detected above the default thresholds in the HUVEC PROPER-Seq dataset.

5.4.5.5 PECAM1 Interactions

In HUVEC cells, two identified interactions for the Platelet Endothelial Cell Adhesion Molecule (PECAM1) protein were verified in the APID database. PECAM1 is involved in cellular adhesion, so it is not surprising that the two verified interaction are with Fibronectin 1 (FN1) and Integrin Subunit Beta 1 (ITGB1), both proteins known for their important roles in cell adhesion. However, the intracellular pathways through which PECAM1 acts are much less well established. Several studies have indicated cytoskeletal reorganization in response to PECAM1 stimulation or deletion[55]. The PROPER-Seq interaction data appears to support this association, as 20 of the 82 identified PECAM1 interactions are cytoskeleton related. These proteins include actin binding proteins (TPM3, TPM4, CALD1, MYL6),

intermediate filament binding proteins (DST, MACF1), and a microtubule binding protein (DYNC1LI2).

Interestingly, as the gene name suggests, there are also 11 interacting proteins that are known members of the platelet degranulation pathway. Most of these proteins are secreted during platelet degranulation process. PECAM1 is a surface protein found on many blood cells (including platelets); it would not be surprising for PECAM1 to interface with proteins secreted from surrounding platelets. In macrophages, PECAM1 has been demonstrated to be part of a recognition pathway that prevents degradation of viable cells[56]. PECAM1 recognition of excretory platelet proteins may perform a similar role.

Among these platelet degranulation proteins, we see two paralogs, APP and APLP2, which have been implicated in the biology of Alzheimer's disease. APLP2 and a second paralog, APLP1 have been postulated to have redundant function to APP, however, they do lack some functional domains found in the APP protein. A functional relationship has been established between APP and PECAM1, as APP induces transmigration of monocytes via PECAM1 signaling. This may play a role in the Alzheimer's pathology, as increased monocyte activation has been observed in the brains of patients with Alzheimer's[57].

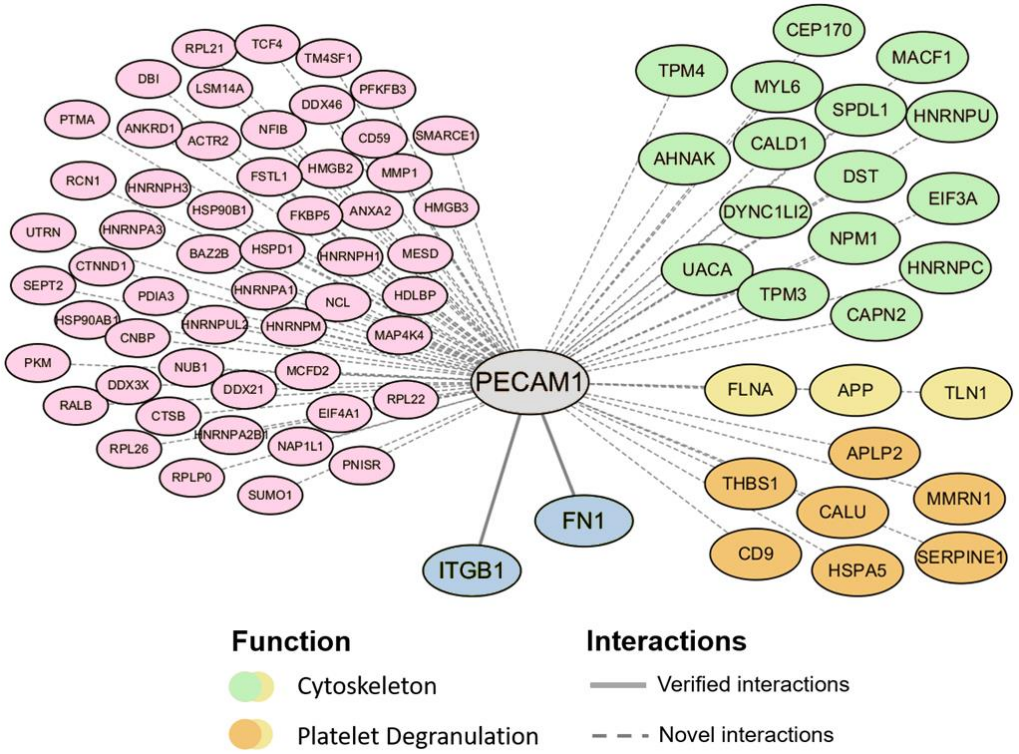


Figure 5-16 PECAM1 Interaction Subnetwork
PECAM1 protein interactions detected above the default thresholds in the HUVEC PROPER-Seq dataset.

5.4.6 Cell-Type Specific Interactions

To demonstrate that PROPER-Seq data is sensitive to cell-type specific biology, the individual cell-type PROPER-Seq datasets (HEK, HUVEC, and Jurkat) are compared to the union dataset (PROPER).

Because the input protein libraries for the PROPER-Seq experiments are derived from cell-type specific mRNA, proteins specifically expressed in certain cell lines should have differentially detected interactions when comparing cell types. The Jurkat cell line is a lineage of T-cells. When the three PROPER-Seq datsets are examined for known T-Cell markers and their interactions, there is a sharp difference of representation in the data from the three cells line, demonstrating a clear cell-type specificity (Table 5.7).

Table 5.7 T-Cell Marker Proteins in PROPER-Seq Interaction Data
This table identifies the number of interactions for T-Cell specific proteins in each of the PROPER-Seq datasets.

Protein	Interactions in Jurkat	Interactions in HEK	Interactions in HUVEC
CD3E	29	0	0
CD3D	67	0	0
CD3G	3	0	0
LCK	48	0	0
ITK	2	0	0
THEMIS	12	0	0

A similar trend was seen in markers specific to endoderm, the HUVEC cell type. Interactions involving the endodermic markers were primarily identified in the HUVEC PROPER-Seq data with very few interactions seen in the Jurkat and HEK data.

Table 5.8 Endodermic Marker Proteins in PROPER-Seq Interaction Data
This table identifies the number of interactions for endoderm specific proteins in each of the PROPER-Seq datasets.

Protein	Interactions in Jurkat	Interactions in HEK	Interactions in HUVEC
PECAM1	1	82	0
VWF	0	4	0
LIPG	0	7	0
VEZF1	0	1	2
EPAS1	0	13	3
PEAR1	0	4	0
ESM1	0	298	0
ECSCR	0	2	0
ICAM2	1	23	0

5.4.6.1 Enriched Networks

Overlap of the interactions found in each of the datasets is illustrated in the upset plot in Figure 5-17. While each of the cell-type specific datasets has significant overlap with the PROPER union set, each cell type has interactions not identified in the other cell types.

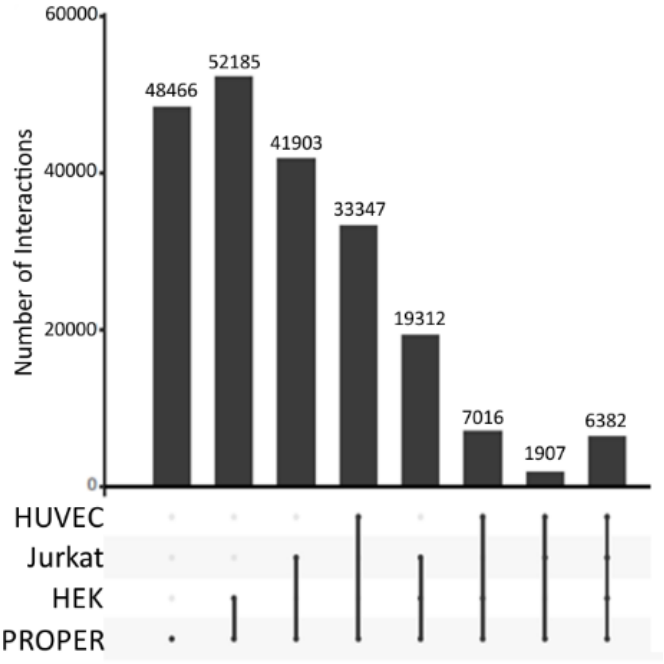


Figure 5-17 PROPER Overlap with Cell-Type Specific Data
This upset plot illustrates the overlap of the union PROPER dataset with each of the individual cell type datasets.

To help elucidate the cell type specific networks in the PROPER network, the enriched GO terms of the PROPER network were identified and the contributions from each cell type specific dataset calculated (Figure 5-18).

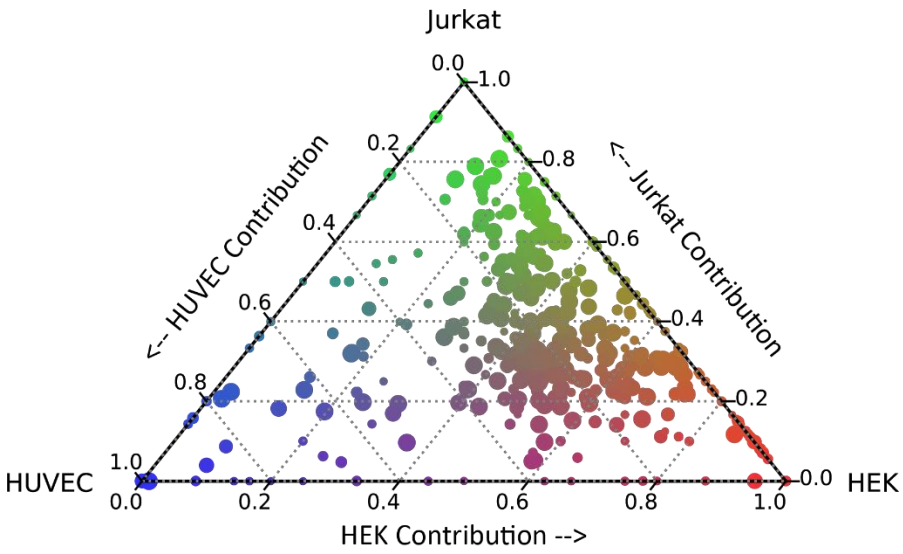


Figure 5-18 Cell Type Specific Contributions to GO Terms Enriched in PROPER
Each axis represents the relative contribution of each cell type to an enriched GO term (sum of all cell type contributions will be one). GO terms that appear near the "HUVEC" apex in blue are dominated by interactions seen primarily in the HUVEC dataset, those that appear near the "HEK" are dominated by interactions from the HEK dataset, ect.

This ternary plot clearly shows GO terms with cell type specificity. Example subnetworks are presented below in Figure 5-19. In panel A, the network for the GO term "T-Cell Receptor Signaling Pathway" is shown with the cell type specific interactions highlighted. In this network, Jurkat derived interaction clearly dominate the network (contribution = 0.80). As the Jurkat cell line is T-Cell derived, this representation aligns with the biological function of these cells. Similarly, the endothelial HUVEC line dominates the "Regulation of Endothelial Cell Migration" network (contribution = 0.92) these subnetworks represent just a small fraction of those currently identified in the data.

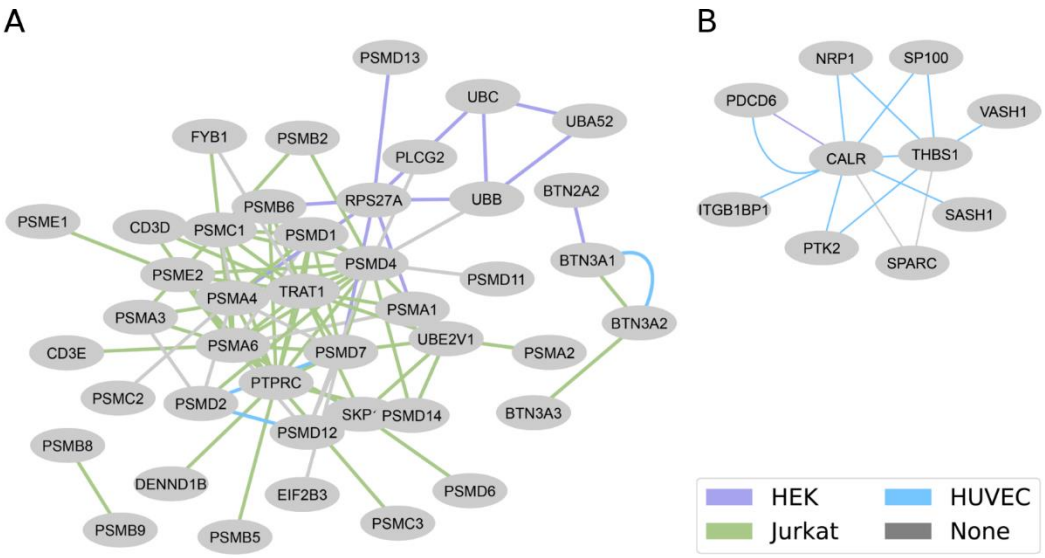


Figure 5-19 Cell Type Specific Contributions to GO Terms Enriched in PROPER
A) PROPER network showing cell type interactions for GO term "T-Cell Receptor Signaling Pathway" (GO:0050852). B) PROPER network showing cell type interactions for GO term "Regulation of Endothelial Cell Migration" (GO:0010594).

All of this data taken together illustrates the power of the PROPER-Seq technique to elucidate cell-specific interactions.

Chapter 5, in part, is currently being prepared for submission for publication of the material. Johnson, Kara; Qi, Zhijie; Wen, Xingzhao; Chen, Chien-ju. The dissertation author was the primary investigator and author of this material.

6 PRIM: High Throughput Identification of Protein RNA Interactions

6.1 Aim

The aim of this technique is to identify interactions between a population of SMART-Display proteins and total RNA.

6.2 Requirements and Control Systems

As in the protein-protein interactions analysis, pairwise interaction data is necessary, as there is not a single 'bait' protein, but potentially thousands. This methodology requires that the barcode of interacting proteins be joined to its partner RNA, and that we select those chimeric fragments for sequencing, as they contain information from both interaction partners. Informatically, both interaction partners need to be identified, and the number and frequency of interactions for each protein enumerated and tested for significance.

6.3 Approach

6.3.1 Design of the Proximity Ligation Method

As mentioned in section 5.3.1, a concern for the efficiency of the system always drives the desire to select the most efficient ligation strategies for the proximity ligation. Here though, one of our interactors is total RNA (also referred to as free RNA to distinguish it from the RNA in the display complex) that is unmodified and therefore has no known sequences, unlike our display nucleic acids. In this case, we do not have the option to convert to DNA and use restriction enzyme ligation, so we directly ligate the total RNA to an interaction linker using a single ended strategy. The efficiency and the specificity of this reaction is increased by using an adenylated linker and an ATP-free ligation reaction.

Because the ligation between the linker and the display nucleic acid can still be performed by sticky end ligation, the interaction liker is designed to be single stranded on one end, and double stranded

with the complementary restriction enzyme site on the other end (for further discussion on the restriction enzyme, see section 5.3.1). As before, the linker contains a biotin for later enrichment of chimeric fragments in the protocol.

The interaction linker provides the same benefits in PRIM as it does in PROPER-Seq. It acts as a tool for discriminating chimeric fragments formed specifically from the desired process, and also acts as a selection marker for the chimeric fragments.

6.3.2 Conversion of Display RNA to DNA

The same strategy is applied here as for PROPER-Seq, and is discussed in section 5.3.2

6.3.3 Interaction Conditions

As for PROPER-Seq, the interaction conditions for this protocol were selected to most closely recapitulate the cellular environment to the greatest degree possible, while also considering ways to reduce background in the experiment. The interaction steps are performed in dilute conditions, to reduce the chance that a protein and an RNA will be in proximity simply due to diffusion. They are also performed at physiological pH and with physiological salt concentration. Finally, interactions are crosslinked to allow for more rigorous washing and the removal of biomolecules that have bound non-specifically in the system. Formaldehyde is used to stabilize the nucleic acid – protein interactions, but also has the ability to stabilize many other macromolecular interactions.

A future goal for the PRIM protocol would be to perform the interaction in a cell-type specific lysate – to best mimic the cellular environment.

6.3.4 Library Preparation

The same strategy is applied here as for PROPER-Seq, and is discussed in section 5.3.4.

6.3.5 Controls

The same strategy is applied here as for PROPER-Seq, and is discussed in section 5.3.5.

6.4 Results and Validation

Preliminary PRIM libraries were generated for HEK cells. One technical replicate was performed alongside the experimental control. All informatics processing was designed in collaboration with and carried out by Zhijie Qi. To see detailed methodology, including the informatics processing and equations, please see Appendix B.

6.4.1 Experimental Features of PRIM Libraries

The HEK experiments were performed with experimental controls. These controls were design such that, when the experiment is successful, the amount of library after the final biotin selection should be less in the controls. Plots of library enrichment during this step confirm enrichment of the positive library relative to the controls, and Bioanalyzer traces illustrate that positive libraries are at higher concentrations than their respective controls, and have a slightly larger average size (Figure 6-1).

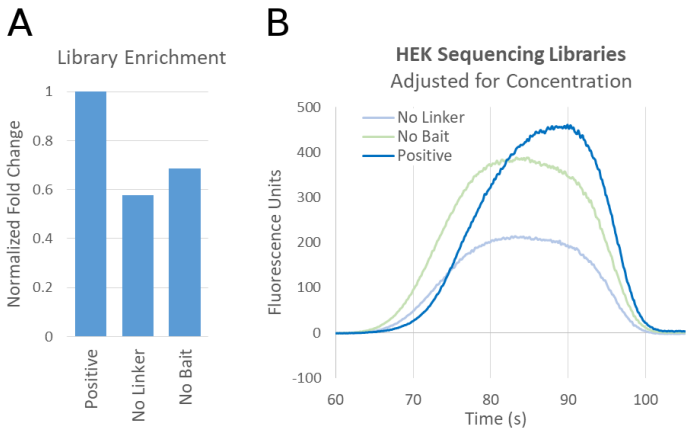


Figure 6-1 Distributions of HEK PRIM Sequencing Libraries
A) Bars show relative enrichment of the positive library over the controls; calculated by dividing the amount of library after biotin selection by the amount present before, and normalizing all those values to the positive library. B and C) The sequencing library distributions as measured by Bioanalyzer.

The resulting sequencing datasets were also checked for their features, including mapping rate and number of chimeric reads identified. The preliminary positive library was sequenced to 305,357,068 reads and 5,932,641 were determined to be chimeric (Table 6.1).

Table 6.1 Read statistics for HEK PRIM Libraries
This table shows the type, total reads, mapped reads, mapping rate, and chimeric reads for each of the HEK PRIM sequenced libraries.

Library ID	Library Type	Total Reads	Mapped Reads	Mapping Rate	Chimeric Reads
HEK1	Positive	305,357,068	158,352,224	51.86%	5,932,641
HEK1_noLinker	No Linker Control	73,158,047	23,675,187	32.36%	869,530
HEK1_noBait	No Bait Control	93,358,221	33,128,523	35.49%	1,408,450

The positive library demonstrated not only a significantly higher mapping rate than the control, but much greater number of chimeric reads. This, along with the specificity of the biotin pull-down and the distribution data suggests that this dataset is a viable candidate for validation and subsequent analysis.

Chapter 6, in part, is currently being prepared for submission for publication of the material. Johnson, Kara; Qi, Zhijie; Wen, Xingzhao; Chen, Chien-ju. The dissertation author was the primary investigator and author of this material.

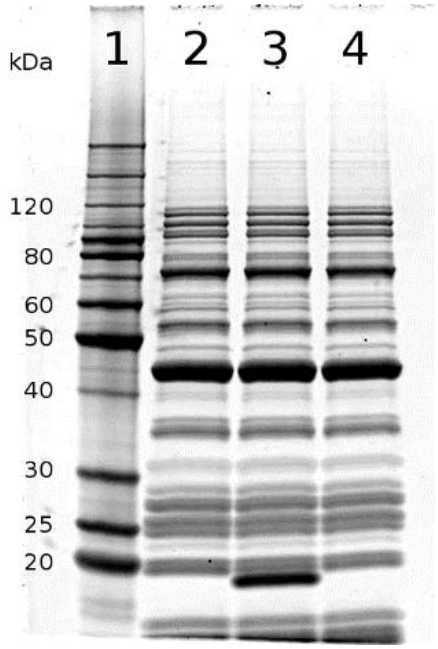
7 Conclusions and Future Work

The techniques presented in this dissertation demonstrate real promise in transforming the manner in which protein interaction networks are explored. Not only have they provided a rapid way to examine a large number of interactions, they allow the process to occur in a cell specific manner. The applications of this technology in human health and biology are numerous. Additionally, SMART-Display can be adapted to a variety of existing protein analysis techniques to increase their throughput or to decrease the time required for the preparation of input materials.

While both SMART-Display and PROPER-Seq have been validated by comparison to literature databases and *in vivo* experiments, datasets resulting from the PRIM technology must still undergo this vigorous qualification. All three technologies demonstrate opportunities to increase yield and decrease non-specific signal to improve the data quality and reduce the overall costs of the process. The proximity ligation and the subsequent enrichment of the generated chimeric fragments remain the two limiting steps in this regard.

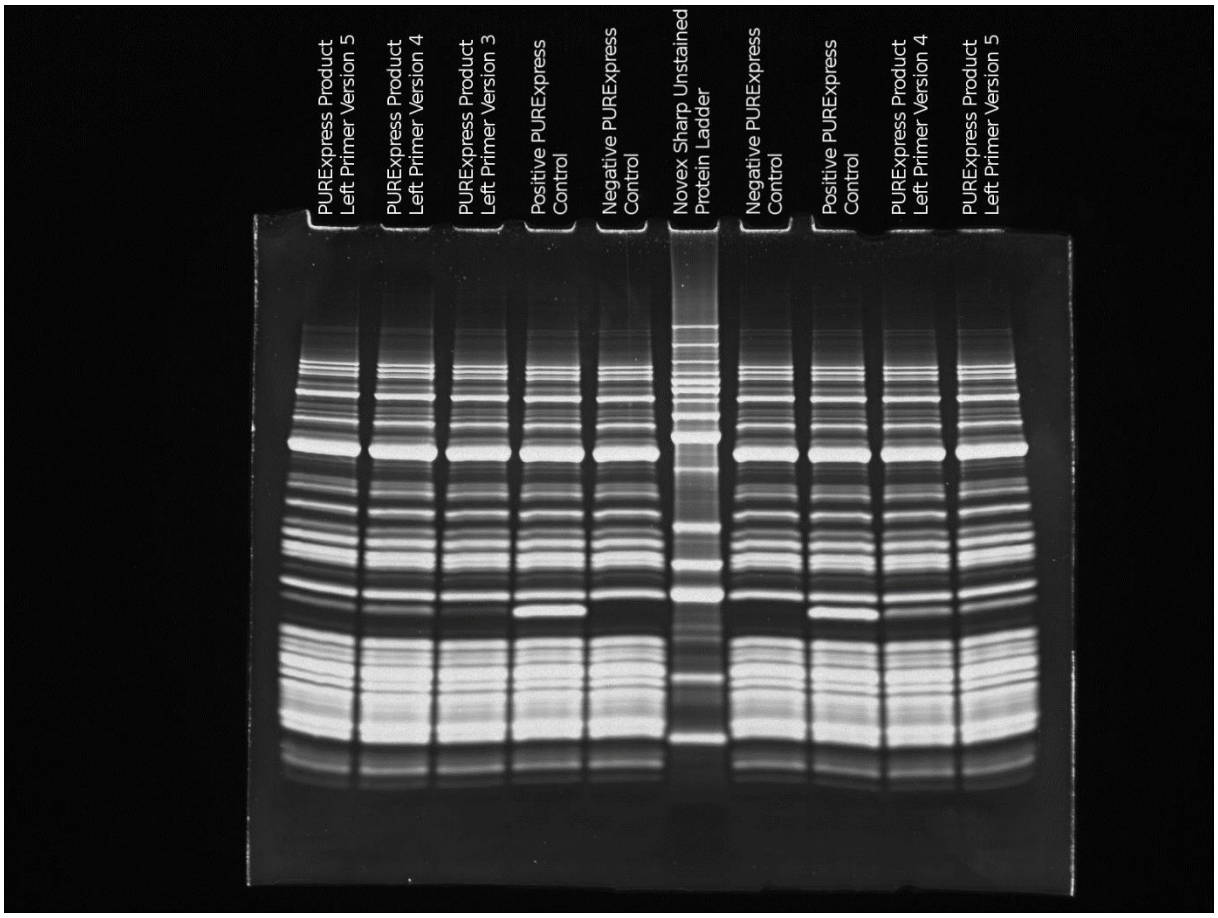
Perhaps the greatest opportunity is to integrate the PRIM and PROPER-Seq data with other high throughput biomolecular interaction techniques. The PRIM and PROPER-Seq networks offer us just a glimpse of the depth and complexity with which we could engage biology by creating a comprehensive, cell-type and state specific network of all biomolecular interactions.

8 Supplementary Materials

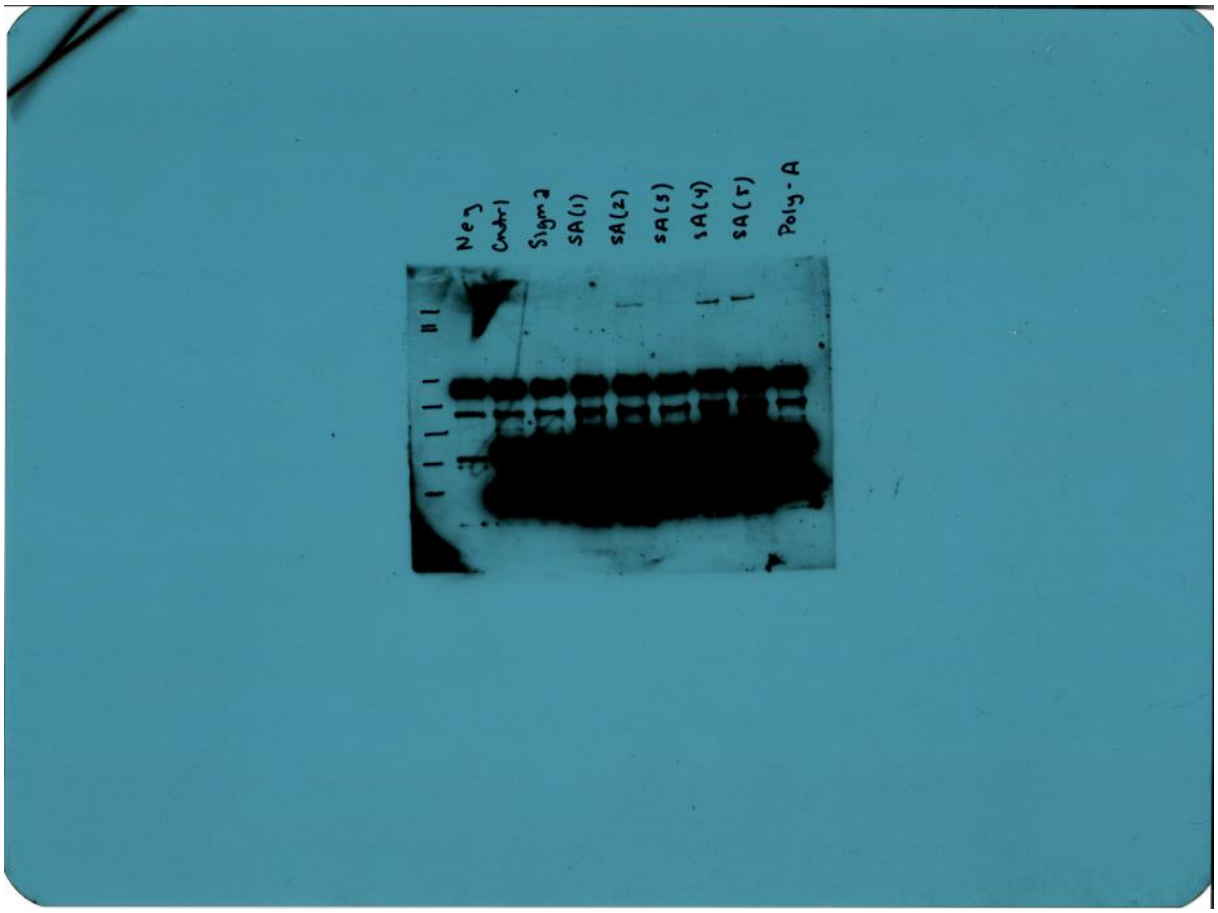


Supplementary Figure 1 PURExpress Products with Universal Plasmid Primers Version One

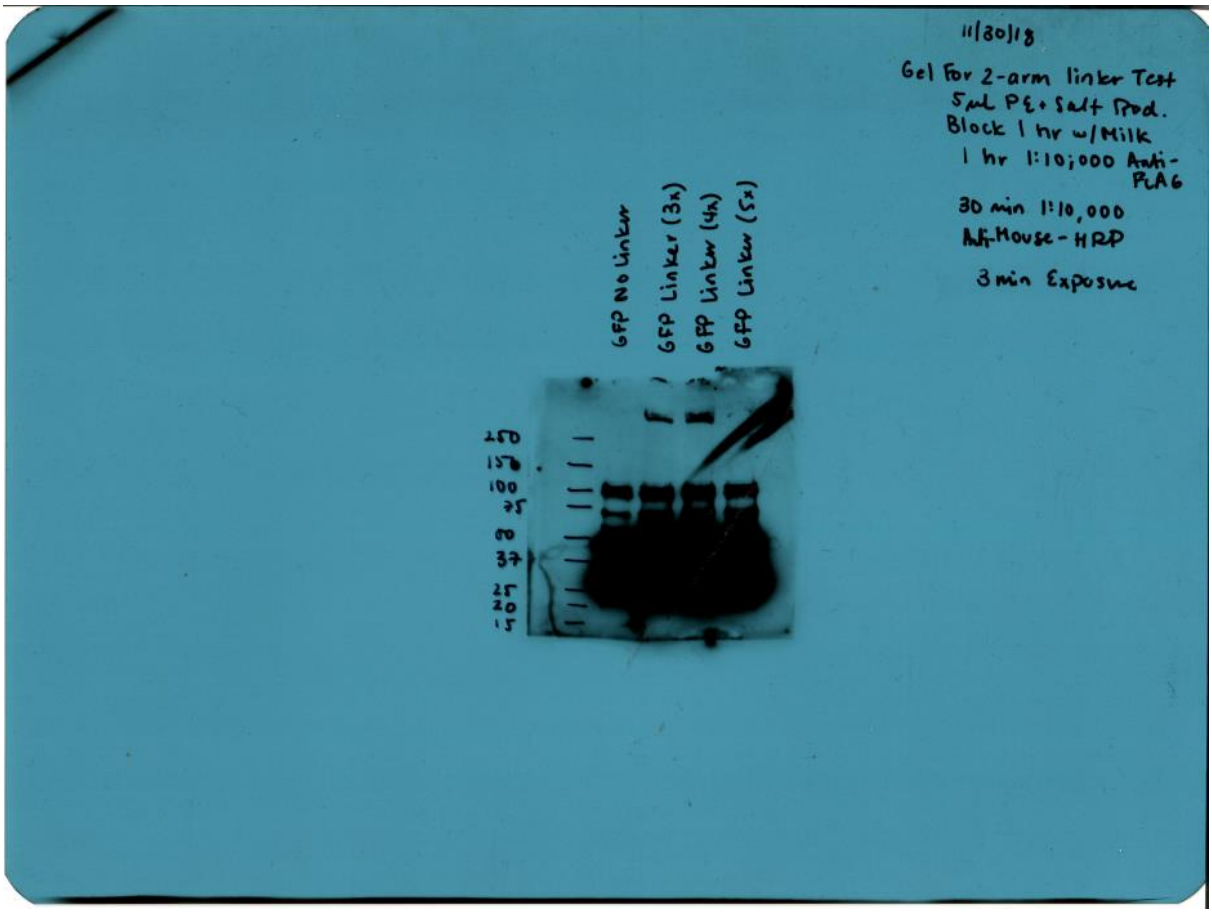
This lanes in this gel are; Lane 1: Benchmark Unstained Protein Ladder, Lane 2: Negative IVT Control (No Template), Lane 3: Positive IVT Control (DHFR Template), Lane 4: MAPK14 IVT Product (MAPK14 template with version 3 primer). The positive control has a strong band around the expected size of 22 kDa; only background bands appear in the sample lane. The success of the positive control and lack of sample translation indicates a poor template or system incompatibility.



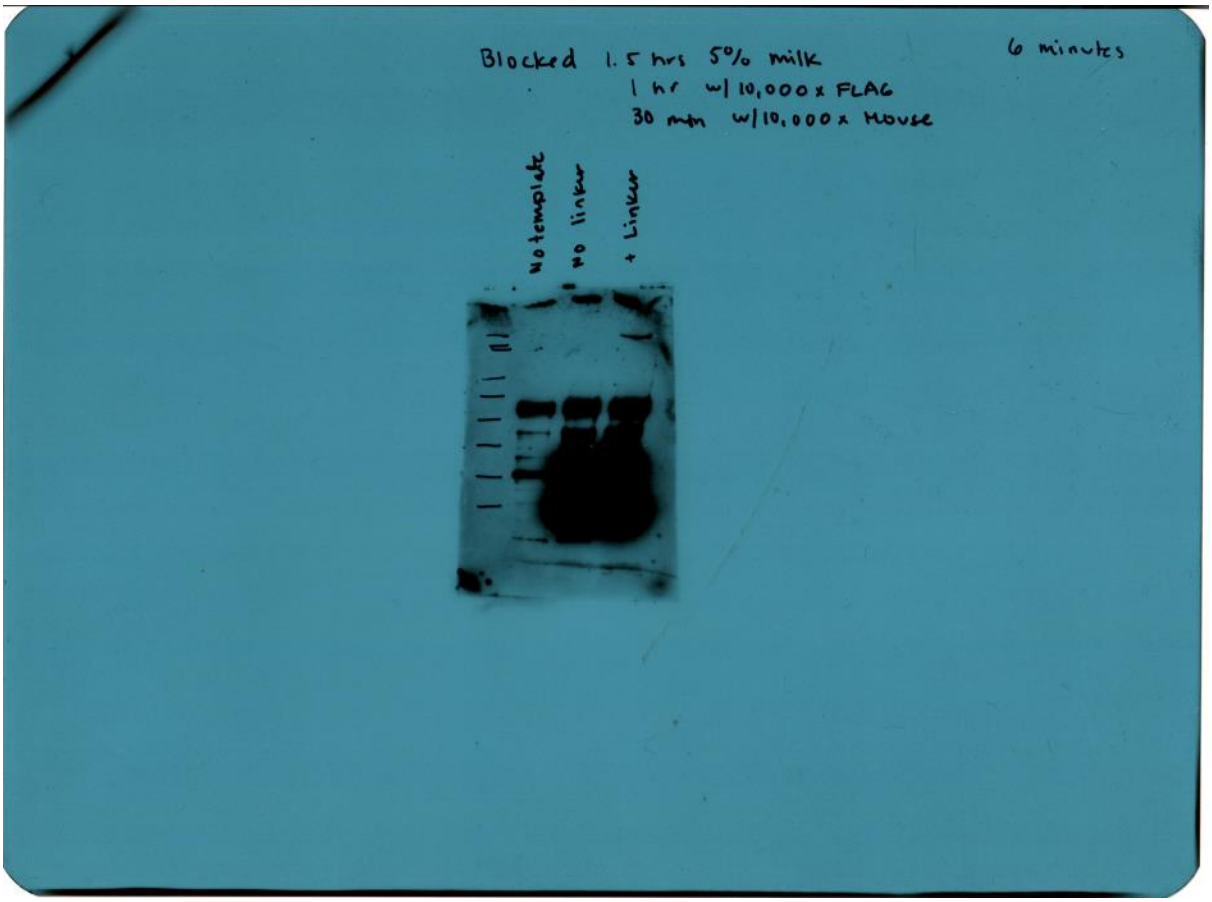
Supplementary Figure 2 Original Image for Figure 4-4 PURExpress IVT Products for Versions of the Universal Plasmid Primer



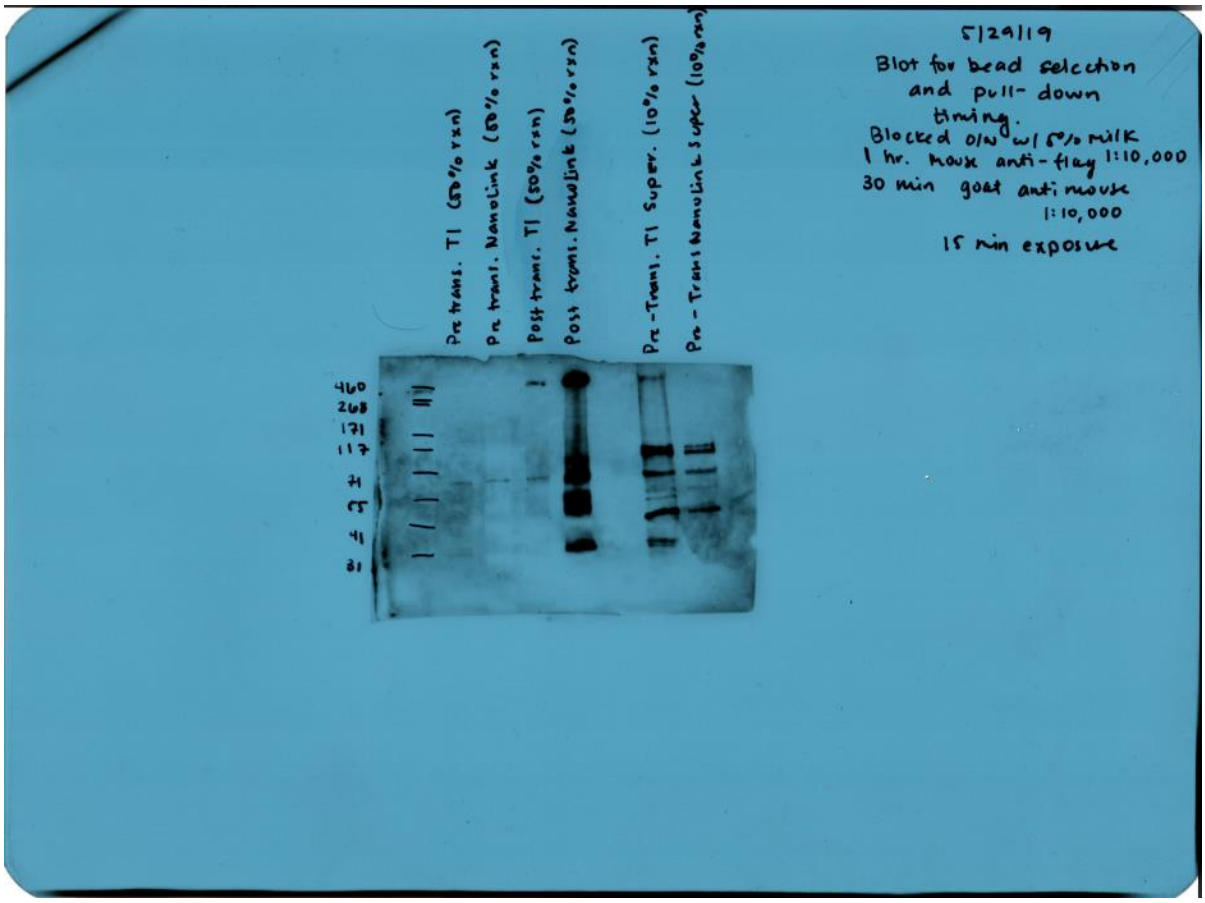
Supplementary Figure 3 Original Image for Figure 4-16 Display Complex Yields for Single-Arm Puromycin Linker Variants



Supplementary Figure 4 Original Image for Figure 4-18 Display Complex Yields for Two-Arm Puromycin Linker Variants



Supplementary Figure 5 Original Image for Figure 4-20 Western Blot for GFP Display ValidationFigure 4-18 Display Complex Yields for Two-Arm Puromycin Linker Variants



Supplementary Figure 6 Original Image for Figure 4-21 Bead Selection and Western Blot for GFP Display ValidationFigure 4-18 Display Complex Yields for Two-Arm Puromycin Linker Variants

APPENDIX

APPENDIX A: PROPER-SEQ METHODS

Experimental Methods

1. 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.

HUVEC and Jurkat cells were cultured in RPMI-1640 Medium (ATCC, 30-2001) supplemented with 10% FBS (Gemini, 100-500), 10 mM HEPES (Sigma-Aldrich, H0887-100ML), and 5,000 U/ml penicillin/streptomycin (GIBCO, 15070063), at 37°C with 5 % CO2.

2. 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, No Longer Available) 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 as assessed with Agilent's Bioanalyzer RNA 6000 Pico Kit.

3. 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 $MgCl_2$ (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.

4. Synthesis of Puromycin Linker

All oligo components of the puromycin linker were reconstituted to 1 mM with 1x PBS pH 7.2 (Thermo Scientific, 20012027). To generate the dI containing puromycin 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 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 a 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 was 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).

5. 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 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).

6. 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 $MgCl_2$ (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.

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).

8. 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 $MgCl_2$, 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.

9. 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 No Bait and No Prey controls were generated by the addition of 5 uLs of Proteinase K (NEB, P8107S) to the appropriate sample. These samples were incubated 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.

10. 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.

11. Interaction Linker Ligation and Release of Prey

Samples with a dI containing puromycin 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 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).

12. Interaction

The sample without dI bases in the puromycin 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 $MgCl_2$, 2mM DTT, 0.2 mM EDTA, 0.1% Tween® 20 (Sigma-Aldrich, P9416-100ML)). The 50 uL of supernatant from the Endonuclease V digestion, containing the prey library, was added the bait samples with the following conventions, Positive Reaction: No Treatment Bait and No Treatment Prey, No Linker Control: No Treatment Bait and No Interaction Linker Prey, and No Bait Control: Proteinase K digested Bait and No Treatment Prey. 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.

13. Crosslinking and Proximity Ligation

Crosslinking was performed at room temperature for 30 minutes with 0.5 mM BS3 (Thermo 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.

14. 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.

Informatic Analysis

All bioinformatic processing was designed in collaboration with and completed by Zhijie Qi.

1. Identification

Linker and adapter sequences are first removed from raw read pairs using Cutadapt. Fastp is then applied to remove low-quality and reads determined to be too short. Read1 and read2 libraries are mapped to transcriptome with BWA separately. The '-a' option is enabled to keep all found alignments using the default threshold of the BWA tool. The mapped read pairs are then deduplicated based on the external coordinates of their primary alignments.

Read pairs whose two ends' primary alignments are mapped to different protein-coding genes are identified and kept. The selected read pairs are then checked to ensure both ends have over 50% of their read bases matched the reference transcriptome, and that the two ends have no shared lesser alignments. The read pairs passing the quality checks above are retained as chimeric read pairs from the library.

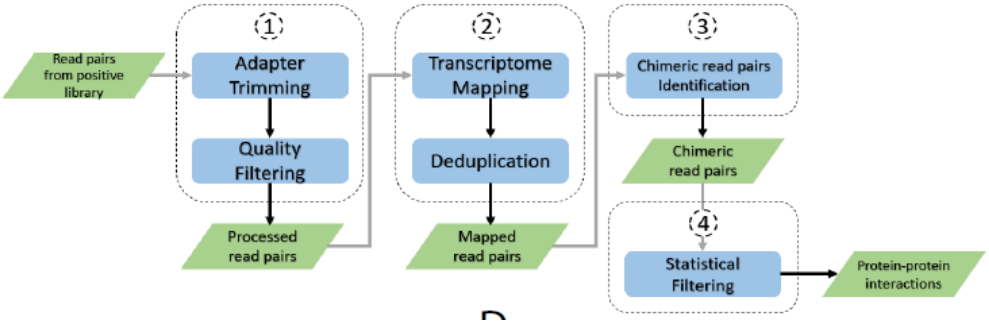


Figure 1 Chi-Suared Table for 'chimericAdj' Protein Interaction Identification
A contingency table showing how chi-square test is applied to a gene pair X-Y for the 'chimericAdj' library.

To determine the 'chimericAdj' protein-protein interaction dataset, for each chimeric gene pair found in the positive library a chi-square test is applied as shown in Figure . Benjamini-Hochberg adjustment is applied to correct all the p-values. Gene pairs with an adjusted p-value less than 0.05 and with an odds ratio larger than 1 are kept. Gene pairs with mapped chimeric read pair count in the positive library larger than 4 times the average number of mapped chimeric read pairs per gene pair in the positive library are kept. The average number of mapped chimeric read pairs per gene pair in the positive library, [X], is computed as

Gene pair X-Y	Number of chimeric read pairs mapped to Gene X	Number of chimeric read pairs NOT mapped to Gene X
Number of chimeric read pairs mapped to Gene Y		
Number of chimeric read pairs NOT mapped to Gene Y		

Total population = Number of chimeric read pairs in the positive library
(Output from process 3 in Panel A)

Figure Chi-Squared Table for 'chimericAdj' Protein Interaction Identification

A contingency table showing how chi-square test is applied to a gene pair X-Y for the 'chimericAdj' library.

Equation 8-1. The kept gene pairs are identified as protein-protein interactions.

Gene pair X-Y	Number of chimeric read pairs mapped to Gene X	Number of chimeric read pairs NOT mapped to Gene X
Number of chimeric read pairs mapped to Gene Y		
Number of chimeric read pairs NOT mapped to Gene Y		

Total population = Number of chimeric read pairs in the positive library (Output from process 3 in Panel A)

Figure Chi-Squared Table for 'chimericAdj' Protein Interaction Identification
A contingency table showing how chi-square test is applied to a gene pair X-Y for the 'chimericAdj' library.

Equation 8-1 Calculation of [X] Statistic
Computation of the average number of mapped chimeric read pairs per gene pair in the positive library, designated as [X].

$$[X] = \frac{\#\ total\ chimeric\ read\ pairs\ in\ the\ positive\ library}{\#\ total\ gene\ pairs\ in\ the\ positive\ library}$$

To determine the 'controlAdj' protein-protein interaction dataset, for each chimeric gene pair found in the positive library a chi-square test is applied as shown in Figure . Benjamini-Hochberg adjustment is applied to correct all the p-values. Gene pairs with an adjusted p-value less than 0.05 and with an odds ratio larger than 1 are kept. Gene pairs with mapped chimeric read pair count in the positive library larger than 4 times the average number of mapped chimeric read pairs per gene pair in the positive library are kept. The kept gene pairs are identified as protein-protein interactions.

Gene pair X-Y	Number of read pairs mapped to X-Y	Number of read pairs NOT mapped to X-Y
Number of chimeric read pairs in the positive library		
Number of chimeric read pairs in the control library		

Total population = Number of mapped read pairs in the positive and the control libraries

Figure 4 Chi-Squared Table for 'ccontrolAdj' Protein Interaction Identification
A contingency table showing how chi-square test is applied to a gene pair X-Y for the 'controlAdj' library.

2. Precision and Sensitivity

Precision and sensitivity were calculated for the different PROPER-Seq datasets using Equation 8-2 and Equation 8-3. Two different reference protein-protein interaction databases were used in these calculations, the Agile Protein Interactomes Data Server (APID) and The Human Reference Protein Interactome Mapping Project (HuRI).

Equation 8-2 Precision Equation
Computation of precision of a query PPI set against a subject PPI set. For example, the precision of HEK set against APID set where HEK is the query set and APID is the subject set.

$$Precision = \frac{\#Overlapped\ PPIs}{\#PPIs\ in\ the\ query\ set} * 100$$

Equation 8-3 Sensitivity Equation
Computation of sensitivity of a query PPI set against a subject PPI set.

$$Sensitivity = \frac{\#Overlapped\ PPIs}{\#PPIs\ in\ the\ subject\ set} * 100$$

PLA Assay

1. 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.

2. Fixation

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).

3. Permeabilization

Cells were washed once with PBS pH 7.2, then permeablized with 200 uLs of 0.1% Triton X-100 (Sigma-Aldrich, T8787-50ML) in PBS for 15 minutes at room temperature with rocking.

4. 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.

5. 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.

Target	Manufacturer	Catalogue Number	Dilution
PARP1	Abcam	Ab227244	1:250
PARP1	Atlas Antibodies	AMAb90959	1:200
Sumo 1	Abcam	Ab32058	1:250
EIF5A2	Atlas Antibodies	HPA029090	1:250
XPO1	Atlas Antibodies	HPA042933	1:500
MATR3	Atlas Antibodies	HPA036565	1:250
IPO5	Santa Cruz Biotechnology	Sc-55527	1:1000
GFP	Thermo Fisher Scientific	A10259	1:250

6. Staining with PLA Probes, Ligation, and Amplification

Slides were wash 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 wash 2x with wash buffer B and 1x with 1:100 wash buffer B.

7. 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).

APPENDIX B: PRIM METHODS

Experimental Methods

1. 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.

2. 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, No Longer Available) 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 as assessed with Agilent's Bioanalyzer RNA 6000 Pico Kit.

3. Synthesis of Interaction Linker

The top and bottom strands of the interaction linker were reconstituted to 200 uM with water. The top strand was adenylated with the 5´ DNA Adenylation Kit (NEB, E2610S) and then purified with Zymo ssRNA/DNA Clean & Concentrator Kit (Zymo Research, D7010).

4. Ligation of Total RNA to Interaction Linker

Purified total RNA was fragmented with the NEBNext® Magnesium RNA Fragmentation Module (NEB, E6150S) for 2 minutes at 94 °C. The fragmented RNA was purified with an RNeasy Mini Column (Qiagen, 74104). The ends of the 200 pmols of RNA were repaired in a 200 uL reaction containing 100 U Quick CIP (NEB, M0525S) and 1x CutSmart buffer. The reaction was incubated at 37 °C for 1 hour then purified with RNeasy Mini Columns.

To ligate the interaction linker to the total RNA, a 200 uL reaction was mixed containing 400 pmols of App-interaction linker, 200 pmol of fragmented total RNA, 4,000 U T4 RNA Ligase 2, truncated KQ (NEB, M0373S), 1x T4 RNA Ligase buffer, and 15% PEG 8000. This reaction was incubated at 16 °C overnight then purified with RNeasy Mini Columns. The bottom strand of the interaction linker was added to the purified sample in a 1:1 molar ratio in 1x 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)]. The solution was heated to 75°C for 5 minutes and cool to 25°C at 0.1°C per second.

5. 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.

6. Synthesis of Puromycin Linker

All oligo components of the puromycin linker were reconstituted to 1 mM with 1x PBS pH 7.2 (Thermo Scientific, 20012027). To generate puromycin linker, 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 a 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 was 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).

7. 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 linker in a 1:1.25 molar ratio in 1x Annealing Buffer, 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.

8. 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.

9. 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).

10. 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.

11. 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 an additional 10 minutes at room temperature.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

APPENDIX C: PROPER-SEQ SAMPLES

All files are located at:

/mnt/extraids/OceanStor-SysCmn-2/qizhijie/ProteinProteinInteraction/PROPERSeq_fastq

Sample ID	PPI-7b_positive	PPI-7b_noLinker	PPI-7b_noBait
Cell Type	HEK	HEK	HEK
Sample Type	Positive	No Linker	No Bait
NxSeq Index	8	9	11
Concentration (ng/uL)	25.2	5.9	5.3
Concentration (nM)	55.08	13.57	13.41
Average Size (bp)	705	670	609
Sequecing Platform	HiSeq 4000	HiSeq 4000	HiSeq 4000
Fastq File	/PPI_7b_positive	/PPI_7b_noLinker	/PPI_7b_noBait
# total reads (1)	343,861,373	69,732,544	87,444,917
# quality filtered input reads (2)	295,124,871	59,266,035	67,419,857
# protein-coding mapped reads (3)	205,881,483	42,197,977	41,828,629
% mapping (3)/(1)*100%	59.87%	60.51%	47.83%
% mapping (3)/(2)*100%	69.76%	71.20%	62.04%
# chimeric reads (4)	12,581,208	2,152,085	1,766,424
% chimeric reads (4)/(1)*100%	3.66%	3.09%	2.02%
% chimeric reads (4)/(2)*100%	4.26%	3.63%	2.62%
% chimeric reads (4)/(3)*100%	6.11%	5.10%	4.22%
# gene pairs	2,622,193	675,280	529,217

PROPER-SEQ SAMPLES, CONTINUED

Sample ID	PPI-8b_positive	PPI-8b_noLinker	PPI-8b_noBait
Cell Type	HEK	HEK	HEK
Sample Type	Positive	No Linker	No Bait
NxSeq Index	6	7	9
Concentration (ng/uL)	16.4	6.04	5.34
Concentration (nM)	36.46	15.21	12.92
Average Size (bp)	693	612	637
Sequecing Platform	HiSeq 4000	HiSeq 4000	HiSeq 4000
Fastq File	/PPI_8b_positive	/PPI_8b_noLinker	/PPI_8b_noBait
# total reads (1)	248,657,713	97,353,678	64,497,521
# quality filtered input reads (2)	230,521,665	87,904,957	59,795,339
# protein-coding mapped reads (3)	173,300,648	64,671,472	46,428,119
% mapping (3)/(1)*100%	69.69%	66.43%	71.98%
% mapping (3)/(2)*100%	75.18%	73.57%	77.65%
# chimeric reads (4)	7,747,982	2,462,181	2,237,573
% chimeric reads (4)/(1)*100%	3.12%	2.53%	3.47%
% chimeric reads (4)/(2)*100%	3.36%	2.80%	3.74%
% chimeric reads (4)/(3)*100%	4.47%	3.81%	4.82%
# gene pairs	2,350,176	937,739	926,592

PROPER-SEQ SAMPLES, CONTINUED

Sample ID	PPI-JKT_positive1	PPI-JKT_positive2
Cell Type	JURKAT	JURKAT
Sample Type	Positive	Positive
NxSeq Index	6	10
Concentration (ng/uL)	11.8	10.7
Concentration (nM)	22.53	20.84
Average Size (bp)	807	791
Sequecing Platform	NovaSeq	NovaSeq
Fastq File	/PPI_JKT_hiseq_positive_1	/PPI_JKT_hiseq_positive_2
# total reads (1)	444,413,111	390,643,931
# quality filtered input reads (2)	336,708,449	299,796,601
# protein-coding mapped reads (3)	262,211,890	236,283,970
% mapping (3)/(1)*100%	59.00%	60.49%
% mapping (3)/(2)*100%	77.88%	78.81%
# chimeric reads (4)	9,988,056	9,385,745
% chimeric reads (4)/(1)*100%	2.25%	2.40%
% chimeric reads (4)/(2)*100%	2.97%	3.13%
% chimeric reads (4)/(3)*100%	3.81%	3.97%
# gene pairs	1,625,773	1,577,795

PROPER-SEQ SAMPLES, CONTINUED

Sample ID	PPI-HUVEC2_positive1	PPI-HUVEC2_positive2
Cell Type	HUVEC	HUVEC
Sample Type	Positive	Positive
NxSeq Index	4	5
Concentration (ng/uL)	7.48	13.1
Concentration (nM)	16	27.17
Average Size (bp)	720	743
Sequecing Platform	NovaSeq	NovaSeq
Fastq File	PPI_huvec2_hiseq_positive_1	PPI_huvec2_hiseq_positive_2
# total reads (1)	359,807,741	483,597,124
# quality filtered input reads (2)	255,700,416	374,288,750
# protein-coding mapped reads (3)	194,690,153	283,434,465
% mapping (3)/(1)*100%	54.11%	58.61%
% mapping (3)/(2)*100%	76.14%	75.73%
# chimeric reads (4)	6,404,274	9,705,398
% chimeric reads (4)/(1)*100%	1.78%	2.01%
% chimeric reads (4)/(2)*100%	2.50%	2.59%
% chimeric reads (4)/(3)*100%	3.29%	3.42%
# gene pairs	894,041	1,309,197

APPENDIX D: PRIM SAMPLES

All files are located at:

/mnt/extraids/OceanStor-SysCmn-2/qizhijie/ RNAProteinInteraction/RPISeq_fastq

Sample ID	PPI-7b_positive	PPI-7b_noLinker	PPI-7b_noBait
Cell Type	HEK	HEK	HEK
Sample Type	Positive	No Linker	No Bait
NxSeq Index	1	2	4
Concentration (ng/uL)	12.5	7.75	11.6
Concentration (nM)	31.57	21.91	32.8
Average Size (bp)	610	545	545
Sequecing Platform	NovaSeq	NovaSeq	NovaSeq
Fastq File	/RPI_4_hiseq_positive	/RPI_4_hiseq _noLinker	/RPI_4_hiseq _noBait
# total reads (1)	409,132,179	104,582,763	146,555,709
# input reads	409,132,179	104,582,763	146,555,709
# quality filtered input reads (1)	305,357,068	73,158,047	93,358,221
# mapped reads (2)	158,352,224	23,675,187	33,128,523
% mapping (2)/(1)*100%	51.86%	32.36%	35.49%
# chimeric reads	16,530,788	2,146,110	3,214,219
# deduped valid chimeric reads (3)	5,932,641	869,530	1408450
% chimeric reads (3)/(1)*100%	1.94%	1.19%	1.51%
# gene pairs	1,553,980	430,569	695,918

WORKS CITED

1. Gu, L., Li, C., Aach, J., Hill, D. E., Vidal, M. & Church, G. M. Multiplex single-molecule interaction profiling of DNA-barcoded proteins. *Nature* **515**, 554–7 (2014).

2. Lewis, J. D., Wan, J., Ford, R., Gong, Y., Fung, P., Nahal, H., Wang, P. W., Desveaux, D. & Guttman, D. S. 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 (2012).

3. Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X., Sahalie, J., Salehi-Ashtiani, K., Hao, T., Cusick, M. E., Hill, D. E., Roth, F. P., Braun, P. & Vidal, M. Leveraging the power of next-generation sequencing to generate interactome datasets. *Nat. Methods* **8**, 478–80 (2011).

4. Darmanis, S., Nong, R. Y., Vänelid, J., Siegbahn, A., Ericsson, O., Fredriksson, S., Bäcklin, C., Gut, M., Heath, S., Gut, I. G., Wallentin, L., Gustafsson, M. G., Kamali-Moghaddam, M. & Landegren, U. ProteinSeq: High-Performance Proteomic Analyses by Proximity Ligation and Next Generation Sequencing. *PLoS One* **6**, e25583 (2011).

5. McGregor, L. M., Jain, T. & Liu, D. R. Identification of ligand-target pairs from combined libraries of small molecules and unpurified protein targets in cell lysates. *J. Am. Chem. Soc.* **136**, 3264–70 (2014).

6. Castello, A., Horos, R., Strein, C., Fischer, B., Eichelbaum, K., Steinmetz, L. M., Krijgsveld, J. & Hentze, M. W. Comprehensive Identification of RNA-Binding Proteins by RNA Interactome Capture. in 131–139 (Humana Press, New York, NY, 2016). doi:10.1007/978-1-4939-3067-8_8.

7. Trendel, J., Schwarzl, T., Horos, R., Prakash, A., Bateman, A., Hentze, M. W. & Krijgsveld, J. The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell* (2018) doi:10.1016/J.CELL.2018.11.004.

8. Nemoto, N., Fukushima, T., Kumachi, S., Suzuki, M., Nishigaki, K. & Kubo, T. A versatile C-terminal specific biotinylation of proteins using both a puromycin-linker and a cell-free translation system for studying high-throughput protein-molecule interactions. *Anal. Chem.* (2014) doi:10.1021/ac501601g.

9. van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J. & Lander, E. S. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* (2010) doi:10.3791/1869.

10. Hammond, M., Nong, R. Y., Ericsson, O., Pardali, K. & Landegren, U. Profiling cellular protein complexes by proximity ligation with dual tag microarray readout. *PLoS One* **7**, e40405 (2012).

11. Nguyen, T. C., Cao, X., Yu, P., Xiao, S., Lu, J., Biase, F. H., Sridhar, B., Huang, N., Zhang, K. & Zhong, S. Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. *Nat. Commun.* **7**, 12023 (2016).

12. Cotten, S. W., Zou, J., Alexander Valencia, C. & Liu, R. Selection of proteins with desired properties from natural proteome libraries using mRNA display. *Nat. Protoc.* **6**, (2011).

13. Devlin, J. J., Panganiban, L. C. & Devlin, P. E. Random peptide libraries: a source of specific protein binding molecules. *Science* **249**, 404–6 (1990).

14. Hanes, J. & Plückthun, A. In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 4937–42 (1997).

15. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci.* **94**, 12297–12302 (1997).

16. Kurz, M., Gu, K., Al-Gawari, A. & Lohse, P. A. cDNA - protein fusions: covalent protein - gene conjugates for the in vitro selection of peptides and proteins. *Chembiochem* **2**, 666–72 (2001).

17. Odegrip, R., Coomber, D., Eldridge, B., Hederer, R., Kuhlman, P. A., Ullman, C., FitzGerald, K. & McGregor, D. CIS display: In vitro selection of peptides from libraries of protein-DNA complexes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2806–10 (2004).

18. Reiersen, H., Løbersli, I., Løset, G. Å., Hvattum, E., Simonsen, B., Stacy, J. E., McGregor, D., Fitzgerald, K., Welschof, M., Brekke, O. H. & Marvik, O. J. Covalent antibody display--an in vitro antibody-DNA library selection system. *Nucleic Acids Res.* **33**, e10–e10 (2005).

19. Kaltenbach, M. & Hollfelder, F. SNAP display: in vitro protein evolution in microdroplets. *Methods Mol. Biol.* **805**, 101–11 (2012).

20. Fields, S. & Song, O. A novel genetic system to detect protein protein interactions. *Lett. to Nat.* (1989).

21. Walhout, a J. & Vidal, M. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24**, 297–306 (2001).

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

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

24. Vermeulen, M., Hubner, N. C. & Mann, M. High confidence determination of specific protein–protein interactions using quantitative mass spectrometry. *Curr. Opin. Biotechnol.* **19**, 331–337 (2008).

25. Roux, K. J., Kim, D. I. & Burke, B. BioID: A Screen for Protein-Protein Interactions. in *Current Protocols in Protein Science* vol. 74 19.23.1-19.23.14 (John Wiley & Sons, Inc., 2013).

26. Larman, H. B., Zhao, Z., Laserson, U., Li, M. Z., Ciccia, A., Gakidis, M. A. M., Church, G. M., Kesari, S., Leproust, E. M., Solimini, N. L. & Elledge, S. J. Autoantigen discovery with a synthetic human peptidome. *Nat. Biotechnol.* **29**, 535–41 (2011).

27. Sidhu, S. S., Fairbrother, W. J. & Deshayes, K. Exploring protein-protein interactions with phage display. *Chembiochem* **4**, 14–25 (2003).

28. Zhu, J., Larman, H. B., Gao, G., Somwar, R., Zhang, Z., Laserson, U., Ciccia, A., Pavlova, N., Church, G., Zhang, W., Kesari, S. & Elledge, S. J. Protein interaction discovery using parallel analysis of translated ORFs (PLATO). *Nat. Biotechnol.* **31**, 331–4 (2013).

29. Fujimori, S., Hirai, N., Ohashi, H., Masuoka, K., Nishikimi, A., Fukui, Y., Washio, T., Oshikubo, T., Yamashita, T. & Miyamoto-Sato, E. Next-generation sequencing coupled with a cell-free display technology for high-throughput production of reliable interactome data. *Sci. Rep.* **2**, 691 (2012).

30. Hammond, M., Nong, R. Y., Ericsson, O., Pardali, K. & Landegren, U. Profiling cellular protein complexes by proximity ligation with dual tag microarray readout. *PLoS One* **7**, e40405 (2012).

31. Kukar, T., Eckenrode, S., Gu, Y., Lian, W., Megginson, M., She, J.-X. & Wu, D. Protein microarrays to detect protein-protein interactions using red and green fluorescent proteins. *Anal. Biochem.* **306**, 50–4 (2002).

32. Stork, C. & Zheng, S. Genome-Wide Profiling of RNA–Protein Interactions Using CLIP-Seq. in *Methods in molecular biology (Clifton, N.J.)* vol. 1421 137–151 (2016).

33. Gilbert, C., Svejstrup, J. Q., Gilbert, C. & Svejstrup, J. Q. RNA Immunoprecipitation for Determining RNA-Protein Associations In Vivo. in *Current Protocols in Molecular Biology* 27.4.1-27.4.11 (John Wiley & Sons, Inc., 2006). doi:10.1002/0471142727.mb2704s75.

34. Transcriptome-wide analysis of protein–RNA interactions using high-throughput sequencing. *Semin. Cell Dev. Biol.* **23**, 206–212 (2012).

35. McHugh, C. A. & Guttman, M. RAP-MS: A Method to Identify Proteins that Interact Directly with a Specific RNA Molecule in Cells. in 473–488 (Humana Press, New York, NY, 2018). doi:10.1007/978-1-4939-7213-5_31.

36. Zeng, F., Peritz, T., Kannanayakal, T. J., Kilk, K., Eiríksdóttir, E., Langel, U. & Eberwine, J. A protocol for PAIR: PNA-assisted identification of RNA binding proteins in living cells. *Nat. Protoc.* **1**, 920–927 (2006).

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

38. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. *Mol. Cell* **44**, 667–678 (2011).

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

40. McMahon, A. C., Rahman, R., Jin, H., Shen, J. L., Fieldsend, A., Luo, W. & Rosbash, M. TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell* **165**, 742–753 (2016).

41. Ramanathan, M., Majzoub, K., Rao, D. S., Neela, P. H., Zarnegar, B. J., Mondal, S., Roth, J. G., Gai, H., Kovalski, J. R., Siprashvili, Z., Palmer, T. D., Carette, J. E. & Khavari, P. A. RNA–protein interaction detection in living cells. *Nat. Methods* **15**, 207–212 (2018).

42. Kretz, M., Siprashvili, Z., Chu, C., Webster, D. E., Zehnder, A., Qu, K., Lee, C. S., Flockhart, R. J., Groff, A. F., Chow, J., Johnston, D., Kim, G. E., Spitale, R. C., Flynn, R. A., Zheng, G. X. Y., Aiyer, S., Raj, A., Rinn, J. L., Chang, H. Y. & Khavari, P. A. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**, 231–235 (2012).

43. Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S. & Sandberg, R. Full-

length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–81 (2014).

44. The Basics: In Vitro Translation. *Thermo Fisher Scientific* https://www.thermofisher.com/us/en/home/references/ambion-tech-support/large-scale-transcription/general-articles/the-basics-in-vitro-translation.html.

45. Ueno, S., Kimura, S., Ichiki, T. & Nemoto, N. Improvement of a puromycin-linker to extend the selection target varieties in cDNA display method. *J. Biotechnol.* **162**, 299–302 (2012).

46. Mochizuki, Y., Biyani, M., Tsuji-Ueno, S., Suzuki, M., Nishigaki, K., Husimi, Y. & Nemoto, N. One-pot preparation of mRNA/cDNA display by a novel and versatile puromycin-linker DNA. *ACS Comb. Sci.* **13**, 478–85 (2011).

47. Prieto, C., De, J. & Rivas, L. APID: Agile Protein Interaction DataAnalyzer. doi:10.1093/nar/gkl128.

48. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Cote, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., Knapp, J. J., Kovács, I. A., Lemmens, I., Mee, M. W., Mellor, J. C., Pollis, C., Pons, C., Richardson, A. D., Schlabach, S., Teeking, B., Yadav, A., Babor, M., Balcha, D., Basha, O., Bowman-Colin, C., Chin, S.-F., Choi, S. G., Colabella, C., Coppin, G., D'Amata, C., Ridder, D. De, Rouck, S. De, Duran-Frigola, M., Ennajdaoui, H., Goebels, F., Goehring, L., Gopal, A., Haddad, G., Hatchi, E., Helmy, M., Jacob, Y., Kassa, Y., Landini, S., Li, R., Lieshout, N. van, MacWilliams, A., Markey, D., Paulson, J. N., Rangarajan, S., Rasla, J., Rayhan, A., Rolland, T., San-Miguel, A., Shen, Y., Sheykhkarimli, D., Sheynkman, G. M., Simonovsky, E., Taşan, M., Tejeda, A., Twizere, J.-C., Wang, Y., Weatheritt, R. J., Weile, J., Xia, Y., Yang, X., Yeger-Lotem, E., Zhong, Q., Aloy, P., Bader, G. D., Rivas, J. D. Las, Gaudet, S., Hao, T., Rak, J., Tavernier, J., Tropepe, V., Hill, D. E., Vidal, M., Roth, F. P. & Calderwood, M. A. A reference map of the human protein interactome. *bioRxiv* 605451 (2019) doi:10.1101/605451.

49. Söderberg, O., Leuchowius, K. J., Gullberg, M., Jarvius, M., Weibrecht, I., Larsson, L. G. & Landegren, U. Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. *Methods* **45**, 227–232 (2008).

50. Sundberg, E., Fasth, A. E. R., Palmblad, K., Harris, H. E. & Andersson, U. High mobility group box chromosomal protein 1 acts as a proliferation signal for activated T lymphocytes. *Immunobiology* **214**, 303–309 (2009).

51. Kirchgessner, H., Dietrich, J., Scherer, J., Isomäki, P., Korinek, V., Hilgert, I., Bruyns, E., Leo, A., Cope, A. P. & Schraven, B. The transmembrane adaptor protein TRIM regulates T cell receptor (TCR) expression and TCR-mediated signaling via an association with the TCR δ chain.

J. Exp. Med. **193**, 1269–1283 (2001).

52. Delehedde, M., Devenyns, L., Maurage, C.-A. & Vivès, R. R. Endocan in Cancers: A Lesson from a Circulating Dermatan Sulfate Proteoglycan. *Int. J. Cell Biol.* **2013**, 11 (2013).

53. Béchard, D., Scherpereel, A., Hammad, H., Gentina, T., Tsicopoulos, A., Aumercier, M., Pestel, J., Dessaint, J.-P., Tonnel, A.-B. & Lassalle, P. Human Endothelial-Cell Specific Molecule-1 Binds Directly to the Integrin CD11a/CD18 (LFA-1) and Blocks Binding to Intercellular Adhesion Molecule-1. *J. Immunol.* **167**, 3099–3106 (2001).

54. Geberhiwot, T., Assefa, D., Kortesmaa, J., Ingerpuu, S., Pedraza, C., Wondimu, Z., Charo, J., Kiessling, R., Virtanen, I., Tryggvason, K. & Patarroyo, M. Laminin-8 (alpha4beta1gamma1) is synthesized by lymphoid cells, promotes lymphocyte migration and costimulates T cell proliferation. *J. Cell Sci.* **114**, 423–33 (2001).

55. Newman, P. J. & Newman, D. K. Signal Transduction Pathways Mediated by PECAM-1 New Roles for an Old Molecule in Platelet and Vascular Cell Biology. (2003) doi:10.1161/01.ATV.0000071347.69358.D9.

56. Brown, S., Heinisch, I., Ross, E., Shaw, K., Buckley, C. O. & Savill, J. Apoptosis disables CD31-mediated cell detachment from phagocytes promoting binding and engulfment. *Nature* **418**, 200–203 (2002).

57. Giri, R., Selvaraj, S., Miller, C. A., Hofman, F., Yan, S. D., Stern, D., Zlokovic, B. V. & Kalra, V. K. Effect of endothelial cell polarity on β-amyloid-induced migration of monocytes across normal and AD endothelium. *Am. J. Physiol. Physiol.* **283**, C895–C904 (2002).