Published in final edited form as:

Methods. 2017 April 15; 118-119: 171–181. doi:10.1016/j.ymeth.2016.10.001.

Integrated analysis of RNA-binding protein complexes using *in vitro* selection and high-throughput sequencing and sequence specificity landscapes (SEQRS)

Tzu-Fang Lou¹, Chase A Weidmann², Jordan Killingsworth², Traci M Tanaka Hall⁴, Aaron C Goldstrohm^{2,3}, and Zachary T Campbell^{*,1}

¹ Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080

² Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109

³ Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN 55455

⁴ Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709

Abstract

RNA-binding proteins (RBPs) collaborate to control virtually every aspect of RNA function. Tremendous progress has been made in the area of global assessment of RBP specificity using next-generation sequencing approaches both *in vivo* and *in vitro*. Understanding how protein-protein interactions enable precise combinatorial regulation of RNA remains a significant problem. Addressing this challenge requires tools that can quantitatively determine the specificities of both individual proteins and multimeric complexes in an unbiased and comprehensive way. One approach utilizes *in vitro* selection, high-throughput sequencing, and sequence-specificity landscapes (SEQRS). We outline a SEQRS experiment focused on obtaining the specificity of a multi-protein complex between *Drosophila* RBPs Pumilio (Pum) and Nanos (Nos). We discuss the necessary controls in this type of experiment and examine how the resulting data can be complemented with structural and cell-based reporter assays. Additionally, SEQRS data can be integrated with functional genomics data to uncover biological function. Finally, we propose extensions of the technique that will enhance our understanding of multi-protein regulatory complexes assembled onto RNA.

1. Introduction

Hundreds of RBPs collaborate to control every step in the life of an RNA, from transcription to processing, export, localization, translation, and ultimately decay (1). The binding sites for these proteins are composed of specific sequences and structured motifs (2, 3). The RBPs

*Corresponding author - zachary.campbell@utdallas.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

that mediate RNA function seldom act in isolation. For example, they can recruit catalytic partners to the RNA and interact with other RNA-binding proteins (3-5). The combinatorial nature of RNA control poses a major problem to reductionist approaches given the inherent complexity of quantitative analysis of interactions between multiple RBPs and RNA. The increasingly common use of functional genomics methods, including massively parallel sequencing, provide a powerful means to understand how protein complexes containing multiple RBPs assemble onto RNA (6).

A variety of methods have been developed to identify the *in vivo* target RNAs of a given RBP including microarray or high-throughput sequencing of RNA isolated by RBP immunoprecipitation (RIP-Chip, RIP-Seq, RIPit-Seq), UV crosslinking and immunoprecipitation (HiTS-CLIP), photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP), or individual-nucleotide resolution crosslinking and immunoprecipitation (iCLIP) (7-10). Each of these approaches involves purification of RNA-protein complexes from cell lysates and determination of the identity of bound RNAs. The use of UV irradiation in HiTS-CLIP, PAR-CLIP, and iCLIP ensures that RNAs in close contact to RBPs before cell lysis are specifically cross-linked. The resolution of binding site determination relative to array-based analysis is enhanced by incorporation of the mutational biases that are found at sites of crosslinking. These methods can capture a tremendous amount of information about bound RNAs and the sequences recognized, but lack the ability to unambiguously differentiate a transient occupancy due to sampling versus a productive regulatory encounter. These techniques are also limited in their ability to discriminate the binding sites and specificities of individual factors that cooperatively recognize target RNAs. *In vitro* techniques that examine the specificity of individual proteins and protein complexes can identify factors that contribute to binding in cells.

Traditionally, RNA-protein interactions were probed using candidate tests such as *in vitro* RNA-binding assays or yeast 3-hybrid analyses (6, 11, 12). While these approaches are powerful in their quantitative precision, they are limited by the depth of sequence space that can be accessed. A number of new methods enable parallel measurements of specificity for many RNAs in a single experiment, as we recently compared and contrasted (6). This information can be tremendously useful for the systematic analysis of RBP specificity. Unbiased RNA selection techniques, originating with systematic evolution of ligands by exponential enrichment (SELEX), allow the isolation of sequences preferentially bound by an RBP or complex (13, 14). High-throughput sequencing has transformed the biochemical utility of SELEX (15, 16). While different methods may be used to analyze the consequences of protein-protein interactions on specificity, we focus here on the use of SEQRS, which is the sole method that has been successfully applied to the problem of combinatorial regulation of RNA (17-21).

As an example for how to design, perform, and analyze data from a SEQRS experiment, we examine PUF (named for *Drosophila* Pum and *fem-3* binding) proteins and their protein partners. PUFs are a conserved family of eukaryotic RBPs that play essential roles in development (22, 23). Structurally, PUFs comprise helical repeat units that assemble into a crescent (19, 24-29). Recognition of substrates by PUFs is highly modular; each repeat may specify the identity of a single RNA base. The SEQRS method has been used to characterize

the evolutionary divergence of the members of this family, detecting a range of binding element lengths ranging from 8-12 bases (17, 19, 30). *In vivo*, PUFs appear to bind hundreds of mRNA targets encompassing 7-16% of the transcriptome and reduce the stability of these mRNAs (31-36).

PUFs require collaboration with a variety of protein partners to exert regulatory effects. For instance, *Drosophila* Pum collaborates with Nanos (Nos), a tandem CCHC Zinc finger (ZF) RBP, to repress maternal *hunchback* (*hb*) mRNA translation (37). Repression of *hb* is required for normal body patterning during embryogenesis (37, 38). Regulation of *hb* mRNA by Pum and Nos is dependent on two binding sites in the 3' untranslated region (UTR) termed Nanos Response Elements (NREs) (23, 39). Each NRE contains a binding site for Pum with the RNA consensus sequence, 5'-UGUAHAUA (where H is A, U or C), the Pumilio Response Element (PRE). Additional nucleotides in the NRE, located 5' of the PRE, are functionally important for *hb* regulation and were proposed to be recognized by Nos (24, 38). By itself, Nos lacked apparent RNA sequence specificity (40).

Here, we describe how SEQRS can be used to analyze the specificity of a multi-protein complex, and we present the Nos-Pum interaction as an example (18). The goal of these experiments was to determine whether addition of Nos resulted in alternative target site selection by the Nos-Pum protein complex as compared to Pum alone. To use SEQRS to analyze the specificity of two physically interacting RNA-binding proteins, we needed to determine the specificity of each individual protein as well as for the complex. We describe the process and subsequent data analysis. Although the approach we describe here is applied to a specific complex, the general methodology is broadly applicable to other systems involving multiple RBPs that interact with RNA.

2. The method

2.1 Overview

Here we outline the steps required to determine the binding preference of a given protein or protein complex for all possible sequences of a given length in a single SEQRS experiment (Fig. 1). To do so, we describe key aspects of experimental design, the selection process, high-throughput sequencing, data analysis, validation, and integration with functional genomics data.

2.2 Experimental design

2.2.1 Considerations for protein production—These are the key criteria of RBP quality for a successful experiment:

- Greater than 500 µg of purified RBP
- Purified to homogeneity
- Free of contaminating ribonucleases (RNases)
- Free of contaminating nucleic acids
- Functional RBP verified by binding assay, if feasible

SEQRS experiments are conducted with purified components and depend upon the availability of soluble, functional, selectable RBPs. As a result, the tractability of obtaining substantial amounts (ca. 500 µg) of purified samples compatible with the selection process must be considered prior to undertaking an experiment. Some important factors to consider are choice and optimization of protein expression system, affinity tags for protein selection, and stability and activity of purified proteins in SEQRS experimental conditions. Additionally, it is imperative that the purified protein be free of contaminating RNase activity.

The choice of protein expression system may influence the protein's utility for SEQRS. Production of eukaryotic proteins in bacterial expression systems is by far the most common approach. That said, eukaryotic expression systems, such as insect virus (i.e. baculovirus) or stable cells lines, may be advantageous. We have found that yields of functional full-length proteins are often better with these systems than with bacterial expression, especially with proteins that are large and/or difficult to express in bacteria. Additionally, eukaryotic expression systems are more likely to produce proteins that are properly post-translationally modified, which may affect their stability and specificity. The major disadvantage of eukaryotic expression is the cost associated with large-scale expression. As a result, we have often used bacterial expression systems, as they are relatively inexpensive and higher throughput.

A wide variety of vectors are available for protein expression in bacteria with different solubility or affinity tags to optimize expression levels and purification. Many of the RBPs we study are relatively unstable when expressed in bacteria without the use of soluble carrier domains, and the fusion partners provide a means to select the RBP and its bound RNAs (see below). We have successfully used glutathione S-transferase (GST) and maltose binding protein (MBP)-fusions. In the example of Pum and Nos described below, we utilized N-terminal HaloTag fusion to purify and immobilize these RBPs to the resin. The covalent attachment of the HaloTag fusion to the Magne HaloTag resin (Promega) is advantageous for selection, as the immobilized protein cannot diffuse throughout the process. A consideration with the use of fusion tags is the potential addition of artifactual sequence specific preferences. These may be detected through inclusion of tag alone controls subjected to SEQRS analysis. The binding motifs obtained from SEQRS for tagged *Drosophila* PUM, Mpt5p, Puf4p, PUM2, and FBF-2 are consistent with independent measurements in cell based assays suggestive of minor contributions originating from the use of fusion tags (17-19).

Standard techniques that optimize heterologous protein expression in bacteria are beneficial for RBP expression. These include adjusting the temperature and time of protein induction, titrating the concentration of inducer molecule (e.g. isopropylβ-D-1-thiogalactopyranoside), use of auto-induction systems, testing alternative growth media (i.e. standard LB broth vs a nutritionally-rich medium such as 2xYT or terrific broth), and optimizing the optical density at which protein expression is induced.

Many RBPs of interest are from organisms that have different codon usage patterns from bacteria, and an important consideration in expressing these proteins in bacteria is codon

optimization. This can be mitigated by two different strategies. First, we use bacterial strains for protein expression that have supplemental genes encoding rare codons (e.g. BL21-Codon plus (Agilent)). Second, for shorter targets, we construct synthetic codon-optimized open reading frames using overlapping gene fragments. The basic strategy, termed Gibson cloning, consists of fusing together overlapping segments of DNA by PCR and introduction of the resulting product into an expression vector using enzymatic recombination (41). Combining both approaches has worked well for many of the constructs we have attempted. Notably, this includes proteins with long stretches of basic amino acids encoded by rare codons that are notoriously challenging to express.

As mentioned above, an important consideration is the choice of affinity tags used for RBP selection in an experiment. In the SEQRS experiments, affinity tags that are compatible with magnetic immobilization following purification are required. Commercial magnetic resins are available to select proteins with GST, MBP, HaloTag, and chitin-binding domain fusions; His, HA, myc, and FLAG tags; or via a specific antibody. Magnetic resins are advantageous for technical and practical reasons. First, in the selection process, a major objective is to separate the RNA-protein (RNP) complex from unbound RNA. The use of magnetic resin allows the RNPs to be trapped on the side of the tube, and liquids that remain in the bottom of the tube are easily removed (as opposed to centrifugation where complete supernatant removal is more difficult). Second, the use of magnetized 96-well plates greatly increases the throughput of the experiments. Third, pipetting errors that cause the resin to be aspirated are easily detected with dark black magnetic beads in an otherwise transparent solution.

If multiple proteins will be analyzed in a single experiment, it is necessary to design the protein fusions to differentially tag each protein of interest. Alternatively, a single tagged protein may be used, but in this case, identical tags on protein partners must be removed. This can usually be accomplished through cleavage by highly sequence-specific proteases (e.g. tobacco etch virus protease - TEV or Sumo protease - ULP). Specific protease cleavage sites can be introduced between affinity tags and proteins of interest, if they are not already present in the expression vector. In the example of Nos and Pum, each protein was purified as a HaloTag fusion. To analyze their cooperative RNA-binding specificity, Nos was cleaved from the resin using TEV protease and then added at equimolar amount to samples containing resin-immobilized HaloTag-Pum. Released affinity tags and uncleaved fusion proteins must be removed after digestion. This may be accomplished using another purification step, such as size exclusion or ion exchange chromatography, or by performing another round of affinity purification. In our work, we addressed this issue by cleaving the purified fusion protein (for instance, HaloTag-Nos fusion) while it was immobilized to the magnetic resin. The cleaved protein was recovered in the liquid phase while the uncleaved fusion or HaloTag alone remained linked to the magnetic resin.

Although N-terminal tags are helpful for protein expression and solubility, one substantial limitation can be the production and co-purification of C-terminally truncated polypeptides, which may arise due to proteolysis or premature translation termination. To mitigate the former, we always use protease-deficient bacterial strains for protein production (e.g. BL21 derivatives or KRX [Promega]) and include protease inhibitor cocktail during purification steps. If truncated polypeptides are present, they must be removed. One strategy is to

engineer the protein expression vectors to encode a C-terminal poly-histidine tag. With this method, the N-terminal tag can be used to purify the protein and then the C-terminal tag can be used to immobilize them, thereby preferentially retaining only full-length proteins.

Although it is desirable to work with highly purified protein samples, this is not always achievable in practice and instead appropriate controls should be designed. When the protein preparation contains multiple bands on an SDS-PAGE gel, contaminating proteins should be identified. Some truncation products can be verified by western blot, probing with an anti-fusion protein antibody, or by mass spectrometric analysis. A crucial concern is to eliminate contaminating RNase activity from the purified protein. Bacteria contain abundant RNases that must be removed during the purification conditions. Moreover, RNases can be unintentionally introduced from environmental contamination of reagents, equipment, and consumables. We strive to maintain an RNase-free environment using pure reagents, ultrapure water, and sterile RNase-free plastics and equipment as well as wearing gloves during all experiments. We subsequently test purified proteins for RNase activity by incubating them with fluorescently-labelled RNA and analyzing the products by gel electrophoresis.

Inclusion of appropriate controls is crucial for interpretation of SEQRS data. To assess background, a negative control selection can be performed with resin alone or resin with bound affinity tag. Ideally, we prefer negative control SEQRS analysis using an RNA-binding defective mutant of the RBP of interest. In the case of Pum, we performed negative control selection using a resin-bound HaloTag fusion of a mutant form of Pum (mutR7) wherein the RNA recognition amino acid residues in the seventh helical repeat were changed to alanine (S1342A N1343A E1346A), thereby preventing RNA binding *in vitro* and loss of regulation *in vivo* (42, 43). This control is extremely informative in understanding and assigning significance to binding elements that are weakly enriched.

For SEQRS to be successful, the proteins of interest must be soluble and functional under selection conditions. Thus, we test purified RBPs using electrophoretic mobility shift assays (EMSAs). We also use the EMSA to optimize the binding conditions that are then used in the SEQRS binding and selection, with minor exceptions. These conditions vary for different types of proteins, and these variations are essential for experimental success. Parameters, such as pH and the concentration of mono and divalent cations, should be optimized before proceeding with SEQRS. For instance, proteins with zinc fingers may require zinc and reducing agent supplementation and the absence of metal chelators during protein production and/or later experimental steps. Special attention should be paid to chemical compatibility of reagents used during purification and the requirements of the RBP and affinity tag. For instance, some common proteinase inhibitors used in affinity purifications are effective chelators (e.g. 1,10-phenanthroline). Additionally, binding of HaloTag fusion proteins to Magne HaloTag resin is inhibited by the common protease inhibitor Pefabloc (4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride, AEBSF). Conducting RNA-binding assays under SEQRS experimental buffer conditions may be helpful to optimize these conditions. In the case of protein complexes that assemble in the absence of RNA, it is prudent to first perform an experiment, such as coimmunoprecipitation, under the conditions used in the SEQRS experiment to confirm that

the conditions are compatible with physical interaction. As an instructive example, we provide the detailed protocol used to express and purify Pum and Nos proteins for SEQRS (18). We feel that users, especially those new to strategies to purify RNase free recombinant RBPs, will find this protocol helpful. This purification strategy will likely require iterative optimization of the parameters described above for different systems.

2.3 Selection and library design

The SEQRS *in vitro* selection strategy is modeled after SELEX (Fig. 2) (13, 14). Briefly, DNA oligonucleotides encoding a randomized 20-mer region are transcribed using T7 RNA polymerase. The resulting pool of RNAs is incubated with purified recombinant protein immobilized on magnetic resin. After repeated washing, bound RNAs are thermally eluted and converted into double-stranded DNA using reverse transcription followed by PCR. This enrichment procedure is repeated for five to seven cycles. We analyze the sequences present after different numbers of cycles using Illumina-based technology. Sequencing adapters and unique barcodes are added prior to high-throughput sequencing. The use of barcodes allows sequencing of multiple samples in parallel, and enables deconvolution of multiplexed data. SEQRS selection involves generating an appropriate randomized RNA library, performing the binding reactions, and conversion of the selected RNAs to cDNAs for sequencing, and each of these steps should be carefully designed in the experimental strategy.

The first step in SEQRS selection is the generation of a library that serves as a template for transcription of the RNA library (Fig. 3). A randomized region is flanked by two constant domains: one provides the binding site for a custom sequencing primer and the opposing end contains a binding site for the reverse transcription primer. After selection, a barcode is attached to the 3' end that will determine the length of sequencing read required in an experiment. The central randomized region is used to probe for the binding site of the protein or protein complex tested. There are a number of sources that can synthesize the randomized regions. We recommend using a vendor that can guarantee the percentages of random bases in the library are close to 25% for each base. Significant biases in the randomized region will diminish library complexity, although these can be accounted for computationally. Hand-mixed random oligonucleotides are an additional potential solution. The library does not need to be purified beyond simple desalting.

The length of the central randomized region should be chosen by balancing two considerations. First is the need for adequate library complexity. A priori, libraries with longer central regions should contain more complex structured elements. Therefore, longer randomized regions allow selection of a wider range of recognition elements. Conversely, the experimental design must be feasible and the second consideration is how the level of library complexity affects practical factors. For example, a longer randomized region requires the addition of a larger amount of RNA in the initial binding reaction for the complexity of the library to be represented. We routinely use a 20-mer randomized region in our experiments, as this level of library complexity (4^{20} sequences) can be sampled in the initial binding reaction through the addition of several hundred nanograms of RNA. However, other groups using a similar approach have reported successful experiments with libraries of 40-mer randomized regions (44). Therefore, it may not be absolutely necessary

to cover the theoretical complexity of the library in the initial binding step to successfully identify binding elements.

Each round of selection begins with *in vitro* transcribing RNA from the library (first round) or selected pool of candidates (subsequent rounds) using T7 RNA polymerase. The synthesis of small RNAs using the T7 enzyme was initially described in the 1980's (45). This enzyme and variant preparations are currently available from multiple commercial sources. However, commercial T7 RNA polymerase is expensive, and given the quantity consumed, contribute substantially to the cost of a SEQRS experiment with multiple rounds of *in vitro* transcription. Recombinant enzyme can also be expressed and purified in the lab. We have found inconsistency in some batches of commercial enzyme, and the activity of enzyme produced in our lab was lower than available products. An alternative approach that we have not yet tested is the use of a specific T7 RNA polymerase mutant construct that significantly improves transcription yields, particularly from “difficult” or unfavorable initial templates (46). Considering cost and activity, we have not identified T7 RNA polymerase from a low-cost supplier that has a comparable activity to T7-Flash® (Epicentre).

Transcribed RNA must then be purified before mixing with immobilized proteins or protein complexes, and different strategies are available. Trizol® extraction is among the most common methods of RNA extraction. Still, we generally use commercial RNA purification kits, as they are considerably faster and effective. The Genejet RNA purification system (Thermo) produces high quality RNA in less than an hr, but we have had acceptable results with kits from other suppliers as well.

The purified RNA library can now be added to the magnetic resin-immobilized RBP(s). We conduct the binding reaction at room temperature for 30 min, but the temperature and incubation time are variables that may be optimized. Protein concentration should be optimized in the experiment based on the approximate binding capacity of the resin and the amount of protein required in the experiment. The goal is to minimize the amount of resin required in order to increase the stringency of binding. As mentioned above, the conditions for this initial binding step are similar to those for RNA-binding assays that require equilibrium binding conditions. It is recommended to conduct EMSAs if nothing is known regarding appropriate binding conditions. In the case of Nos and Pum, a well-documented RNA element was known, the so-called Nos Response Element, and we used this RNA as a ligand to confirm binding activity of the purified proteins and to establish optimal binding conditions. Additives may be included, such as bovine serum albumin (BSA) and tRNA as competitors or carriers, small molecules that suppress aggregation (*e.g.* Tween-20), and reducing agents (*e.g.* DTT, 2-mercaptoethanol, TCEP). Chemicals that interfere with reverse transcription should be avoided (*e.g.* residual diethylpyrocarbonate, DEPC). The RNPs with bound RNAs are isolated by immobilization of the magnetic resin, and then the resin is washed repetitively with ice-cold buffer to remove unbound RNAs. After washing, the RBP-RNA complexes are thermally eluted from the resin.

After elution, the selected RNAs must be reverse transcribed to cDNA and then amplified to dsDNA using primers corresponding to the constant regions that were engineered into the library (Fig. 3). Our standard reverse transcription reaction protocol has been optimized for

reproducibility and includes an initial 5 min incubation at 22°C followed by 1 hr at 40°C. The cDNA is amplified to dsDNA by PCR for 20-25 cycles. The goal is to obtain a clear band without over-amplification. We have found that 25 cycles is usually sufficient to recover products without over amplification. However, the number of PCR cycles should be optimized to avoid loss of library complexity by over amplification, as the dsDNA is used as the transcription template for the next round of selection. We monitor the selection progress using agarose gel electrophoresis. Given the small size of the PCR products, we have found that 2% agarose gels run with sodium borate buffer work well for visualization of the DNA. There is typically variation in the intensity of bands for different samples over the course of an experiment. However, the intensity level is not correlated with experimental success provided that a band is detected at the end of the experiment. If a sample does not have a band present, additional rounds of PCR are required.

Following PCR, there are several ways to purify the DNA. Because throughput is essential, we use a commercial kit to purify our DNA (Genejet, Thermo). Note that column binding conditions must be modified; isopropyl alcohol must be added to the DNA solution for the small product to be efficiently retained on the purification column. The elution volume should be optimized to yield a concentration of at least 100 ng/μL for the subsequent transcription reaction to be successful.

The number of rounds of selection required in an experiment is an important variable that must be determined empirically. The goal in a SEQRS experiment is to sample a broad range of bound elements of a given RBP or RBP complex. In nature, weak (μM) and medium (nM) affinity sites confer regulatory potential in addition to high affinity (sub-nM) sites. As a result, we try to balance detecting high affinity sites with retaining more weakly bound elements. We have found that in early rounds of selection, the signal-to-noise ratio is quite low, as higher affinity elements have not been amplified at the expense of the more weakly bound elements (17). We typically terminate the selection process after five rounds, when low and medium affinity sequences are still present in the enriched pool of RNAs. In principle, as selection is continued, the highest affinity sequences are most enriched and eventually dominate. We have found that one round of selection is consistently insufficient to yield biologically meaningful selection. Generally, five rounds of selection are required to obtain useful results. As a result, selection is conducted for five to seven rounds. However, the optimal number of rounds must be determined case-by-case by sequencing and analyzing selected RNAs from different rounds of selection (see below). A crucial aspect of a successful experiment is consistency at all steps of selection. This includes using frozen stock solutions of RBPs, buffers and master mixes wherever possible. Some proteins do not tolerate freezing and thawing and in rare cases must be stored at 4°C or prepared immediately before the SEQRS experiment. Similarly, RNA must be handled carefully to avoid degradation by RNases and/or divalent metal ions. Filter barrier pipette tips, gloves, and clean work areas that are treated with products that reduce RNase contamination prior to each experiment are essential.

We provide a highly optimized protocol for PUF protein complexes as an example. This should be used as a template for the development of protocols amenable to other RBPs and

complexes therein. The variables we describe above can be used to guide the optimization process.

2.3.1 Selection protocol—Buffers and reagents:

GoTaq® Green Master Mix (Promega)

Genejet PCR Purification Kit (Thermo)

AmpliScribe T7-Flash Transcription Kit (Epicentre)

Yeast tRNA (Ambion)

SEQRS Buffer (see Section 2.2.2 for composition)

RNA elution buffer

1 mM Tris pH 7.5

1. Prepare transcription template by amplifying by PCR the 20-mer library (1619) using the T7-containing reverse (1615) and forward (1614) primers at a concentration of 0.2 μ M (Fig. 3). Attach the adapters with 10-15 rounds of PCR.
2. Purify the PCR products with the Genejet PCR Purification Kit (Thermo). To ensure the small products are efficiently retained, add isopropyl alcohol to the binding buffer.
3. Transcribe the initial RNA library using 1 μ g of input dsDNA template with the AmpliScribe T7-Flash Transcription Kit (Epicentre) following manufacturer's instructions.
4. Purify the RNA transcripts with the Genejet RNA Purification Kit (Thermo).
5. For each binding reaction, add 200 ng of the purified RNA library, 200 ng of yeast tRNA competitor (Ambion) and 0.4 U of RNase inhibitor to the thawed protein-resin complex. Adjust the final volume of each solution to 100 μ L with SEQRS buffer. We conducted binding reactions with immobilized HaloTagged Pum alone and with Nos (in duplicate), HaloTagged RNA-binding defective Pum alone and with Nos, and HaloTagged Nos alone.
6. Incubate the samples at room temperature for 30 min.
7. Place the strip tubes on a 96-well magnetic block to capture the immobilized protein-RNA complexes.
8. Aspirate the reaction buffer and add 150 μ L of ice-cold SEQRS buffer to the beads and incubate for 1 min. For optimal results, the resin should not be discarded over the course of the experiment, as productive RNPs are present on the beads.
9. Repeat step 8-9 three more times for a total of 4 washes.

10. After the final wash, resuspend the resin in 20 μ L elution buffer (1 mM Tris pH 7.5) containing 10 pmol of the reverse-transcription primer (1614). Incubate samples at 65°C for 10 min to elute the RNPs, then cool on ice for 2 min.
11. Place the cooled strip tubes back on the 96-well magnetic block.
12. Carefully remove 5 μ L of the eluted sample and add it to 5 μ L of 2x ImProm-II reverse-transcription reaction (Promega).
13. Incubate the reverse transcription reaction for 5 min at 22°C and then 1 hr at 40°C.
14. Amplify the reverse transcription products. Add 10 μ L of the cDNA product as the template in a 50 μ L PCR (25 cycles) with the T7 reverse 1615 primer and the forward/reverse-transcription primer.
15. Optional QC step – visualize 5 μ L of each sample on 2% agarose gels run with sodium borate buffer stained with ethidium bromide. If faint bands or no bands are observed, add additional rounds of amplification (2-5) and visualize again.

NOTE: We use a ChemiDoc Touch (Bio-Rad) system to image
16. Purify the PCR products with a Genejet PCR Purification Kit (Thermo). Add isopropyl alcohol to the binding buffer, as before.
17. Elute the dsDNA samples into 15 μ L of water.
18. For the next round of transcription, use 5 μ L of the eluted dsDNA sample as a template in a 15 μ L transcription reaction (AmpliScribe T7-Flash Transcription Kit, Epicentre).
19. Incubate at 37°C overnight.
20. Add 1 μ L RNase-free DNase (Epicentre) to each sample and incubate at 37°C.
21. Repeat additional rounds of selection from step 5 with fresh proteins samples until the desired number of rounds has been conducted. We conducted five rounds total for the Nos-Pum SEQRS experiment.
22. Attach Illumina sequencing adapter, primer binding site, and barcode elements with identifiers for each individual experiment by PCR (1616 and 1618). Add 1616 and 1618 to a final concentration of 0.2 μ M.

2.4 Sequencing

To reduce costs, sequencing libraries are generated from a pool of uniquely labeled individual experiments. In the last step of the selection process, short six-base ‘barcode’ sequences were added to the 3’ ends for each sample as experimental identifiers. These barcodes allow multiple samples to be combined in a sequencing library by enabling deconvolution of the multiplexed library. The sequences differ from one another by at least two base pairs to reduce incorrect assignment. The sequencing depth should approach the complexity of the length of sequences being analyzed. For an RNA library with a 20-mer randomized region, this translates into a depth of approximately one million reads per

experiment. In practice, acceptable results are possible with fewer reads, but sampling of the full complexity of the RNA library will be incomplete. The targeted number of reads per experiment is one million, but in practice, this value is rarely achieved. Due to the high degree of variability in clustering on the sequencing flow cell, we routinely obtain a range of reads per sample between 50,000-2,500,000. The input material is also sequenced for normalization purposes. Sequencing the input material enables correction for biases such as coupling efficiencies that are crucial for data normalization. The number of samples that can be combined in an experiment is dependent on the number of reads yielded by the sequencing machine but can approach 100-200 on an Illumina HiSeq 2000.

The type of instrument utilized for the experiment should be chosen to fit the number of samples that will be analyzed. We have utilized Illumina technology in all of our experiments. There are strengths and weaknesses to each platform. For example, the HiSeq 2000 has the capacity for a large number of reads, but if the demand on the sequencing facility for longer read lengths is low, filling all of the lanes on a flow cell requires several weeks causing a delay in obtaining results. The NextSeq 500, which has a smaller flow cell, overcomes the need for a queue and the problem of wait times and produces acceptable results. The cost per read is comparable between HiSeq and NextSeq technology. The MiSeq is not ideal as the cost per read is higher than the aforementioned platforms.

The data processing for SEQRS is virtually identical to the cognate site identification (CSI) procedure and has recently been described in depth elsewhere (47). Briefly, three key files are yielded from data processing – a single de-multiplexed file per barcode, a list of all 10-mers and their occurrence in the raw data, and a list of 10-mers that has been normalized to a negative control such as the input material. The scripts for CSI are available from Aseem Ansari's group at University of Wisconsin, Madison (<https://biochem.wisc.edu/faculty/ansari>).

2.5 Data analysis and visualization

The data analysis and visualization steps require identifying selected motifs and displaying them graphically. The generation of motifs begins with counting individual sequences. Although the library contains random 20-mers, the sequence data are processed into shorter subsequences of length k by searching for all possible k -mer sequences. For the Nos-Pum SEQRS experiment, we searched for 10-mers. The numbers of 10-mer reads in the experimental sets are counted and normalized vs the numbers of reads in the input library to determine the relative enrichment. There are three major reasons for reducing the complexity. First, many RNA-binding proteins bind relatively short sequence elements (fewer than 10 nucleotides in length) (48). Second, searching for and counting shorter subsequences increases the redundancy. Greater numbers of occurrences of a subsequence increases the signal-to-noise ratio. Third, a subsequence of 10 nucleotides generates 4^{10} different sequences, which is roughly the limit of what is tractable on a reasonable time scale (hrs as opposed to days and weeks) with current computational capacity. After counting, typically, 100-300 of the most abundant sequences are used to generate sequence logos.

To determine the specificity of Nos and Pum individually, we examined the most abundant 10-mers for binding elements using positional weight matrices (PWMs) (Fig. 4). PWMs can

be displayed using a graphical logo where the size of any given nucleotide is proportional to the level of enrichment at each position in a binding motif. These logos can be generated using tools such as multiple em for motif elicitation (MEME) (49). This program requires a list of sequences in FASTA format and input parameters for the length of the motif and number of motifs present in the experiment. It is crucial to set the motif discovery on the sense strand only, as the tool was developed to also process motifs present in DNA.

As an example, consider the motifs present in each round of selection for Pum alone (Fig. 4A). The enrichment process yields degenerate motifs in the first two rounds that possess the 5' UGU trinucleotide typically bound by Pum and other PUF proteins. By the third round and through the fifth round, the motif we detected for Pum is remarkably similar to motifs previously reported for homologues including human Pumilio and *C. elegans* PUF-8 (17, 36, 50). In the presence of Nos, degenerate motifs persist to round 3, and the specificity of Pum only becomes apparent after the fourth round of selection where the motif appears to include upstream bases (Fig. 4B). Nos alone did not yield a discrete motif, consistent with prior observations that it is apparently binds RNA non-sequence specifically in isolation (40) (Fig. 4C). Although PWM analysis identifies high-affinity sites in our example, it does not capture alternate sites even when highly populated, nor does it detect context-dependent sequence features (51).

High-throughput sequencing allows sampling of millions of different sequence permutations. For the Nos-Pum SEQRS experiments, we analyzed ~50,000-100,000 reads per experiment resulting in 500,000-1,000,000 observed 10-mers. We utilized this depth of information to ascertain the global landscape of sequences selected by Pum compared to the Nos-Pum complex using sequence specificity landscapes (SSLs) (Fig. 5A) (51). The goal of the SSL is to generate a plot representing the relative enrichment for each sequence in a series of concentric rings (47). The innermost ring contains positions for all sequence permutations that exactly match a consensus motif provided by the user. For the Nos-Pum SEQRS experiment, the consensus sequence for Pum was UGUAWAUA or for the Nos-Pum complex, WWDUGUA (Fig. 5B-C, W=A or U and D=A, G or U). Rings radiating outward contain positions for sequences with increasing numbers of mismatches to the innermost seed motif. The z-axis (height) corresponds to the number of reads of a particular sequence, normalized to the starting library. Areas of the landscape that lack peaks have little enrichment in the dataset, whereas high peaks represent sequences that are observed many times in an experiment.

The consensus motif in the innermost ring is generally derived from PWM analysis, but it can be refined iteratively with SSL analysis. The optimization process involves generating a motif that maximizes the density of sequences with peaks in the innermost ring and minimizes large populations of sequences that do not strictly adhere to the seed motif in the outer rings. The sequences are arranged by motif in alphabetical order where permutations are allowed and then the flanking regions outside of the motif are considered. The sequences are arranged by motif and then the flanking regions are considered from 5' to 3'. For example, in the SEQRS analysis of Pum, the origin sequence is UGUAAAUAAA (Fig. 5A). This ordering system is maintained in subsequent rings first by the positions of the

mismatches and then by the substituted nucleotide at the mismatch. As a result, the order of sequences is consistent both within and between rings.

The motifs for Pum alone and the Nos-Pum complex agree well across all of the sequencing data as evidenced by the dense population of sequences in the inner most concentric ring of the motif (Fig. 5 B-C). Conversely, the degenerate C-rich motif detected for Nos alone by MEME analysis (Fig. 4C) is weakly populated, and the overall intensity of enrichment across the SSL is low (Fig. 5D). This indicates that the degenerate C-rich motif does not describe a pattern of sequence enrichment in the experiment. The examples of Pum and Nos SEQRS analysis show how SSLs enhance the utility of information obtained by PWMs through comprehensive analysis of all potential sequence permutations. The distinct specificities of Pum alone and Nos-Pum are apparent in their SSLs. Degenerate motifs yielded by MEME can be assessed and discarded as we have shown with Nos, which lacks a clear sequence preference in isolation.

3 Integration

To strengthen the results of a SEQRS experiment, multiple independent lines of evidence are highly desirable. The two main questions raised by the motif we observed for the Nos-Pum complex are whether the 3' end is less important for binding and whether there are biases in the upstream nucleotides that favor A/U sequence composition. While there are many ways to address these questions, we examined candidate sequences present in the SEQRS motif using reporter assays *in vivo*, EMSAs *in vitro*, and structural approaches with candidates from the former experiments (18). Remarkably, the *in vivo* consensus sequence for the Nos-Pum complex is situated in a known biological target (*hb*) and conformed to the consensus motif yielded by SEQRS. This served as an ideal candidate for validation purposes.

Cell-based, 3' UTR reporter gene assays have been instrumental in understanding post-transcriptional regulatory mechanisms, including for human and *Drosophila* PUF proteins. For example, the use of reporter assays has facilitated identification of repressive regions outside of the PUF domain and sites of interaction with mechanistically relevant protein partners (18, 42, 43, 52, 53). To facilitate implementation and analysis of reporter genes, we refer readers to our recent methods article, which focuses on use of dual luciferase reporter assays to measure RBP regulatory function (52). This approach involves transient transfection of two luciferase reporters. One reporter contains a putative control element for a given regulatory factor, and the other is devoid of control elements and is used as a control for normalization. After an incubation period where reporter expression is allowed to proceed, the levels of both reporters are determined contemporaneously. As the substrate and cofactor requirements of the bioluminescent reporter proteins differ; the enzymatic activity of each can be unambiguously measured in a sequential fashion. The ratio of the reporters provides an indirect but highly quantitative measure of protein expression. Reporter assays are an essential tool to determine if elements yielded by SEQRS are regulated in a cellular context. We employed this assay to validate the SEQRS-derived binding motif of the Nos-Pum protein complex (18). A *Renilla* luciferase reporter gene was created by inserting the Nos-Pum motif into the 3' UTR. This motif corresponded to the NRE-2 element (AAAUUGUACAUAAGCC, wherein the Nos-Pum derived SEQRS motif is underlined)

from the *hunchback* gene, a known target of Nos and Pum. We found that the Nos-Pum motif conferred repression or *Renilla* luciferase reporter expression when introduced into *Drosophila* S2 cells in a Nos and Pum dependent manner (18). Importantly, mutations in the Nos-Pum motif blocked regulation. Once established, the reporter assay provides a facile, powerful means to interrogate the RBP-RNA interaction and the mechanism of post-transcriptional control.

To understand why there are additional upstream nucleotides in the SEQRS motif for the Nos-Pum complex, we determined the structure of the Nos-Pum complex using X-ray crystallography (18). This experiment marks the first successful attempt to determine structurally how PUF proteins interact physically with protein partners. We found that Nos makes extensive contacts with Pum but also embraces the RNA substrate. This structure provides an explanation for why Nos and Pum fail to interact in the absence of RNA. Additionally, contacts between Nos and nucleotides upstream of the UGU suggest a mechanism for the apparent requirement of A/U bases upstream of the core Pum motif. Thus, SEQRS and structural analysis defined the Nos-Binding Site, which we then corroborated by EMSA analysis of RNA ligands with wild-type and mutant motifs.

The dataset generated by SEQRS analysis provide a powerful tool for prediction of target RNAs. One approach is to search the transcriptome to identify RNAs that bear the SEQRS motif and to map the prevalence and location of the motif. To determine whether there were differences in RNAs bound by Pum versus the Nos-Pum complex in the *Drosophila* transcriptome, we mined existing data computationally. This approach was facilitated by the fact that Pum target mRNAs were previously identified in both embryos and adults using RIP-chip (36). We reasoned that this dataset should contain a combination of Pum and Nos-Pum targets, as the conditions of the IP would likely preserve protein-protein interactions. As a first step in our analysis, we built a search model based on the SEQRS motifs and searched for either the Pum consensus element or the Nos-Pum motif using the grep command-line utility. We then determined the extent of overlap between the two lists (*ca.* 25%). The distinct patterns of regulatory elements suggested that mRNAs were usually regulated by either Pum or the Nos-Pum complex. The next challenge was to use the set of Pum targets to confirm the known biological function of the Nos-Pum complex or possibly uncover new potential biological functions. To do so, the gene lists from each search were then used for gene ontology analyses. A powerful way to identify enriched biological themes is through the use of Database for Annotation, Visualization and Integrated Discovery (DAVID) (54). This tool requires a gene list as an input and returns significantly enriched biological functions.

3. Concluding remarks

The general approach we describe is useful for understanding the specificity of either individual or multi-protein complexes assembled onto RNA. The combinatorial nature of RNA regulation is broadly recognized, yet our ability to probe RNA-binding proteins as part of multi-protein complexes is constantly evolving. A limitation of the SEQRS approach is a window of ten nucleotides. This small sequence length may be insufficient for RNPs with longer binding sites, such as Cas9, or proteins that bind to exceptionally structured elements.

An additional concern with the current protocol is the inability to quantitatively determine the efficiency of each step in the process. For example some rounds of selection disproportionately improve motif quality and expectation scores (*e.g.* Fig. 4B rd. 3 vs. 4). This could be overcome potentially through incorporation of spike-in controls at each round of the selection process (55). To date, we have applied this method to identify examples where protein partners exert substantial effects on the specificity and affinity of RNA-binding proteins (17, 18). In both cases, the action of the protein complex is linked to control of protein expression. However, complexes containing RBPs control each step of RNA function, from assembly of RNPs, to RNA processing, transport and localization, and combinatorial mechanisms are likely to be pervasive. The role of protein-protein complexes in defining specificity in these contexts is largely unclear. Additionally, many of the classic examples of cooperative interactions involve other types of associations, such as binding of small molecules that impact oligomeric states (56, 57). Comparable studies on RBPs should be tractable using the SEQRS approach with minor modifications. Placing data obtained by *in vitro* selection and sequencing into a robust biological context rests on synergy with independent lines of evidence. The integration of cell-based assays, functional genomics, and structural approaches provides a powerful level of understanding of multi-protein complexes in RNA control.

Acknowledgements

We thank Robert Wine and Shruti Sasaki for assistance in preparing the manuscript. This research was supported in part by NIGMS R01GM105707 (A.C.G.); NIH NRSA 5T32GM007544 (C.W.); a Research Scholar Grant, RSG-13-080-01-RMC, from the American Cancer Society (A.C.G.); and the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (TMTHT).

References

1. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* 2014; 15:829–845. [PubMed: 25365966]
2. Mitchell SF, Parker R. Principles and properties of eukaryotic mRNPs. *Mol. Cell.* 2014; 54:547–558. [PubMed: 24856220]
3. Iadevaia V, Gerber AP. Combinatorial Control of mRNA Fates by RNA-Binding Proteins and Non-Coding RNAs. *Biomolecules.* 2015; 5:2207–2222. [PubMed: 26404389]
4. Goldstrohm AC, Wickens M. Multifunctional deadenylase complexes diversify mRNA control. *Nat. Rev. Mol. Cell. Biol.* 2008; 9:337–344. [PubMed: 18334997]
5. Piqué M, López JM, Foissac S, Guigó R, Méndez R. A combinatorial code for CPE-mediated translational control. *Cell.* 2008; 132:434–448. [PubMed: 18267074]
6. Campbell ZT, Wickens M. Probing RNA-protein networks: biochemistry meets genomics. *Trends Biochem. Sci.* 2015; 40:157–164. [PubMed: 25636997]
7. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature.* 2008; 456:464–469. [PubMed: 18978773]
8. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T. Transcriptome-wide identification of RNA-binding protein and microRNA target. *Cell.* 2010; 141:129–141. [PubMed: 20371350]
9. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual. *Nat. Struct. Mol. Biol.* 2010; 17:909–915. [PubMed: 20601959]

10. Keene JD, Komisarow JM, Friedersdorf MB. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein. *Nat. Protoc.* 2006; 1:302–307. [PubMed: 17406249]
11. Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature protocols.* 2007; 2:1849–1861. [PubMed: 17703195]
12. Ryder SP, Recht MI, Williamson JR. Quantitative analysis of protein-RNA interactions by gel mobility shift. *Methods in molecular biology.* 2008; 488:99–115. [PubMed: 18982286]
13. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science.* 1990; 3:505–510.
14. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. *Nature.* 1990; 346:818–822. [PubMed: 1697402]
15. Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J. DNA-binding specificities of human transcription factors. *Cell.* 2013; 152:327–339. [PubMed: 23332764]
16. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, Mann RS. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell.* 2011; 147:1270–1282. [PubMed: 22153072]
17. Campbell ZT, Bhimsaria D, Valley CT, Rodriguez-Martinez JA, Menichelli E, Williamson JR, Ansari AZ, Wickens M. Cooperativity in RNA-protein interactions: global analysis of RNA binding. *Cell Rep.* 2012; 1:570–581. [PubMed: 22708079]
18. Widmann CA, Qiu C, Arvola RM, Lou TF, Killingsworth J, Campbell ZT, Tanaka Hall TM, Goldstrohm AC. Drosophila Nanos acts as a molecular clamp that modulates the RNA-binding and repression activities of Pumilio. *eLife.* 2016 In press.
19. Wilinski D, Qiu C, Lapointe CP, Nevil M, Campbell ZT, Tanaka Hall TM, Wickens M. RNA regulatory networks diversified through curvature of the PUF protein. *Nat. Commun.* 2015; 6
20. Campbell ZT, Valley CT, Wickens M. A protein-RNA specificity code enables targeted activation of an endogenous human. *Nat. Struct. Mol. Biol.* 2014; 21:732–738. [PubMed: 24997599]
21. LeGendre JB, Campbell ZT, Kroll-Conner P, Anderson P, Kimble J, Wickens M. RNA targets and specificity of Staufen, a double-stranded RNA-binding protein in *Caenorhabditis elegans*. *J. Biol. Chem.* 2013; 288:2532–2545. [PubMed: 23195953]
22. Zhang B, Gallegos M, Puoti A, Durkin E, Fields S, Kimble J, Wickens M. A conserved RNA binding protein that regulates patterning of sexual fates in the *C. elegans* hermaphrodite germ line. *Nature.* 1997; 390:477–484. [PubMed: 9393998]
23. Murata Y, Wharton R. Binding of pumilio to maternal hunchback mRNA is required for posterior patterning in *Drosophila* embryos. *Cell.* 1995; 80:747–756. [PubMed: 7889568]
24. Edwards TA, Pyle SE, Wharton RP, Aggarwal AK. Structure of Pumilio reveals similarity between RNA and peptide binding motif. *Cell.* 2001; 105:281–289. [PubMed: 11336677]
25. Wang X, Zamore PD, Hall TM. Crystal structure of a Pumilio homology domain. *Molecular cell.* 2001; 7:855–865. [PubMed: 11336708]
26. Wang X, McLachlan J, Zamore PD, Hall TM. Modular recognition of RNA by a human pumilio-homology domain. *Cell.* 2002; 110:501–512. [PubMed: 12202039]
27. Miller MT, Higgin JJ, Hall TM. Basis of altered RNA-binding specificity by PUF proteins revealed by crystal structures of yeast Puf4p. *Nat. Struct. Mol. Biol.* 2008; 15:397–402. [PubMed: 18327269]
28. Zhu D, Stumpf CR, Krahn JM, Wickens M, Hall TM. A 5' cytosine binding pocket in Puf3p specifies regulation of mitochondrial. *Proc. Natl. Acad. Sci. U S A.* 2009; 106:20192–20197. [PubMed: 19918084]
29. Wang Y, Opperman L, Wickens M, Hall TM. Structural basis for specific recognition of multiple mRNA targets by a PUF. *Proc. Natl. Acad. Sci. U S A.* 2009; 106:20186–20191. [PubMed: 19901328]
30. Valley CT, Porter DF, Qiu C, Campbell ZT, Hall TM, Wickens M. Patterns and plasticity in RNA-protein interactions enable recruitment of multiple proteins through a single site. *Proc. Natl. Acad. Sci. U S A.* 2012; 109:6054–6059. [PubMed: 22467831]

31. Gerber AP, Herschal D, Brown PO. Extensive association of functionally and cytologically related mRNAs with Puf. *PLoS Biol.* 2004; 2:16.
32. Suh N, Crittenden SL, Goldstrohm A, Hook B, Thompson B, Wickens M, Kimble J. FBF and its dual control of *gld-1* expression in the *Caenorhabditis elegans* germline. *Genetics.* 2009; 181:1249–1260. [PubMed: 19221201]
33. Goldstrohm AC, Seay DJ, Hook BA, Wickens M. PUF protein-mediated deadenylation is catalyzed by Ccr4p. *J. Biol. Chem.* 2007; 282:109–114. [PubMed: 17090538]
34. Hook BA, Goldstrohm AC, Seay DJ, Wickens M. Two yeast PUF proteins negatively regulate a single mRNA. *J. Biol. Chem.* 2007; 282:15430–15438. [PubMed: 17389596]
35. Kershner AM, Kimble J. Genome-wide analysis of mRNA targets for *Caenorhabditis elegans* FBF, a conserved stem cell regulator. *Proc. Natl. Acad. Sci. U S A.* 2010; 107:3936–3941. [PubMed: 20142496]
36. Gerber AP, Luschnig S, Krasnow MA, Brown PO, Herschlag D. Genome-wide identification of mRNAs associated with the translational regulator. *Proc. Natl. Acad. Sci. U S A.* 2006; 103:4487–4492. [PubMed: 16537387]
37. Irish V, Lehmann R, Akam M. The *Drosophila* posterior-group gene *nanos* functions by repressing *hunchback* activity. *Nature.* 1989; 338:646–648. [PubMed: 2704419]
38. Sonoda J, Wharton RP. Recruitment of Nanos to *hunchback* mRNA by Pumilio. *Genes Dev.* 1999; 13:2704–2712. [PubMed: 10541556]
39. Wharton R, Struhl G. RNA regulatory elements mediate control of *Drosophila* body pattern by the posterior morphogen *nanos*. *Cell.* 1991; 67:955–967. [PubMed: 1720354]
40. Curtis D, Treiber DK, Tao F, Zamore PD, Williamson JR, Lehmann R. A CCHC metal-binding domain in Nanos is essential for translational regulation. *EMBO J.* 1997; 16:834–843. [PubMed: 9049312]
41. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA 3rd, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods.* 2009; 6:343–345. [PubMed: 19363495]
42. Weidmann CA, Raynard NA, Blewett NH, Van Etten J, Goldstrohm AC. The RNA binding domain of Pumilio antagonizes poly-adenosine binding protein and accelerates deadenylation. *RNA.* 2014; 20:1298–1319. [PubMed: 24942623]
43. Weidmann CA, Goldstrohm AC. *Drosophila* Pumilio protein contains multiple autonomous repression domains that regulate mRNAs independently of Nanos and brain tumor. *Mol. Cell. Biol.* 2012; 32:527–540. [PubMed: 22064486]
44. Lambert NJ, Robertson AD, Burge CB. RNA Bind-n-Seq: Measuring the Binding Affinity Landscape of RNA-Binding Proteins. *Methods Enzymol.* 2015; 558:465–493. [PubMed: 26068750]
45. Milligan JF, Groebe DR, Witherell GW, Uhlenbeck OC. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA. *Nucleic Acids Res.* 1987; 15:8783–8798. [PubMed: 3684574]
46. Guillerez J, Lopez PJ, Proux F, Launay H, Dreyfus M. A mutation in T7 RNA polymerase that facilitates promoter clearance. *Proc Natl Acad Sci U S A.* 2005; 102:5958–63. [PubMed: 15831591]
47. Tietjen JR, Donato LJ, Bhimisaria D, Ansari AZ. Sequence-specificity and energy landscapes of DNA-binding molecules. *Methods Enzymol.* 2011; 497:3–30. [PubMed: 21601080]
48. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecnas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LO, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR. A compendium of RNA-binding motifs for decoding gene regulation. *Nature.* 2013; 499:172–177. [PubMed: 23846655]
49. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 2006; 34:W369–373. [PubMed: 16845028]
50. Laver JD, Li X, Ray D, Cook KB, Hahn NA, Nabeel-Shah S, Kekis M, Luo H, Marsolais AJ, Fung KY, Hughes TR, Westwood JT, Sidhu SS, Morris Q, Lipshitz HD, Smibert CA. Brain tumor is a

- sequence-specific RNA-binding protein that directs maternal mRNA clearance during the *Drosophila* maternal-to-zygotic transition. *Genome biology*. 2015; 16:94. [PubMed: 25962635]
51. Carlson CD, Warren CL, Hauschild KE, Ozers MS, Qadir N, Bhimsaria D, Lee Y, Cerrina F, Ansari AZ. Specificity landscapes of DNA binding molecules elucidate biological function. *Proc. Natl. Acad. Sci. U S A*. 2010; 107:4544–4549. [PubMed: 20176964]
52. Van Etten J, Schagat TL, Goldstrohm AC. A guide to design and optimization of reporter assays for 3' untranslated region. *Methods*. 2013; 63:110–118. [PubMed: 23643865]
53. Van Etten J, Schagat TL, Hrit J, Weidmann CA, Brumbaugh J, Coon JJ, Goldstrohm AC. Human Pumilio proteins recruit multiple deadenylases to efficiently repress. *J. Biol. Chem*. 2012; 287:36370–36383. [PubMed: 22955276]
54. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID. *Nat. Protoc*. 2009; 4:44–57. [PubMed: 19131956]
55. Chen K, Hu Z, Xia Z, Zhao D, Li W, Tyler JK. The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. *Mol Cell Biol*. 2016:662–667.
56. Edelstein SJ. Cooperative interactions of hemoglobin. *Annu. Rev. Biochem*. 1975; 44:209–232. [PubMed: 237460]
57. Antonini E, Wyman J, Brunori M, Bucci E, Fronticelli C, Rossi-Fanelli A. Studies on the Relations between Molecular and Functional Properties Of Hemoglobin. IV. The Bohr Effect in Human Hemoglobin Measured By Proton Binding. *J. Biol. Chem*. 1963; 238:2950–2957. [PubMed: 14081909]

Figure 1.
An overview of the steps required for a SEQRS experiment focused on a protein complex.

Figure 2.

The SEQRS workflow. The DNA library contains a randomized region of 20 bp (yellow) between two 20 bp constant regions (blue and orange). Transcription with T7 RNA polymerase yields a pool of RNAs used as the starting material for selection. A sufficient quantity of RNA to cover all possible 20-mer sequences is incubated with recombinant protein. The protein (red circles) is immobilized on resin to enable capture of the RNA-protein complexes. After repeated wash steps, the remaining RNA is heat eluted from the protein. The RNA is reverse transcribed (RT) back to DNA using a primer complementary to the constant region denoted as primer B. The single-stranded DNA is amplified using a primer set that reintroduces the T7 promoter (green boxes) upstream of primer A. After the desired number of rounds, aliquots of double-stranded DNA are amplified with primers containing adapters appropriate for Illumina sequencing and unique barcode identifiers.

1619 - SEQRS 20-mer library template (N is any base in the random region)
 5' CTGATCCTACCATCCGTGCT NNNNNNNNNNNNNNNNNNNN CACAGCTTCGTACCGAGCGG

Primers for selection cycles
 1614 - Forward (and reverse transcription primer)
 5' CTGATCCTACCATCCGTGCT
 1615 - Reverse (contains the T7 promoter)
 5' TAATACGACTCACTATAGGCGCGCTCGGTACGAAGCTGTG

Primers for use at the sequencing step
 1616 - Forward (contains Illumina flow cell adapter and binding site for 1617)
 5' AATGATACGGCGACCAACGAGATCTGCTCTCCGATCTCTGATCCTACCATCCGTGCT
 1618 - Barcode primer (contains Illumina adapter and a custom six base identifier)
 5' CAAGCAGAAGACGGCATACGAGATNNNNNTCTTCCGATCCGCTCGGTACGAAGCTGTG

Custom sequencing primer
 1617
 5' GCTCTTCGATCTCTGATCCTACCATCCGTGCT

Final library
 5' AATGATACGGCGACCAACGAGATCTGCTCTCCGATCTCTGATCCTACCATCCGTGCTNNNNNNNNNN
 NNNNNNNNNCACAGCTTCGTACCGAGCGGATCGGAAGANNNNNNATCTCGTATGCCGTCTTCTGCTTG

Figure 3.

SEQRS library and primer sets used for selection and sequencing. In the 1619 library template, two constant regions that are adjacent to the random portion of the library (yellow) are colored blue at the 5' end and orange at the 3' end. The library is assembled with amplification of 1619 using primers 1614 and 1615 (areas of sequence complementarity to the library template are colored in the same way). This results in the addition of the T7 RNA polymerase promoter (magenta) at the 3' end of the library. After transcription and selection, 1614 is used as a reverse transcription primer. Prior to sequencing, each sample is amplified with primers 1616 and 1618 to attach the Illumina adapters (dark blue) and the binding site for the custom sequencing primer 1617 (green) to produce the final sequencing library.

Figure 4. Sequence logos for Pum and the Nos-Pum complex at different rounds of selection. PWMs were generated using MEME (49). The height of each letter is proportional to the information content of that base at the indicated positions. A – Logos for Pum after each round of selection. B – Logos for the Nos-Pum complex after each round of selection. C - Logos for Nos after each round of selection.