Supplementary Information

In vivo mRNA display enables Large-scale Proteomics by Next Generation Sequencing

P. Oikonomou[1,2,3], R. Salatino[2], S. Tavazoie[1,2,3*]

[1]Department of Biological Sciences, Columbia University, New York City, New York, United States

[2]Department of Systems Biology, Columbia University, New York City, New York, United States

[3]Department of Biochemistry and Molecular Biophysics, Columbia University, New York City, New York, United States

* Correspondence: P.O.: po2194@columbia.edu; S.T.: st2744@columbia.edu

Table of Contents

www.pnas.org/cgi/doi/10.1073/pnas.2002650117

Methods

Plasmid Construction

The backbone for all *in vivo* mRNA display plasmids and respective controls is pSH100 (URA3 selection marker; Addgene #45930) (1). MS2 coat protein (MCP) was PCR amplified from pSH100 (1), while stem-loop sequences from pDZ415 (1) were ordered as Blocks from IDT; defective MCP (MCP*) mutations were introduced via overlap PCR. For cloning, PCR insert fragments for MCP variants and a Gateway cloning ccdB cassette were amplified using Q5 polymerase (NEB M0491). The backbone was digested using restriction enzymes (Dataset S1), combined with PCR inserts for Gibson Assembly according to the manufacturer (NEB E2611), and transformed into One Shot ccdB Survival Cells (Invitrogen A10460). The resulting Destination vectors (pIPOIVMD156,160,155) allow for Gateway cloning of ORFs flanked by Gateway attL sites in to the display constructs. The *in vivo* mRNA display constructs in this study are under the control of a *MET25* promoter (induced in methionine dropout media). Super folder GFP (2) and mCherry (1) were amplified with flanking attB sites and Entry Vectors were generated via a BP reaction. Individual yeast ORFs were amplified from the ORFeome collection purchased from Dharmacon (3) using the available flanking attB sites. Plasmids expressing SAM2- and ARC40-GFP fusion baits were constructed using Gibson assembly based on the pSH62 (4) backbone (HIS3 selection marker; pIIVMD495, 496). DH5α competent cell were used for all bacterial transformations (NEB C2989K). See Dataset S1A-B for a complete list of plasmids in this study, as well as primers and restriction enzymes used for construction.

Yeast Strains

The BY4742 S288c MATα laboratory deletion strain was used as the starting strain for all strains harboring *in vivo* mRNA display constructs. All plasmids were transformed using the LiAc-PEG-ssDNA method (5), and selected in 2% glucose –URA dropout media. EY0986 MATa deletion strains expressing genomically integrated SAM2- and ARC40-GFP fusions (6) were purchased from Thermo Fischer (Catalog no. 95701; HIS3 selection marker). Plasmids expressing SAM2- and ARC40-GFP fusion baits were transformed into BY4741 S288c MATa laboratory deletion strain and selected on 2% glucose –HIS dropout media. For diploid strains, MATa and MATα haploids were mated in YPD at 30 ºC with vigorous shaking for 1-2 hours and selected on 2% glucose –URA, -HIS dropout media. Dropout media supplement powders from ForMedium and

US Biological were used interchangeably. Single strain selections were plated on appropriate 2% hard agar SC dropout plates. See Dataset S1C for a list of all strains used in this study.

In vivo mRNA display Library Generation

E.coli strains from the yeast ORFeome plasmid collection were outgrown, pooled and pelleted. Pooled plasmid was extracted from the pellets using a Qiagen Maxi-prep kit (#12963). Yeast ORFs were PCR amplified using the attB1 and attB2 flanking sequences. A two-step recombination reaction (first a BP recombination into pDONR221, followed by an LR reaction) was used to transfer the sequence into the Gateway cloning site of the in vivo mRNA display Destination Vector (pIVMD156). BP and LR reactions were transformed in DH5α cells (NEB C2989K) and colonies were selected in semi-liquid soft agarose gel (7) (0.3% Lonza Seaprep #50302) LB media with the appropriate antibiotics (kanamycin and ampicillin for BP and LR reactions respectively). More than 1 million colonies were collected for the BP reaction (~200x coverage) and over 125,000 colonies (~25x coverage) for LR reactions. The final *in vivo* mRNA display library was transformed into BY4742 using the LiAc-PEG-ssDNA method (5), and selected in 2% glucose SC–URA semi-liquid soft agarose gel (7) (0.3% Lonza Seaprep #50302). Over 500,000 colonies were collected, outgrown for 6 hours in SC-URA and stored at -80 ºC in 15% glycerol media. Mated diploid libraries were selected in 2% glucose SC–URA-HIS semi soft agar (7) at similar coverage. For both bacterial and yeast libraries, colony counts were assessed by plating a dilution on hard agar SC dropout media. See Dataset S1D for a list of all libraries generated in this study.

Yeast cell culture

S. cerevisiae strains were cultured in the appropriate SC dropout media (-HIS, -URA or –HIS and -URA) supplemented with 2% glucose at 30 ºC and shaken at 220rpm. Overnight cultures were induced by seeding 0.1 OD_{600}/ml into a new liquid culture with a similar SC dropout media additionally lacking methionine (-MET). Strains were outgrown for 6-8 hours to 0.6-0.8 OD_{600}/ml and collected by centrifugation, washed twice with ultrapure water, and split in aliquots equivalent to 10 to 40 OD_{600} units of cultured cells. Pelleted cells were flash frozen in a dry ice ethanol bath and stored at -80 ºC until further processing.

For in vivo mRNA display yeast libraries, biological replicates were independently revived from frozen stock and outgrown in semi-liquid soft agarose gel as colonies to avoid any growth biases. All resulting colonies were pooled and outgrown in liquid SC dropout media as described above.

Excess Coat Protein

Unless otherwise noted, we provided an excess of MS2 coat protein for all high throughput experiments in order to titrate any non-specific interactions. To this end, we mixed into each sample an excess culture of yeast cells that express an MCP fusion that does not display its own mRNA. Additionally, the MCP is not isolated specifically in any given protein assay and its mRNA is not processed during first strand and second strand synthesis.

For excess coat protein expression, we constructed three strains (scIVMD115, 118, 217: MCP-mCherry-FLAG, MCP-mCherry-MYC, and MCP-BFP-HA). Upon induction, the equivalent of 30 OD_{600} units of cells were mixed with 10 OD_{600} of *in vivo* mRNA display library cells immediately prior to or immediately after freezing. If purifying a FLAG- or MYC-tagged *in vivo* mRNA display library, the anti-FLAG or anti-MYC tagged excess coat proteins were excluded, respectively. For 6xHIS and GFP tag purifications, all three strains were mixed in equal proportions.

Moreover, lysis and all purification steps were performed at 4 ºC since increased temperatures also compromise precision, possibly due to partner exchange (Fig. S10-11).

Non-Specific Functional Controls for *in vivo* mRNA display

We included a set of *in vivo* mRNA display constructs in every library that function as internal negative and positive controls for a given protein purification assay. Their mRNA frequencies provide a background with respect to which we normalize the frequencies of each ORF (Fig. S8-S10). For that, we chose a small set of reporter genes and peptides that should not participate in any biological interactions inside the cell. These control ORFs included GFP (2), mCherry (1), BFP (Addgene #44839), acGFP (pBI-CMV2; Clontech), Firefly Luciferase, Renilla Luciferase (psiCHECK-2; Promega) as well as short peptides derived from these reporter genes. All ORFs were cloned in vectors with an N-terminal MCP, a downstream SL and various purification tags (MYC, FLAG, 6xHIS, HA-tag) such that a subset of them works as a non-specific control set for every protein purification. Moreover, during anti-GFP magnetic bead purifications, the GFP construct works as a positive control for the assay. Additionally, we designed a smaller set of 7 control proteins that harbored an MCP but no SL such that their mRNA progenitors function as non-displaying controls. These proteins are variants of mCherry with 6 additional bases on the 5' and 3' ends of the ORF, so that they can be identified during sequencing. Control strains were added at various concentrations to represent a wide range of construct frequencies and allow us to assess any representation biases and determine read depth cut offs. Overall, the controls

represented 2-5% of the total processed cell culture. See Dataset S1E for a complete list of strains and stoichiometry of the control mix.

Additionally, these constructs allow us to assess RNA integrity and the efficiency of our assay post purification but prior to sample preparation for sequencing by means of Quantitative PCR (see relevant Qunatitative PCR Method section).

Whole cell lysate preparation

Frozen cell pellets were re-suspended in 750 µl of ice cold Lysis Buffer(8) (20 mM HEPES pH 7.5, 140 mM KCl, 1.5 mM $MgCl_2$, 1% Triton X-100, 1× Complete Mini Protease Inhibitor EDTA-free, 0.2 U/µl SUPERase RNase Inhibitor) and added on top of 250µl of pre-chilled acid washed glass beads (Sigma G8772) . After this point, samples were kept at 4 ºC throughout all purification steps. The samples were homogenized (9) in a Fast-Prep24 5G instrument (10 rounds of a 30 sec disruption pulse at 6m/s followed by 5 minutes of rest in contact with ice cold ethanol packs in between disruptions). Glass beads were removed by a 1 min centrifugation at 7,000 g, and lysate was transferred to a new tube and further cleared by a 30 second spin at 11,000 g. We set aside roughly 100 µl of the resulting sample, which is referred to as the lysate.

***In vivo* mRNA display Library Purification & Protein bait purification**

We used magnetic beads for all tagged protein purifications. Purifications were performed at 4 ºC. For 6x HIS tagged proteins, we used His-Tag Isolation Dynabeads (Invitrogen 10103D). For MYC tagged proteins, we used Anti-c-Myc Beads (Pierce #88842). For FLAG tagged proteins, we used Anti-FLAG M2 beads (Sigma M8823). For GFP and mCherry tagged proteins, we used GFP- and RFP-Trap beads respectively (ChromoTek gtma and rtma). All beads were washed 3 times before use with 200 µl of Wash Buffer (50mM Sodium Phosphate pH8, 300 mM NaCl, 0.01% Tween-20, 0.02 U/µl SUPERase RNase Inhibitor). Beads in 200µl of Wash Buffer were added to 400 µl of whole cell lysate and incubated on a roller (8 min for 6xHIS, 30 min for MYC, 2 hours for FLAG, 1 hour for GFP and RFP tagged proteins). Then beads were washed 4 times with 300 µl Wash Buffer and resuspended in 100 µl of Storage Buffer (same as Wash Buffer with 0.2 U/µl SUPERase RNase Inhibitor).

For 6xHIS purifications, 10 mM of imidazole was added in both Lysis and Wash Buffers. Additionally, proteins were eluted in 300mM of imidazole and re-purified using a fresh aliquot of

His-Tag Isolation Dynabeads (Invitrogen 10103D). To that end, 100 µl of eluted sample was mixed with 1000 µl of Wash Buffer and incubated on a roller, and washed 4 times.

See Dataset S1D for a list of purification assays each library was submitted to.

Plate reader measurements

When purification involved cultures predominantly expressing GFP and/or mCherry, 10-20µl of the lysate and 10-50µl of the purified sample was added to 150µl of ultrapure water on a flat-bottom 96-well plate (Corning #3631). A SynergyMx (Biotek; Winooski, VT) plate reader was used to measure fluorescence (mCherry: 544/612; GFP: 500/530 excitation/emission nm).

Crude Mitochondrial Isolation

We processed frozen library pellets (scIVMD580) equivalent to 20 OD_{600} units of cells per replicate. There was no excess coat protein culture added to the yeast libraries for these purifications. We performed a crude mitochondrial isolation using a commercially available kit from Sigma (MITOISO3). The frozen library cell pellets were resuspended in 2 ml of ice cold Buffer A (Sigma B3311 with 0.02 U/µl SUPERase RNase Inhibitor) and incubated for 15 min at 30 ºC with gentle shaking. Next, they were centrifuged at 1,500 g for 5 min, resuspended in 1ml of Buffer B (Sigma B3186 with 0.02 U/µl SUPERase RNase Inhibitor), and supplemented with 40 units of Lyticase Solution (Sigma L2524). Spheroplasts were formed by incubating at 30 ºC with gentle shaking for roughly 10 min (until OD_{600} decreases to 30% of the initial value). The reaction was stopped by centrifuging at 1,200 g for 5 min at 4 ºC. Spheroplasts were homogenized in 1 ml of Storage Buffer (Sigma S9689 with 0.2 U/µl SUPERase RNase Inhibitor) with 10 strokes using a pre-chilled sterile Dounce homogenizer at 4 ºC (Sigma T2690; P1110). To remove nuclei, samples were centrifuged at 600 g for 10 min at 4 ºC. The supernatant was transferred to a new Eppendorf tube and centrifuged at 6,500 g for 10 min at 4 ºC. The supernatant was saved for further processing (75 µl for RNA extraction). Storage buffer was added to the pellet and the sample was centrifuged at 6,500 g for an additional 10 min at 4 ºC. Supernatant was discarded and the final pellet was saved for further processing.

RNA extraction

We extracted RNA from all protein samples (50 µl of whole cell extract; up to 100µl of purified protein bound on beads; 75 µl of 6,500 g supernatant from crude mitochondrial isolation; or the complete 6,500 g pellet) using TRIzol (Invitrogen 15596026). We added 750 µl of TRIzol reagent

to each sample, vortexed and incubated at RT for 5 min. We added 150 µl of chloroform (Sigma C2432), vortexed and incubated for 2 min at RT. Samples were centrifuged at 12,000 g for 15 min at 4 ºC. The aqueous phase was transferred to a new tube and mixed with an equal volume of 100% ethanol. The mixture was applied to a spin column, washed and concentrated according to the manufacturer's recommendations (Zymo Research, RNA Clean & Concentrator R1015). RNA from lysates was eluted in 50 µl, while RNA from purified protein samples was eluted in 15 µl of RNase free water with 0.2 U/µl SUPERase RNase Inhibitor.

cDNA synthesis

For whole cell extract samples, we used 8 µg of purified total RNA as input. For purified protein samples, we used the whole sample. We treated extracted RNA with dsDNase (Thermo Scientific EP0771) at 37 ºC for 5 min in a 20 µl reaction (volumes doubled from manufacturer's recommendation). DNase treated RNA was reverse transcribed using Maxima H Minus RT (Thermo Scientific EP0752) and a construct specific primer (prIVMD212) binding downstream of the in vivo mRNA display construct ORF (Fig. S3). The samples were incubated at 65 ºC for 5 min with RT primer (prIVMD212) and dNTP mix per the manufacturer's recommendations. RT Buffer and RT Enzyme were added and samples were incubated for 30 min at 50 ºC. The reaction was terminated at 85 ºC for 5 min. When random hexamers were used (Fig.1D-E) the 50 ºC incubation was preceded by a 10 min incubation at 25 ºC. Next, we hydrolysed remaining RNA by adding 8 µl of 500 mM EDTA and 8 µl of 1N NaOH per 40 µl of 1st Strand Synthesis samples and incubating at 65 ºC for 15 min. cDNA was cleaned and concentrated using a Zymo Research spin column kit (D4013) by adding 7 volumes of binding buffer and washing twice. Samples were eluted in 20 µl of DNase free water.

For second strand synthesis, we performed a PCR amplification using construct specific primers upstream and downstream of the *in vivo* mRNA display ORF (prIVMD113 & prIVMD212, Fig. S3) and PrimeSTAR GXL DNA polymerase (Clontech R050B). We set up 50µl reactions according to the manufacturer's recommendations for the Rapid PCR protocol (2x enzyme) with annealing at 58 ºC and 90 second extension for 8 cycles. Second strand synthesis samples were purified using a Zymo Research spin column kit (D4013) by adding 5 volumes of binding buffer and washing twice. Samples were eluted in 20 µl of RNase free water.

Quantitative PCR

We assessed extracted RNA for quality and *in vivo* mRNA display efficiency using qPCR. Quantitive PCR (PerfeCTa SYBR Green FastMix, QuantaBio 95073-012; on an Applied Biosystems QuantStudio5 384-well instrument) was used to determine the relative abundance of mCherry and GFP transcript in each sample. Protein purification experiments were designed such that either GFP or mCherry is co-purified in the experiment (specific positive control) and the other is washed away (non-specific reference). We calculated a ΔC_t value for each sample and a $-\Delta\Delta C_t$ between purified sample and input lysate. Therefore the Log_2 Fold Enrichment is:

$$-\Delta\Delta C_t = \left[C_t^{Specific} - C_t^{Non-Specific}\right]^{IP} - \left[C_t^{Specific} - C_t^{Non-Specific}\right]^{LYS}.$$

For random hexamer RT (Fig.1D), the relative abundance to ACT1 was quantified as a reference. For each sample, technical duplicate replicate measurements were made, and if they were inconsistent, they were repeated for quadruplicates resulting in the reporting of average values and standard deviations as error bars (Fig. S1, S10, S11, S12). For Fig. 1D-E, purification experiments were conducted in biological replicates, and averages of biological replicates are reported as bars and replicate values as grey dots. See Dataset S1A for qPCR primers.

In vivo mRNA display Library Sequencing Preparation

Restriction Enzyme Digestion: To prepare samples for sequencing we used 20 µl of double stranded cDNA as input. Each sample was split in half for two 20 µl restriction enzyme digestion reactions (Fig. S3-5). One half was treated with HinP1I (NEB R0124) and AciI (NEB R0551), while the other half was treated with MspI (NEB R0106) and HpyCH4IV (NEB R0619). Each 20 µl digestion contained 1 µl of each restriction enzyme and 2 µl of CutSmart Buffer (NEB B7204) and was incubated at 37 ºC for 3-6 hours and heat inactivated at 65 ºC for 20 min. Reactions were combined and purified using a Zymo Research spin column kit (D4013) by adding 7 volumes of binding buffer and washing twice. Each sample was eluted in 9 µl of DNase free water. All restriction enzymes generate a CG overhang used for linker ligation.

Y-Linker Annealing: Per 8 samples, we used 8 µL of HPLC purified 100 µM YCG5 and 8 µL of 100 µM YCG3 primer, combined with 2 µl of DNase free water and 2 µL of 10x Annealing Buffer (1 M NaCl, 100 mM Tris-HCl ph8, 10 mM EDTA pH8) (10). Samples were placed in a thermocycler and with a starting temperature of 94 ºC and slowly cooled to 25 ºC (reduced by 2 ºC every 30 seconds).

Y-Linker Ligation: For each sample, 9 µl of cleaned up digestion was mixed with 2.5 µl of annealed Y-Linker, 1 µl of Quick Ligase (NEB M2200) and 12.5 µl of 2x Quick Ligase Buffer. The reaction was incubated at room temperature for 10 min. We added 1 µl of 500 mM EDTA to stop the reaction and purified using a Zymo Research spin column kit (D4013).

Multiplexing and NGS adapter addition: We set out to amplify the ligated 5' and 3' ends on each ORF. For 5' fragments, one primer lands on the universal sequence of *in vivo* mRNA display constructs *upstream* of the ORF (Fig. S3) and the other lands on the ligated Y-Linker. For 3' fragments, one primer lands on the universal sequence of *in vivo* mRNA display constructs *downstream* of the ORF and the other lands on the ligated Y-Linker. In the process of amplification, Illumina adapters are added for Next Generation Sequencing and samples are multiplexed. We perform this amplification in two rounds of PCR amplification.

PCR amplification Round 1: During the first round, custom-designed identifying index sequences of varying length were included on the end of the PCR that would be sequenced, as well as partial Illumina adapter sequences on both ends. The custom-designed indexes are used to multiplex samples but also to stagger the library sequences to achieve the necessary variability in the initial bases (because all our library sequences included an identical universal adaptor at each end of the ORF). Two PCRs are set up for every sample: one amplifying the 5' end of every ORF and one amplifying the 3' end of every ORF in the library. For each 5' or 3' ORF PCR, one primer lands on the universal construct sequence that is upstream or downstream of the 5' or 3'end of each ORF, respectively, while the other PCR primer lands on the Y-Linker (Fig. S3). See Dataset S1A for primer sequences. PCR amplification was performed for 7 cycles using the Q5 High Fidelity Polymerase (NEB M049; a two PCR program with annealing at 62 ºC for the first 3 cycles and 67 ºC for the remaining 4 cycles and 2 minute extension throughout). Reactions were set up as per the manufacturer's recommendations. Upon completion of thermocycling reaction, we combined 5' and 3' PCRs and used Ampure XP beads for DNA cleanup (A63881, Beckman Coulter, Brea, CA) at a 1.7x ratio. We eluted fragments in 25 µl of water.

PCR amplification Round 2: During the second round, Illumina Adapter sequences were extended while Illumina indexes were added to each sample for further multiplexing. See Dataset S1A for primer sequences. Reactions were set up using the Q5 High Fidelity Polymerase (NEB M04) as per the manufacturer's recommendations. A 40 µl reaction was set up side by side with a smaller 10 µl reaction additionally including ROX Low Reference Dye (KK4602, Kapa Biosystems, Wilmington, MA) and SYBR dye (EvaGreen; 31000, Biotium, Fremont, CA) in 1x concentrations. The smaller reaction was split in two technical replicates and cycled on a qPCR machine.

Amplification was observed to determine the number of cycles needed or the amplification to reach the exponential phase (or roughly 30% of the maximum signal) and the number of cycles were noted (11). The remaining 40 µl PCR reaction was thermocycled for the same number of cycles as noted from the qPCR. A two step PCR program was employed for both qPCR and regular PCR with annealing at 65 ºC for the first 3 cycles and 68 ºC for the remaining cycles and 90 second extension throughout. We used Ampure XP beads (A63881, Beckman Coulter, Brea, CA) at a 1.3x ratio. We eluted fragments in 25 µl of water.

The concentration of each sample was measured using the Qubit dsDNA HS Assay Kit (Q32854, Invitrogen) and/or the Agilent Bioanalyzer High Sensitivity DNA kit (5067-4626, Agilent, Santa Clara, CA). Libraries were sequenced for 75 cycles with the NextSeq 500/500 High Output Kit v2.5 (20024906, Illumina) either single-end or pair-end depending on the needs of other libraries on the lane. For pair-end sequenced samples, cycles were allocated as follows: 58 cycles read 1, 17 cycles read 2. Only read 1 was utilized for data analysis (read 2 contains a universal Y-Linker sequence).

In vivo mRNA display Sequencing Data Analysis

After de-multiplexing Illumina indexes, we used Cutadapt (12) to trim low-quality sequences and trim universal 3' adapter sequences corresponding to the Y-Linker primers. Cutadapt was also used to de-multiplex internal custom indexes, and remove universal 5' adapter sequences. Surviving reads of sufficient length (>20 nt) were mapped to the 5' and 3' ends of all yeast ORFs using Bowtie (13). If the sum of all reads mapping to the 5' end (or the 3' end) of each ORF i is ${R_i}^{5'}$ $(or\ {R_i}^{3'})$, we can calculate an average log frequency, for the 5' fragments, the 3' fragments and the average:

$$f_i^{5'} = \log_2 \left[\frac{R_i+1}{\sum_j R_j}\right]^{5'},$$

$$f_i^{3'} = \log_2 \left[\frac{R_i+1}{\sum_j R_j}\right]^{3'},$$

$$f_i = 0.5\left(f_i^{5'} + f_i^{3'}\right),$$

where f_i represents the frequency for ORF i , and $\sum_j R_j$ represents the total sum of reads for all the 5' (or 3') end reads . If C is the set of constructs in the non-specific functional control library, we calculate a log normalized frequency, h^i, for each ORF i with respect to the control set:

$$h_i{}^{5'} = f_i{}^{5'} - \frac{\sum_{k\in C} f_k{}^{5'}}{|C|},\ h_i{}^{3'} = f_i{}^{3'} - \frac{\sum_{k\in C} f_k{}^{3'}}{|C|}\ ,\ h_i = f_i - \frac{\sum_{k\in C} f_k}{|C|}.$$

The Display Score for each ORF between two matched samples is calculated as the difference of the log normalized frequencies between the samples. For example, the Display Scores for ORF i for a protein purification experiment are:

$$DS_i{}^{5'} = h_i^{5',Pur} - h_i^{5',Lys}\ ,$$

$$DS_i{}^{3'} = h_i^{3',Pur} - h_i^{3',Lys}\ ,$$

$$DS_i = h_i^{Pur} - h_i^{Lys}\ ,$$

where h_i^{Pur} is the log normalized frequency in the purified protein sample and h_i^{Lys} is the log normalized frequency in the input whole cell extract. An ORF is considered to be present in an experiment only if it had more than 8 reads in either the input or the purified sample (for Fig. 2D-G a threshold of 4 reads was set based on the distribution of reads). An ORF is considered present in an assay with replicates, if it is present in half or more of the 3' and 5' samples of all the replicates. The Display score represents an enrichment ($DS_i > 0$) or depletion ($DS_i < 0$) of the reads of ORF i in the purified sample compared to the lysate with respect to the non-specific functional controls. The distribution of the non-specific functional controls can be used to calculate a z Score for the Display Score:

$$Z_i = \frac{DS_i}{\sigma_{DS}^{controls}}\ ,$$

where $\sigma_{DS}^{controls}$ is the standard deviation of the Display Scores of the non-specific functional controls, and $\mu_{DS}^{controls} = 0$ by definition. Z Scores for biological replicate experiments were averaged using the Stouffer rule. Display Score p-values for biological replicates (Fig. 2D-G, Fig. 3 and Fig. 4) were calculated by comparing the distribution of $DS_i^{5'}$ and $DS_i^{3'}$ measurements for

every ORF to the distribution of Display Scores for all the non-specific functional controls using a Mann-Whitney U test.

Western Blots

We treated 10µl of Lysate and 40% of the purified protein bound beads in 2xSDS Sample Buffer (Invitrogen Invitrogen LC2676) buffer and 10% β- mercaptoethanol at 95 ºC for 10 min. The beads were separated on a magnetic stand. Each sample was split in half and each half was loaded on an Invitrogen WedgeWell 8 to 16% Tris-Glycine Mini Gel (Invitrogen XP08162BOX) for electrophoresis. Resolved proteins were transferred onto a nitrocellulose membrane, followed by 1 hour blocking with 1% milk TBS-T. One membrane was incubated overnight at 4 ºC with one of the primary antibodies against GFP (Rat monoclonal ChromoTek 3H9; 1:1000), and the other against RFP (Mouse monoclonal ChromoTek 6G6; 1:2000). The membranes were washed and incubated at room temperature for 1 hour with HRP-conjugated secondary anti-Mouse and anti-Rat (Jackson Immuno Research 115035072 and 112035072). Proteins in blots were detected using the KwikQuant Detection Kit (Kindle Biosciences, R1004). RFP gels were stripped and re-probed with an α-Tubulin antibody conjugated to HRP (Rat monoclonal YOL1/34; Santa Cruz Biotechnology 53030; 1:500) and visualized again. Western Blot images were converted to greyscale and image colors were inverted.

Mass Spectrometry

Immunoprecipitation and in-gel digestion: SAM2- and ARC40-GFP were purified as described above from in vivo display libraries using anti-GFP magnetic beads. Samples were processed in biological duplicate. Protein-bound beads were washed three times with ultrapure water. Samples were processed at the Proteomics and Macromolecular Crystallography Core Facility at Columbia University Medical Campus. Immunoprecipitated samples were separated on 4-12% gradient SDS-PAGE, and stained with SimplyBlue (Thermo fisher Scientific). Protein gel slices were excised and in-gel digestion was performed. Gel slices were washed with 1:1 (Acetonitrile: 100 mM ammonium bicarbonate) for 30 min, Gel slices were then dehydrated with 100% acetonitrile for 10 min until gel slices were shrink and excess acetonitrile was removed and slices were dried in speed-vac for 10 min at no heat. Gel slices were reduced with 5 mM DTT for 30 min at 56 ºC in an air thermostat and chilled to room temperature, then alkylated with 11 mM IAA for 30 min in the dark. Gel slices were washed with 100 mM ammonium bicarbonate and 100 % acetonitrile for 10 min each. Excess acetonitrile was removed and dried in speed-vac for 10 min at no heat and

gel slices were rehydrated in a solution of 25 ng/µl trypsin in 50 mM ammonium bicarbonate on ice for 30 min on ice. Digestions were performed overnight at 37ºC in an air thermostat. Digested peptides were collected and further extracted from gel slices in extraction buffer (1:2 vol/vol) 5% formic acid/acetonitrile) at high speed shaking in an air thermostat. Supernatant from both extractions were combined and dried down in a speed-vac. Peptides were dissolved in 3% acetonitrile/0.1% formic acid.

LC-MS/MS analysis: Thermo Scientific™ UltiMate™ 3000 RSLCnano system and Thermo Scientific EASY Spray™ source with Thermo Scientific™ Acclaim™ PepMap™100 2 cm x 75 µm trap column and Thermo Scientific™ EASY-Spray™ PepMap™ RSLC C18 50 cm x 75 µm ID column were used to separate desalted peptides with a 5-30% acetonitrile gradient in 0.1% formic acid over 100 min at a flow rate of 250 nL/min. The column temperature was maintained at a constant 50 ˚C during all experiments. Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer was used for peptide MS/MS analysis. Survey scans of peptide precursors were performed from 400 to 1500 m/z at 120K FWHM resolution (at 200 m/z) with a 2 x 105 ion count target and a maximum injection time of 50 ms. The instrument was set to run in top speed mode with 3 s cycles for the survey and the MS/MS scans. After a survey scan, tandem MS was performed on the most abundant precursors exhibiting a charge state from 2 to 6 of greater than 5 x 103 intensity by isolating them in the quadrupole at 1.6 Th. CID fragmentation was applied with 35% collision energy and resulting fragments were detected using the rapid scan rate in the ion trap. The AGC target for MS/MS was set to 1 x 104 and the maximum injection time limited to 35 ms. The dynamic exclusion was set to 45 s with a 10 ppm mass tolerance around the precursor and its isotopes. Monoisotopic precursor selection was enabled.

Data Analysis: Raw mass spectrometric data were processed and searched using the Sequest HT search engine within the Proteome Discoverer 2.2 (PD2.2, Thermo Fisher) with a reference Saccharomyces cerevisiae proteome database downloaded from SGD. The default search settings used for protein identification in PD2.2 searching software were as follows: two mis-cleavages for full trypsin with fixed carbamidomethyl modification of cysteine and oxidation of methionine and demaidation of asparagine and glutamine and acetylation on N-terminal of protein were used as variable modifications. Identified peptides were filtered for maximum 1% false discovery rate using the Percolator algorithm in PD 2.2. PD2.2 output combined folder uploaded in Scaffold (Proteome Software) for data visualization. Spectral counting was used for analysis to compare samples and p-values for the enrichment of every protein were calculated using Fisher's exact test for the combined counts of every protein in the SAM2 and the ARC40 samples.

GO Term Analysis for Crude Mitochondrial Isolation

We considered mutually exclusive GO Term categories. For cytosolic, cytoplasmic and nuclear proteins, we filtered out the genes common to other organelle and membrane fractions (GO:0016020, 'GO:0005739', 'GO:0005740', 'GO:0007005', 'GO:0005773', 'GO:0005777', 'GO:0016020', 'GO:0005783', 'GO:0005794', 'GO:0005635', 'GO:0005618', 'GO:0009277', 'GO:0005811', 'GO:0005768', 'GO:0005886', 'GO:0005743', 'GO:0005741', 'GO:0005759', 'GO:0005758') and vice versa. P-values for enrichments and depletions were calculated using the hypergeometric test between the number of ORFs with significant Display Scores in each category compared to the significant Display Scores present in the assay (Fig. 3). We calculated calculate enrichment of genes in organelle and membrane categories with respect to cytosolic (GO:0005829) proteins.

Supplementary Datasets

Supplementary Dataset S1

(Dataset S1A) Catalogue of primers used for *in vivo* mRNA display processing and sequencing, primers for plasmid construction, and primers for Quantitative PCR analysis.

(Dataset S1B) Catalogue of plasmids used in this study.

(Dataset S1C) Catalogue of strains used in this study.

(Dataset S1D) Catalogue of *in vivo* mRNA display libraries generated for this study.

(Dataset S1E) Non-specific functional controls for *in vivo* mRNA display.

Supplementary Dataset S2

Assessment of *in vivo* mRNA display precision NGS sequencing results. Log_2 Reads for Lysates and Purified Samples, Log Normalized frequencies, Display Scores and Display z Scores per each ORF. ORFs specific and non-specific to each purification are noted. Data corresponds to Fig. 2A-C.

(Dataset S2A) anti-HIS Purification (scIVMD385).

(Dataset S2B) anti-MYC Purification (scIVMD385).

(Dataset S2C) anti-FLAF Purification (scIVMD385).

Supplementary Dataset S3

Yeast *in vivo* mRNA display library purification NGS sequencing results. Log_2 Reads for Lysates and Purified Samples, Log Normalized frequencies, Display Scores and Display z Scores per each ORF. Data corresponds to Fig. 2D-G.

(Dataset S3A) Replicate 1 (scIVMD263).

(Dataset S3B) Replicate 2 (scIVMD263).

(Dataset S3C) Replicate 3 (scIVMD263).

(Dataset S3D) Replicate 4 (scIVMD263).

(Dataset S3E) Average Display Scores, Average Display z Scores, Display p-values and q-values.

Supplementary Dataset S4

Yeast *in vivo* mRNA display library crude mitochondrial purification NGS sequencing results. Log_2 Reads for Lysates and Purified Samples, Log Normalized frequencies, Display Scores and Display z Scores per each ORF. Data corresponds to Fig. 3A-E.

(Dataset S4A) Replicate 1 (scIVMD580).

(Dataset S4B) Replicate 2 (scIVMD580).

(Dataset S4C) Replicate 3 (scIVMD580).

(Dataset S4D) Average Display Scores, Average Display z Scores, Display p-values, q-values and whether a gene belongs to any of the categories specific to the crude mitochondrial enrichment according to the Gene Ontology, Huh et al., Morgenstern et al, and UniProt annotations.

Supplementary Dataset S5

Yeast *in vivo* mRNA display library negative control co-purification NGS sequencing results. Log_2 Reads for Lysates and Purified Samples, Log Normalized frequencies, Display Scores and Display z Scores per each ORF. Data corresponds to Fig. 4A-E.

(Dataset S5A) Replicate 1 (scIVMD263, no bait purification, only *in vivo* mRNA display).

(Dataset S5B) Replicate 2 (scIVMD263, no bait purification, only *in vivo* mRNA display).

(Dataset S5C) Replicate 3 (scIVMD263, no bait purification, only *in vivo* mRNA display).

(Dataset S5D) Replicate 4 (scIVMD580, GFP bait).

(Dataset S5E) Replicate 5 (scIVMD580, GFP bait).

(Dataset S5F) Average Display Scores, Average Display z Scores, Display p-values and q-values.

Supplementary Dataset S6

Yeast *in vivo* mRNA display library SAM2 co-purification NGS sequencing results. Log_2 Reads for Lysates and Purified Samples, Log Normalized frequencies, Display Scores and Display z Scores per each ORF. Data corresponds to Fig. 4A-E .

(Dataset S6A) Replicate 1 (scIVMD292, SAM2-GFP bait, genomically integrated).

(Dataset S6B) Replicate 2 (scIVMD579, SAM2-GFP bait, episomally expressed).

(Dataset S6C) Replicate 3 (scIVMD579, SAM2-GFP bait, episomally expressed).

(Dataset S6D) Replicate 4 (scIVMD292, SAM2-GFP bait, genomically integrated).

(Dataset S6E) Average Display Scores, Average Display z Scores, Display p-values and q-values.

Supplementary Dataset S7

Yeast *in vivo* mRNA display library ARC40 co-purification NGS sequencing results. Log_2 Reads for Lysates and Purified Samples, Log Normalized frequencies, Display Scores and Display z Scores per each ORF. Data corresponds to Fig. 4A-E.

(Dataset S7A) Replicate 1 (scIVMD294, ARC40-GFP bait, genomically integrated).

(Dataset S7B) Replicate 2 (scIVMD578, ARC40-GFP bait, episomally expressed).

(Dataset S7C) Replicate 3 (scIVMD578, ARC40-GFP bait, episomally expressed).

(Dataset S7D) Replicate 4 (scIVMD294, ARC40-GFP bait, genomically integrated).

(Dataset S7E) Average Display Scores, Average Display z Scores, Display p-values and q-values.

Supplementary Figures

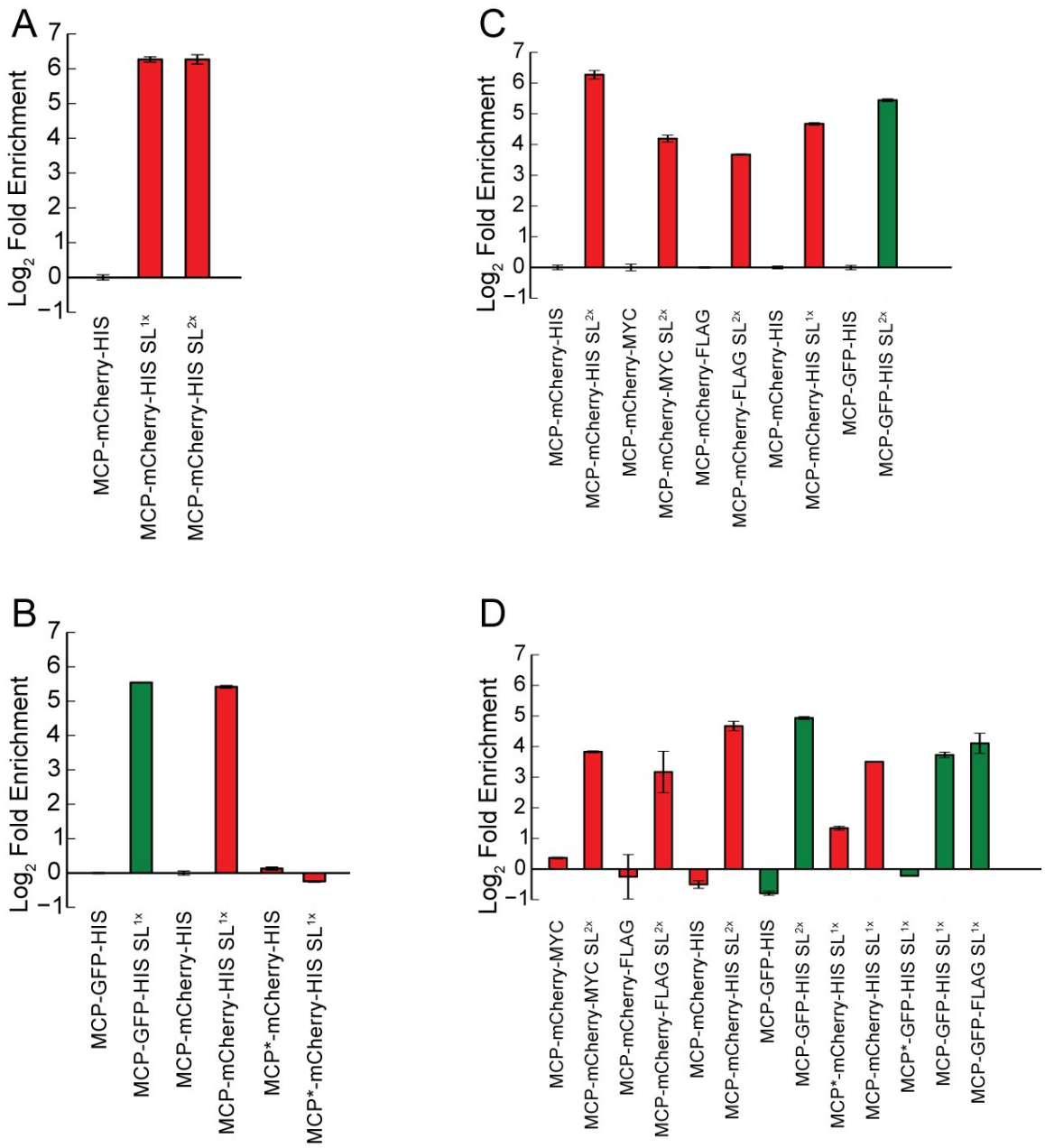


Figure S1. in vivo mRNA Display proteins co-purify their cognate mRNA for a variety of constructs and purification tags

(A-C) Log Fold Enrichments for purified proteins were calculated with respect to a reference (ACT1) in each purified sample and normalized to the construct with no hairpin loop. (A) MCP fusion constructs (mCherry) with no hairpin, one and two stem- loops (SLs). Samples with SLs display significantly more than the no-stem-loop sample. There is no significant difference between one and two stem-loops. (B) MCP-mCherry (red) and -GFP (green) constructs with no stem-loop, as well as their *in vivo* mRNA display counterparts, and defective MCP fusions (MCP*).

While the presence of a SL allows for *in vivo* mRNA display in the MCP constructs, it has no effect for MCP*. (C) *In vivo* mRNA display is independent of purification tags used. Constructs (mCherry and GFP) with no stem-loop and a single hairpin purified with anti-HIS, -MYC, -FLAG, -RFP, and -GFP magnetic beads. (D) Similar to C, but Log Fold Enrichments were calculated with respect to a housekeeping gene in each purified sample and normalized to the lysate. qPCR averages are shown as bars and SD of technical replicates as errors.

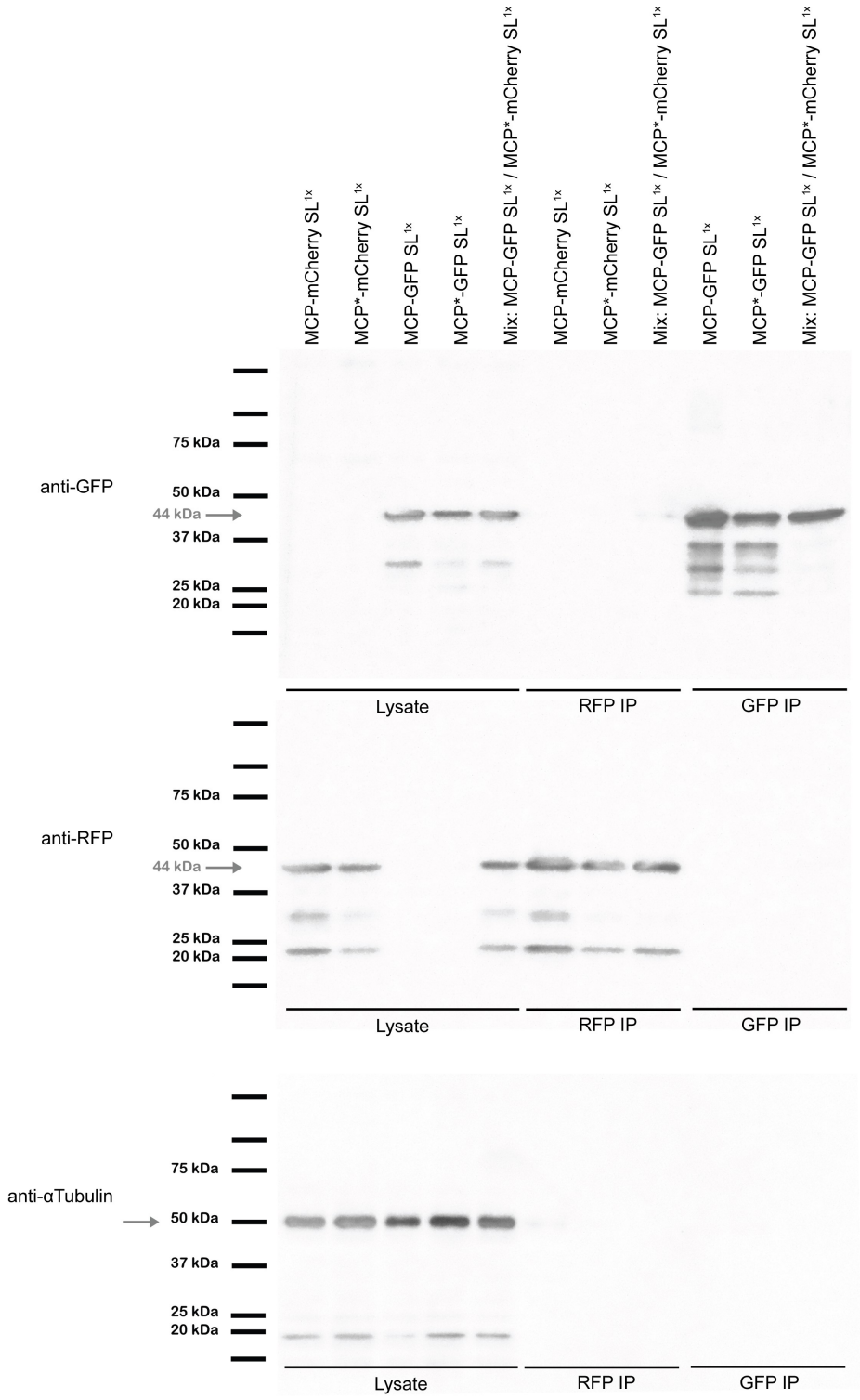


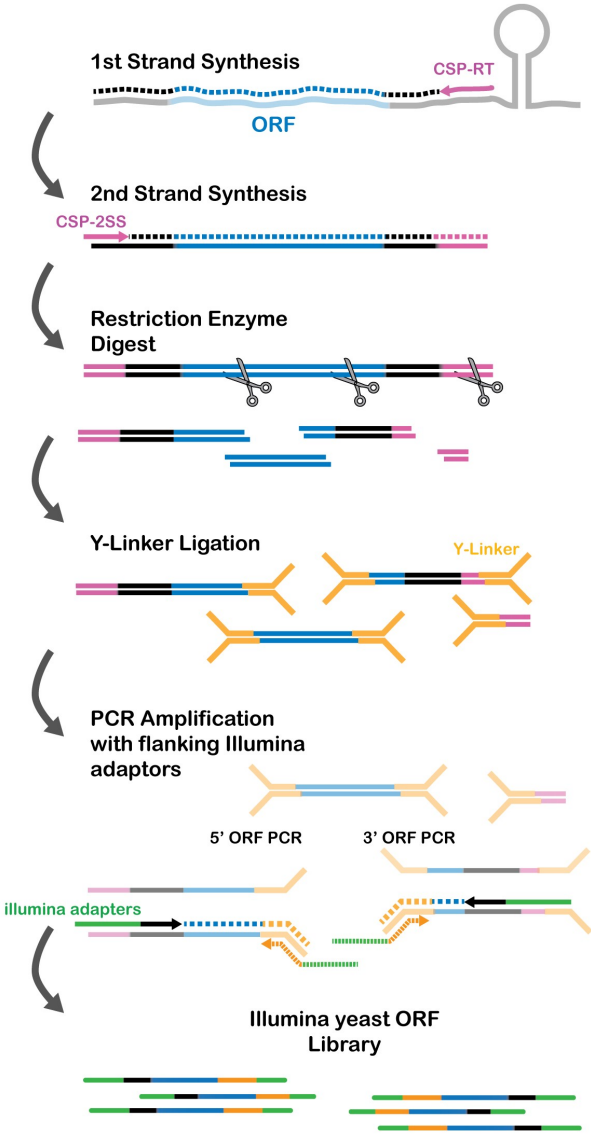
Figure S2. Western blots for Fig.1D-E

Complete images of samples presented in Fig.1D-E. Samples were run on an SDS-PAGE gel and probed with GFP, RFP and α-Tubulin antibodies as described in Methods.

Library Preparation:

Process lysate and displayed RNA in parallel to generate libraries for illumina sequencing.

Using a y-Linker based method, only 5′- and 3′-ends of the ORFs will be quantified.



RT with Construct Specific Primer (CSP)
- treat with DNAse I @37°C for 5min
- inactivate with 10mM DTT @55°C for 5min
- incubate with CSP @65°C for 5min
- reverse transcibe with Maxima H @55°C for 30min
- innactivate @85°C for 5min
- denature with NaOH
- spin-column clean-up (Zymo-5)

2nd Strand Synthesis with CSP
- 8 cycles of GXL Polymerase extension using an upstream CSP
- spin-column clean-up (Zymo-5)

RE Digestion
- split sample in two and set up two digestions:
  - HindP1I and AciI
  - MspI and HpyCH4IV
- digest @37°C for 3hrs, inactivate @65°C for 15min
- combine & spin-column clean-up (Zymo-5)

y-Linker ligation
- ligate with Quick Ligase for 10min @25°C
- innactivate with 100mM EDTA
- spin-column clean-up (Zymo-5)

1st PCR amplification: partial illumina adaptors
- split sample in two and set up two PCRs with a y-Linker primer and:
  - a CSP for the 5′-end of the ORF
  - a CSP for the 3′-end of the ORF
- internal staggered indexes and partial illumina adaptors added at this step
- 8 cycles of PCR; Q5 Polymerase
- combine & clean-up with 1.7x Ampure XP beads

2nd PCR amplification: complete library prep
- illumina indexes added at this step and adaptors are completed
- determine number of cycles: cycles that correspond to 1/3 of max intensity measured by qPCR & addition of SYBR/ROX
- Q5 Polymerase
- combine & clean-up with 1.3x Ampure XP beads

Figure S3. Pipeline for high-throughput sequencing of in vivo mRNA library

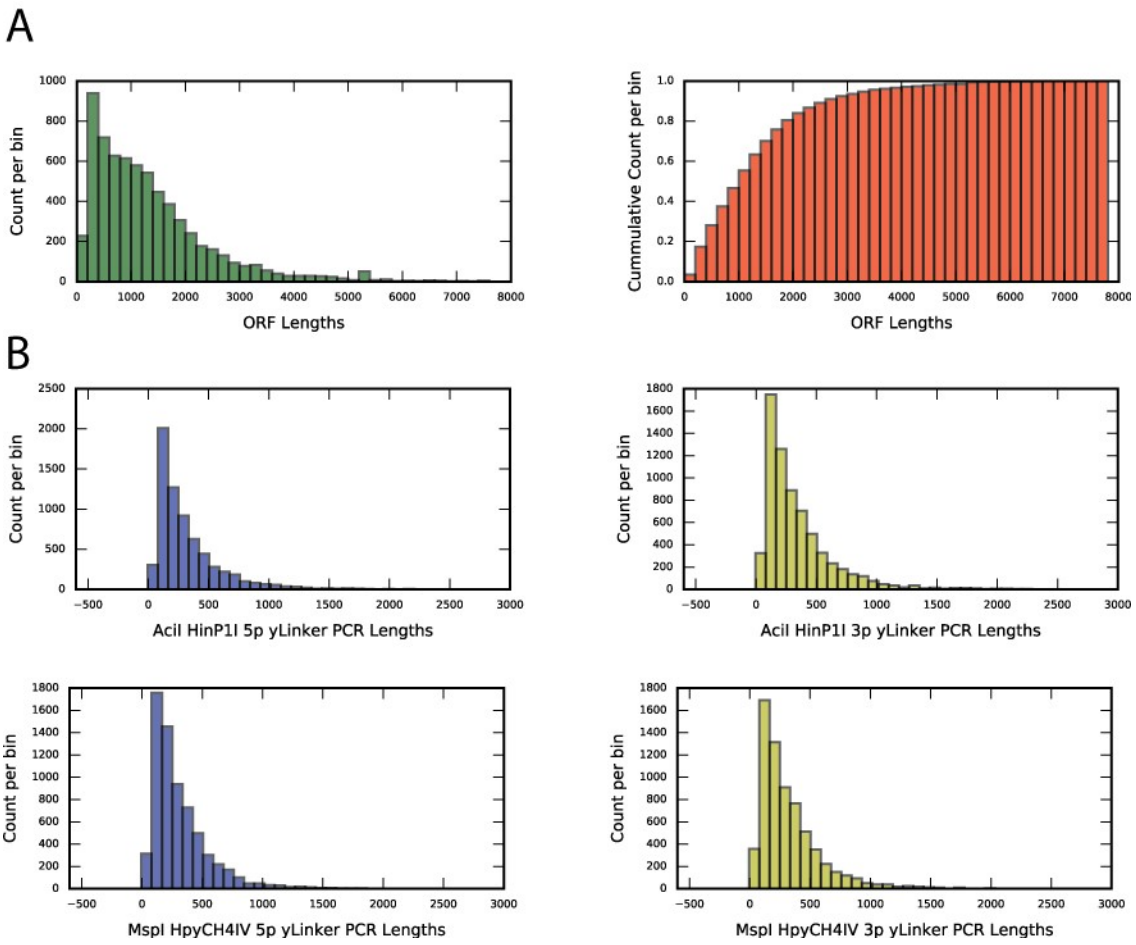


Figure S4. Restriction Enzyme digestion generates a tighter distribution of fragment lengths for the yeast proteome

(A) Distribution of ORF lengths for the yeast proteome (B) distribution of 3' and 5' fragments after cDNA synthesis and RE digestion with the two enzyme mixes (top: AciI and HinP1I; bottom: MspI and HpyCH4IV). Histograms are plotted on the left, while cumulative distributions are plotted on the right.

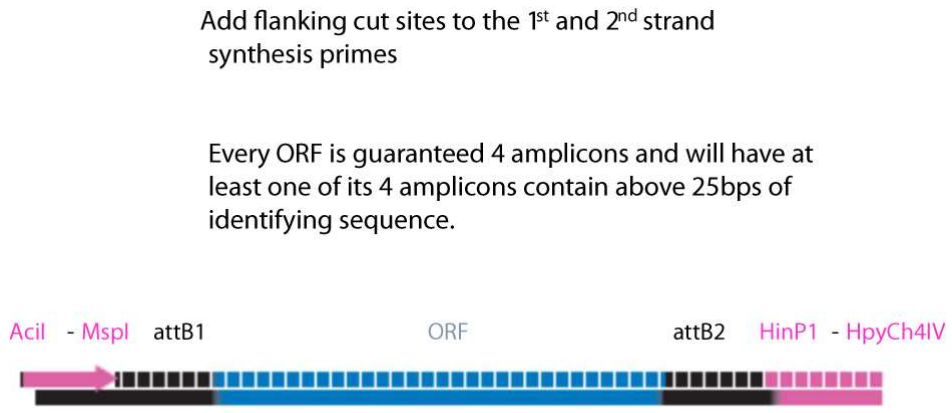


Figure S5. Restriction Enzymes in universal sequences flanking *in vivo* mRNA display ORFs.

Introduction of additional cut sites flanking each ORF to ensure representation of every yeast protein during sequencing preparation.

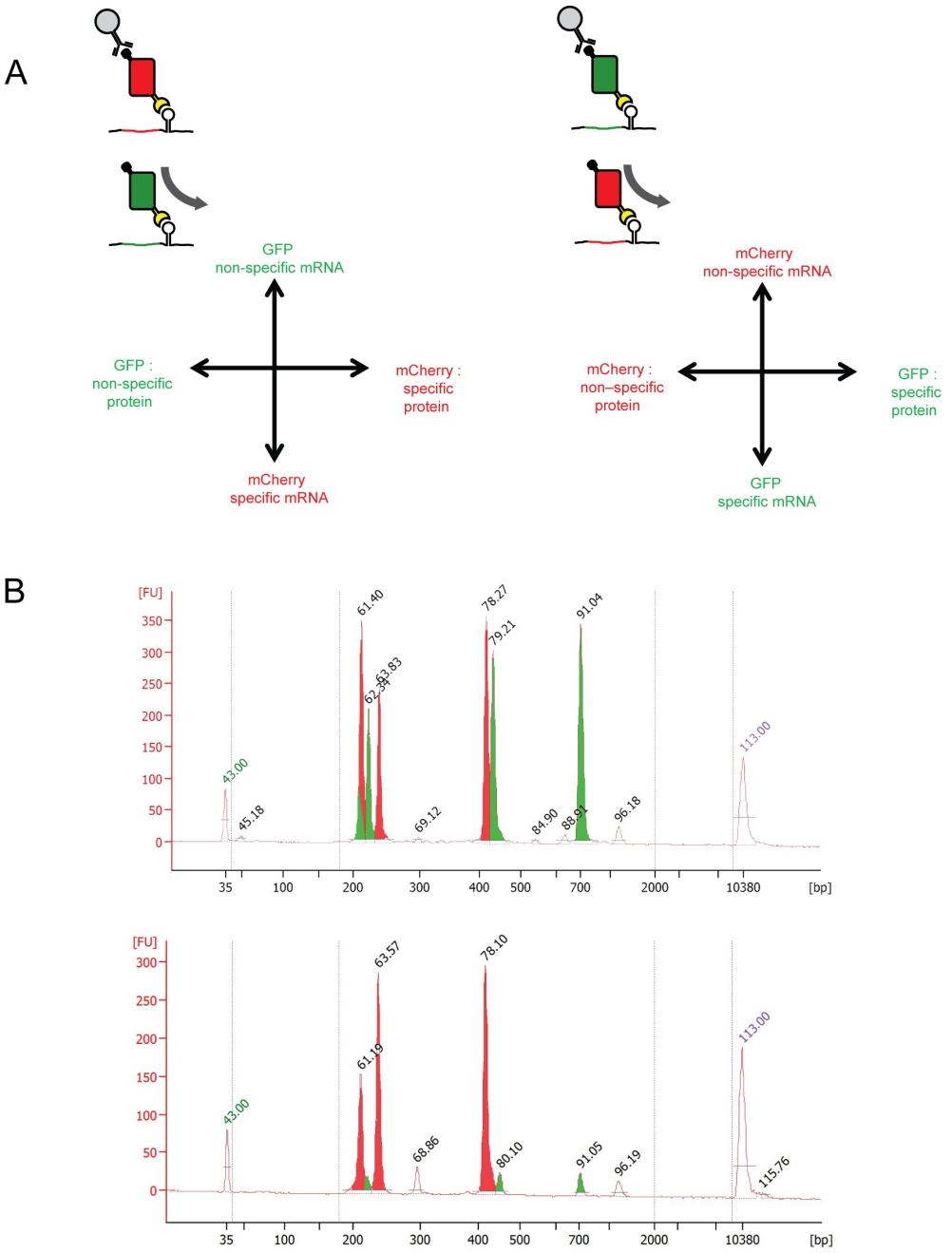


Figure S6. One-on-one completion of *in vivo* mRNA display constructs: scheme and sequencing preparation fragment enrichment.

(A) Specific and non-specific mRNA for two construct purification experiments in Fig. 1E. (B) Bioanalyzer quantification of fragments from the two color competition in Fig.1E. RNA libraries were prepared according to the sequencing pipeline. mCherry corresponding fragments are shown in red and GFP fragments in green. The top panel corresponds to the frequency of

fragments from the lysate while the bottom corresponds to the purified sample. mCherry fragments are ~8x enriched with respect to GFP fragments in agreement with the qPCR data.

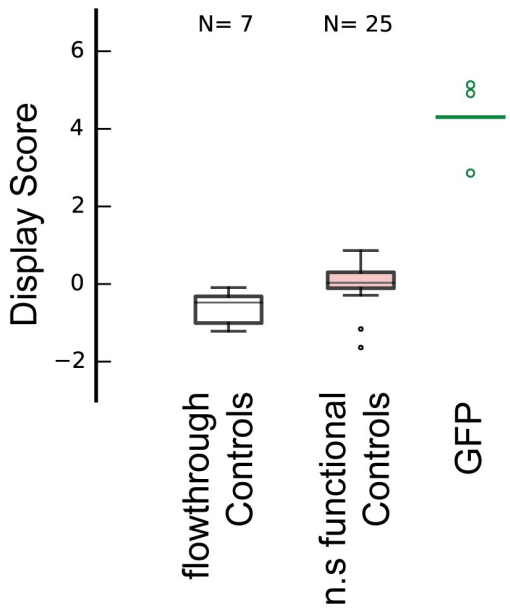


Figure S7. GFP Purification from a library with 25 non-specific functional controls

GFP mRNA fragments are enriched in the purified sample compared to the non-specific functional controls and 7 flow-through control mRNAs. Boxplot distributions are shown. The box extends from the lower to the upper quartile values, while whiskers extend 1.5×(Q3-Q1) outside the box and outliers are shown as individual points.

Display Quantification:

Demultiplex the samples: map illumina and internal staggered indexes.

Trim CSP sequences and any y-Linker sequence that might appear inside the read length.

Map to a database of all 5′- and 3′- ORF sequences and calculate counts for both RNA samples (e.g. IP and Lysate). Calculate a read frequency for every ORF (5′ and 3′ separately). For each ORF, *j*, Take the average of the $\log_2$-frequencies:

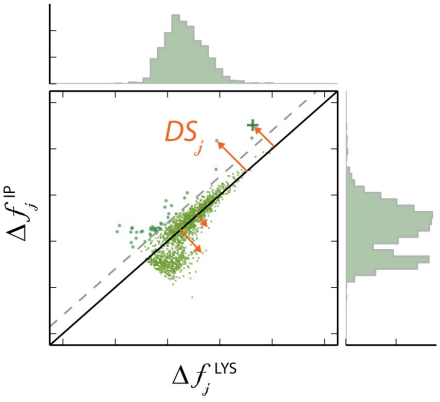
$$f_j = \log_2[\,(\,\text{reads}_j + 1\,) \,/\, \Sigma_k \,\text{reads}_k\,]$$

Normalize to the average frequence of a pool of negative controls:

$$\Delta f_j = f_j - \langle f^{Cntr} \rangle$$

Calculate the Display enrichment Score (DS) for every ORF by normalizing the IP signal to the lysate:

$$DS_j = (\,\Delta f_j^{\,IP} - \Delta f_j^{\,LYS}\,)$$



and a read depth parameter:

$$q = (\,\Delta f_j^{\,IP} + \Delta f_j^{\,LYS}\,) \,/\, 2^{1/2}$$

and define z Score based on the negative functional controls included in the library :

$$\text{Display z Score} = (\,DS_j - \mu_{DS}^{\,cntrl}\,) \,/\, \sigma_{DS}^{\,cntrl}$$

Figure S8. Post-sequencing data analysis pipeline.

See Methods for details.

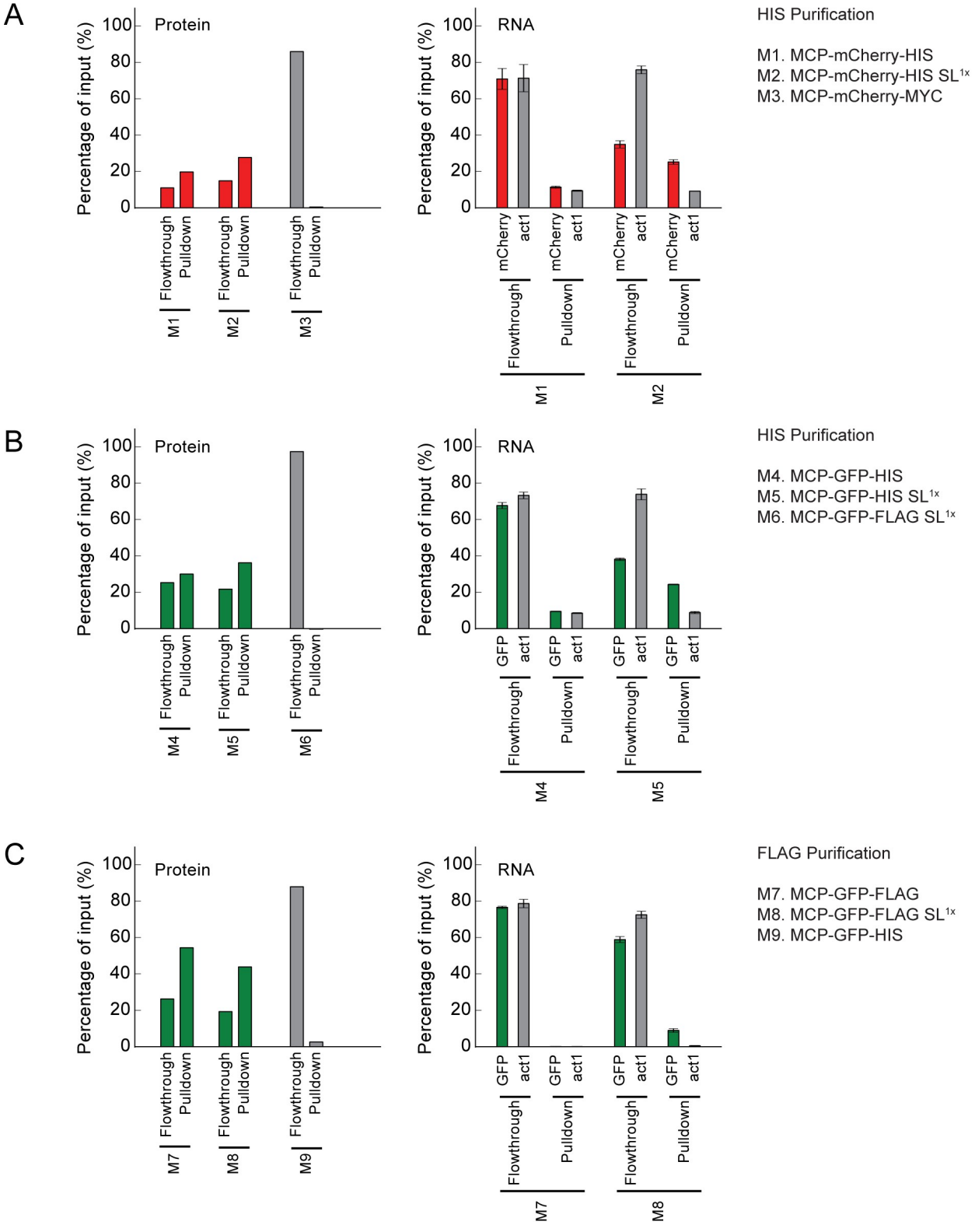

A
Protein
Percentage of input (%)
100
80
60
40
20
0
Flowthrough
Pulldown
M1
M2
M3
RNA
mCherry
act1
Flowthrough
Pulldown
HIS Purification
M1. MCP-mCherry-HIS
M2. MCP-mCherry-HIS SL1x
M3. MCP-mCherry-MYC
B
Protein
RNA
GFP
act1
M4
M5
M6
HIS Purification
M4. MCP-GFP-HIS
M5. MCP-GFP-HIS SL1x
M6. MCP-GFP-FLAG SL1x
C
Protein
RNA
M7
M8
M9
FLAG Purification
M7. MCP-GFP-FLAG
M8. MCP-GFP-FLAG SL1x
M9. MCP-GFP-HIS

Figure S9. *in vivo* mRNA Display proteins co-purify a fraction of their cognate mRNA

Percentage of protein and mRNA in the flow-through and purified fractions with respect to the levels in the input sample. We tested a single step purification for two HIS-tagged (A, B) and one FLAG-tagged (C) construct using reduced salt concentration to avoid excessive loss of RNA and protein. Percentages of protein and RNA levels were calculated with respect to the total input whole cell extract for each strain (using fluorescence and qPCR respectively). Tagged protein constructs were purified in similar amounts with specificity independent of the presence of a stem loop (left-most column). To assess the percentage of cognate mRNA co-purified during the isolation process, we quantified RNA levels of specific and background RNA for constructs with and without stem loops in their 3'UTR. For all three panels, the MCP fusion constructs with no stem-loop co-purified similar levels of construct specific and ACT1 mRNA. In contrast, for the MCP constructs with a stem-loop, purification of the protein resulted in enriched construct specific mRNA with respect to both the reference and the construct with no stem loop. In addition, for the MCP constructs with a stem-loop, construct specific mRNA was depleted in the first flow-through (unbound fraction) with respect to the references.

(A) anti-HIS purification of MCP-mCherry fusion construct with no stem-loop (HIS tag), one stem-loop (HIS tag) and a no HIS tag construct. ~30% of isolated protein co-purified ~25% of mRNA carrying a stem loop. However, ~10% of background RNA with no SL is also co-purified, resulting in an excess ~15% that can be specifically attributed to stem loop binding. Therefore, we estimate that ~30% of isolated protein specifically co-purified ~15% of its cognate mRNA. This percentage of RNA amounts to ~%50 of the RNA that proportionally corresponds to the purified protein.

(B) anti-HIS purification of MCP-GFP fusion constructs with no stem-loop (HIS tag), one stem-loop (HIS tag) and no HIS tag constructs. ~36% of isolated protein co-purified ~25% of mRNA carrying a stem loop. However, ~10% of background RNA with no SL is also co-purified, resulting in an excess ~15% that can be specifically attributed to stem loop binding. Therefore, we estimate that ~36% of isolated protein specifically co-purified ~15% of its cognate mRNA. This percentage of RNA amounts to ~%40 of the RNA that proportionally corresponds to the purified protein.

(C) anti-FLAG purification of MCP-GFP fusion constructs with no hairpin (FLAG tag), one stem-loop (FLAG tag) and no FLAG tag constructs. ~4045% of isolated protein co-purified ~10% of mRNA carrying a stem loop. However, background RNA with no SL is co-purified in small amounts (<0.05%), resulting in a specific co-purification. Therefore, we estimate that ~4045% of isolated protein specifically co-purified ~10% of its cognate mRNA. This percentage of RNA amounts to ~%20 of the RNA that proportionally corresponds to the purified protein.

Overall, we estimate that isolated protein co-purifies roughly 20-50% of its corresponding mRNA with specificity. At the same time, while ~80% of the displayed protein is missing from the flow-through, the construct specific RNA is depleted in an excess of 20-40% compared to both a no SL control and a housekeeping reference gene. The protein and RNA that is not present in the first flow-through will either be purified or removed during the wash steps. Protein levels were assayed by means of fluorescence using a plate reader (Synergy MX, BioTek). Here, protein constructs were bound using the respective magnetic beads and washed with a reduced salt Wash Buffer (150mM NaCl). RNA was precipitated from every sample using TRIzol in order to avoid inconsistent losses on the spin columns used otherwise in this manuscript. RNA levels were

assessed using relative standard curves for each primer set (for mCherry, GFP, ACT1). Percentages of protein and RNA levels were calculated with respect to the total input whole cell extract for each strain.

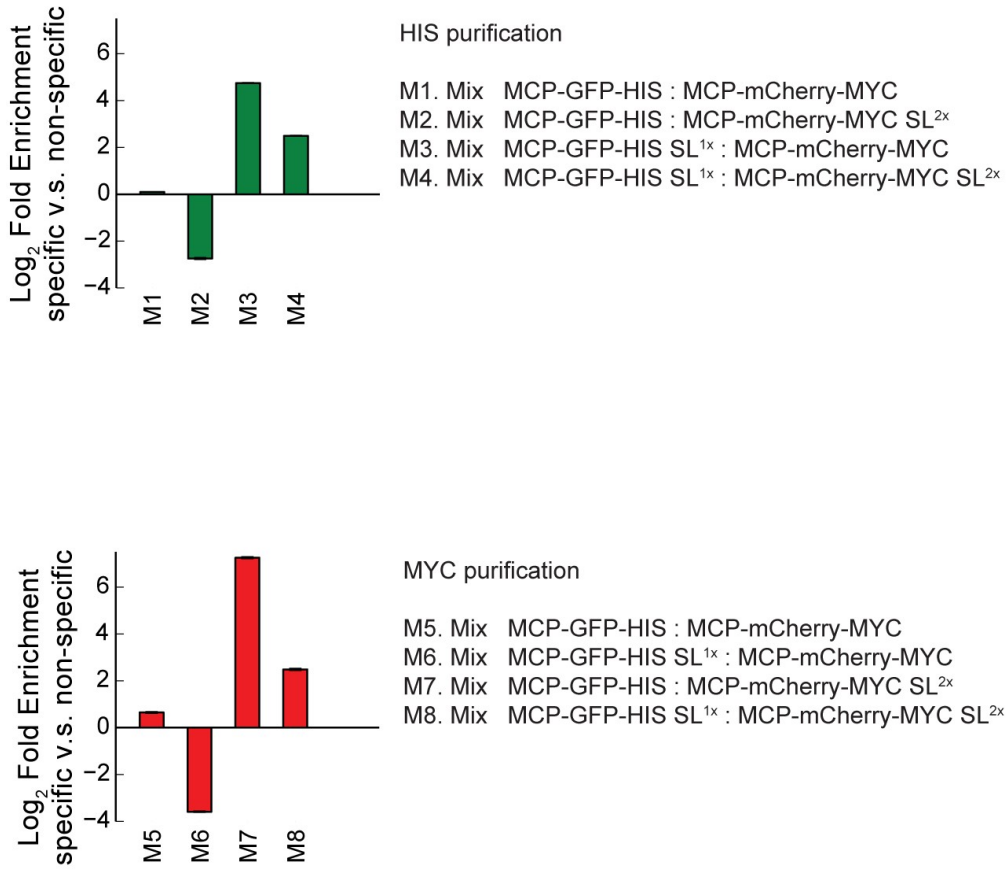


Figure S10. Specificity concerns: Mixed populations with constructs lacking a stem-loop.

Log Fold Enrichment of specific over non-specific mRNA for purifications from mixed populations. Constructs with no hairpin loop are mixed with functional *in vivo* mRNA display constructs. *In vivo* display efficiency is lower in the presence of another functional construct (compare M4 to M3 and M8 to M7). Additionally, when we purify no HL constructs in the presence of functional constructs (M2 and M6), the non-specific mRNA is enriched which is an additional concern. qPCR averages are shown as bars and SD of technical replicates as errors.

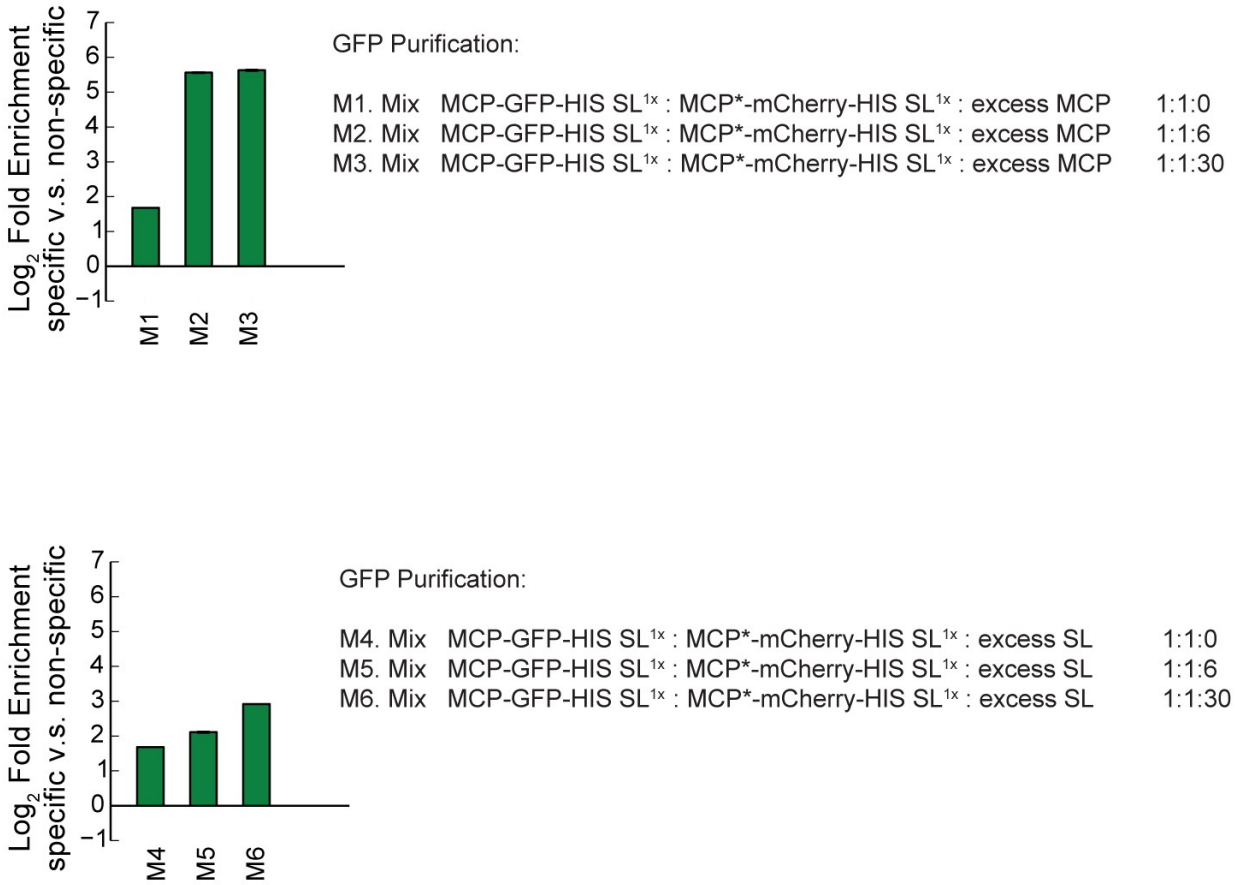


Figure S11. Excess Coat Protein or Stem-loop for increased display enrichment.

In vivo mRNA display GFP constructs were mixed with defective coat protein mCherry constructs and the Log Fold Enrichment of specific over non-specific mRNA was quantified for samples purified with anti-GFP magnetic beads. qPCR averages are shown as bars and SD of technical replicates as errors.

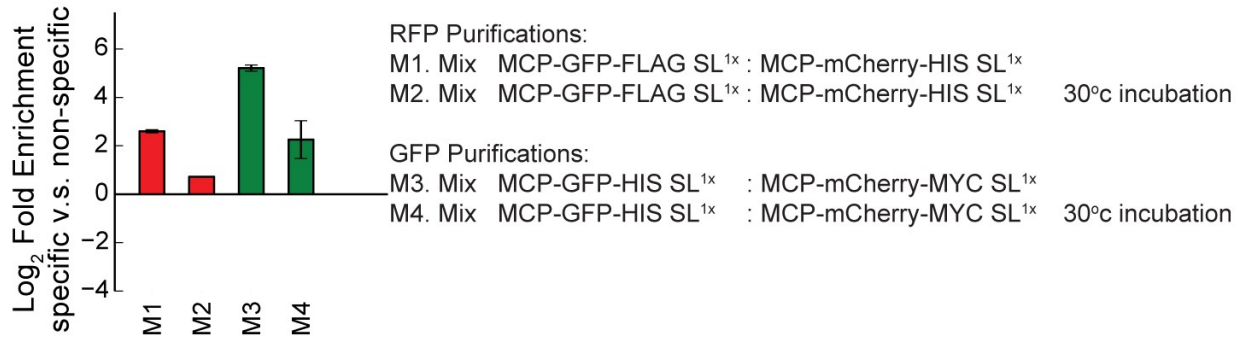


Figure S12. Increasing temperature decreases precision of in vivo mRNA display assay.

In vivo mRNA display GFP and mCherry constructs were mixed together and the Log Fold Enrichment of specific over non-specific mRNA was quantified for samples purified with anti-RFP (M1, M2) and anti-GFP (M3, M4) magnetic beads. For M1 and M3, samples were kept at 4ºC throughout purification. For M2 and M4, samples were incubated at 30ºC for 30 min post lysis. qPCR averages are shown as bars and SD of technical replicates as errors.

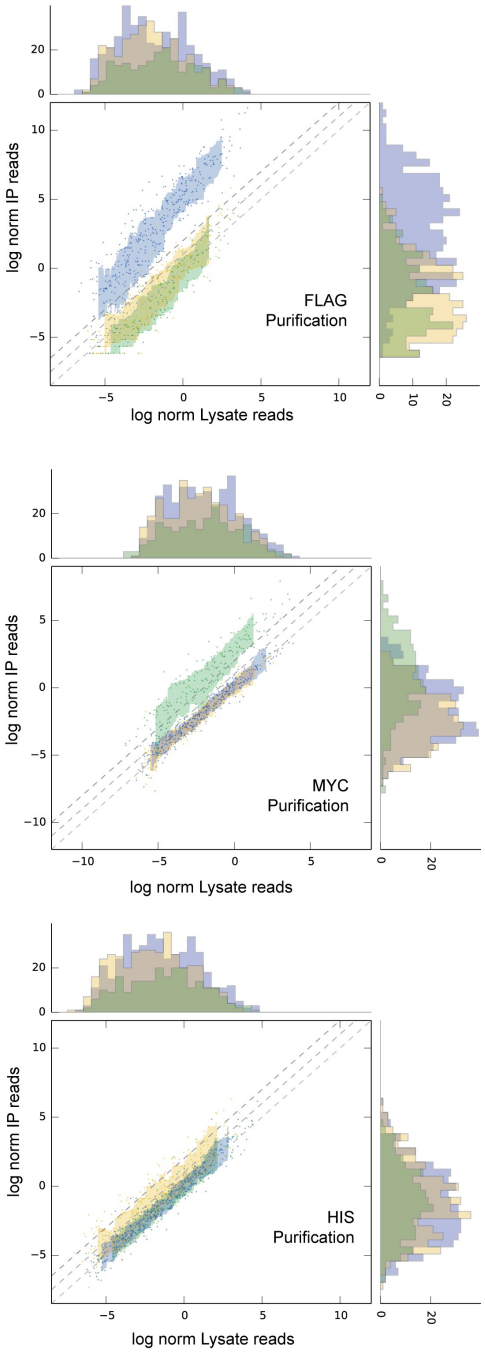


Figure S13. Assessment of in vivo mRNA precision by purification of specific protein subpopulations.

Anti-FLAG (A), anti-MYC (B), anti-HIS (C) immunoprecipitation from a mixed population containing HIS (yellow), MYC (green) and FLAG (blue) tagged yeast in vivo display sub-populations. Scatter plot for log normalized reads for the lysate (x-axis) against the purified samples (y-axis). Reads for each sample were normalized by the mean of non-specific functional

controls. For each population, the area between a rolling 10^{th} and 90^{th} percentile is shaded with the respective color.

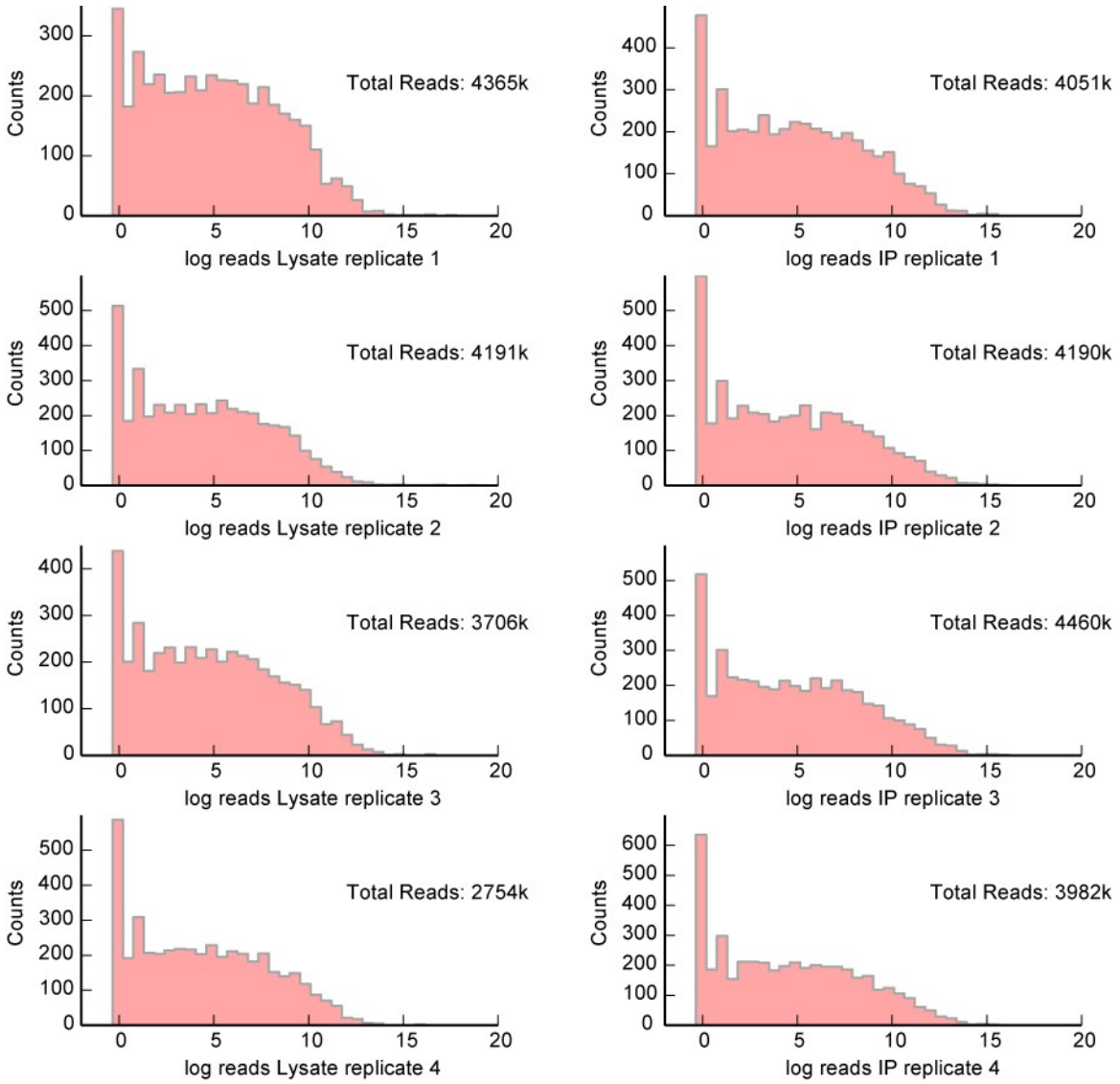


Figure S14. Distribution of reads for *in vivo* mRNA display yeast library purification.

Distribution of $\log_2$(reads+1) for every sample of the yeast library replicates (Fig. 2D-G; Left: Lysates; Right: Purified samples). Total number of reads indicated in thousands for each sample.

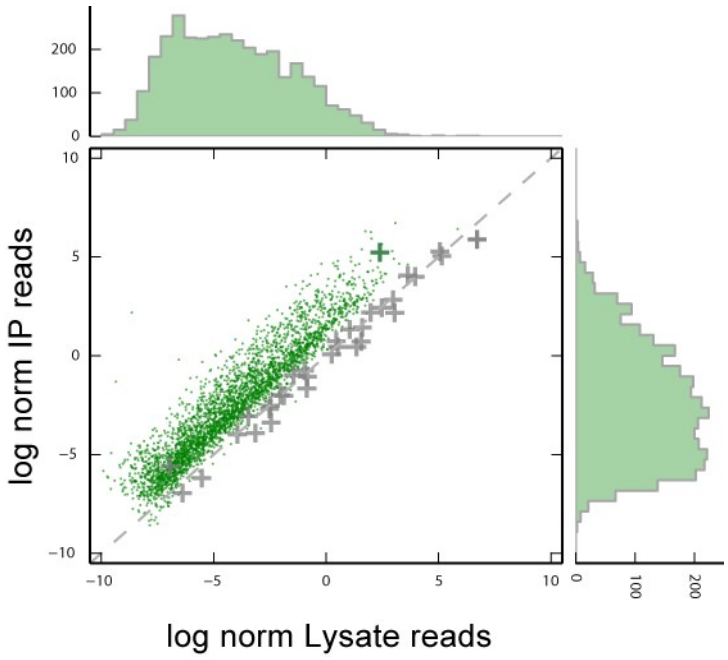


Figure S15. Lysate vs. Purified reads for *in vivo* mRNA display yeast library purification.

Scatter plot for average log normalized reads for the lysate (x-axis) against the purified samples (y-axis) for the purified yeast library (Fig. 2D-G). Reads for each sample were normalized by the mean of non-specific functional controls (grey crosses). GFP is an specific functional positive control (green cross). The area between a rolling 10^{th} and 90^{th} percentile is shaded with the respective color. Purified ORFs are enriched in the purified sample compared to the non-specific functional controls.

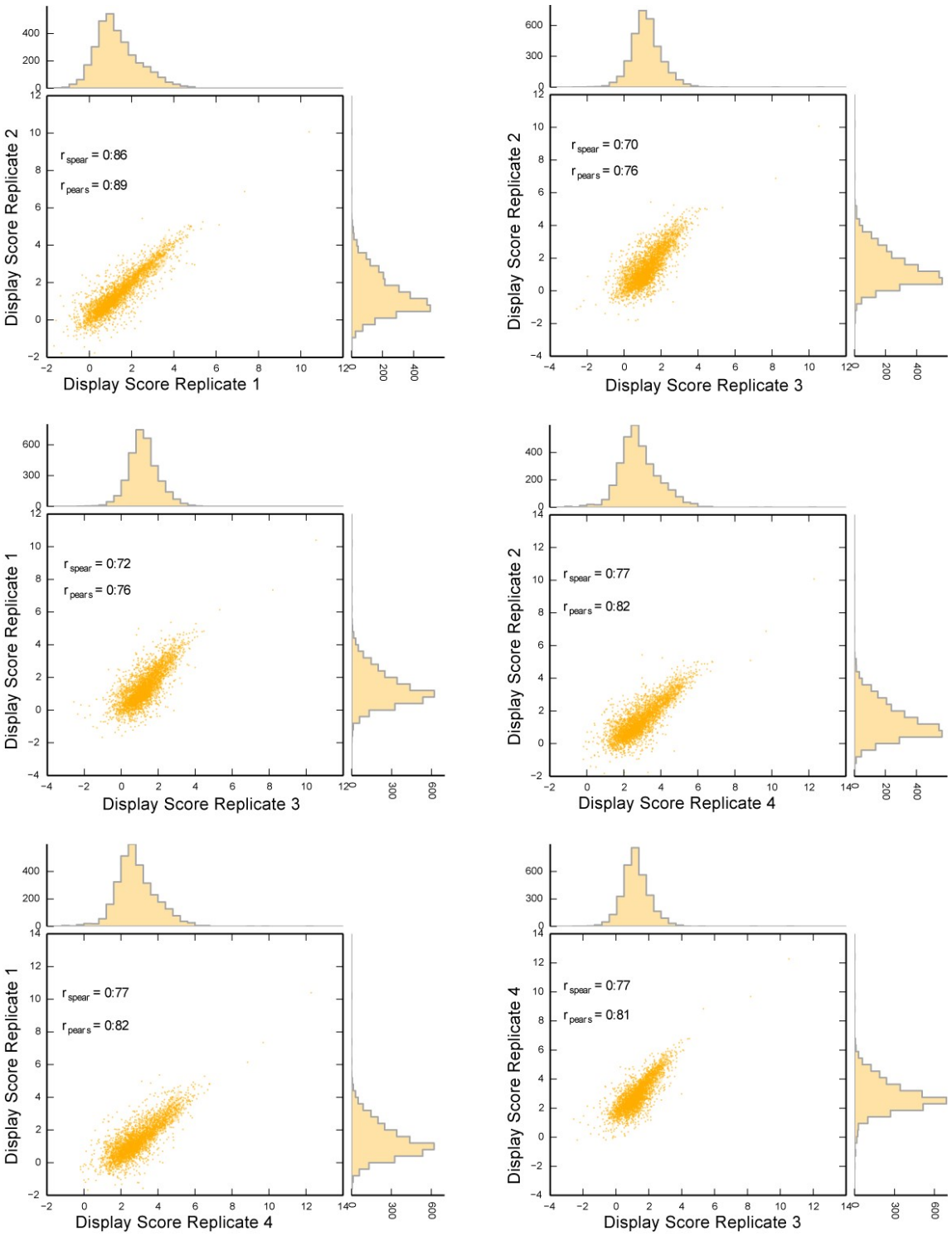


Figure S16. Display Scores for *in vivo* mRNA display yeast library purifications are reproducible.

Scatter plot for Display Scores between all yeast library purification biological replicates (Pearson and Spearman correlations reported).

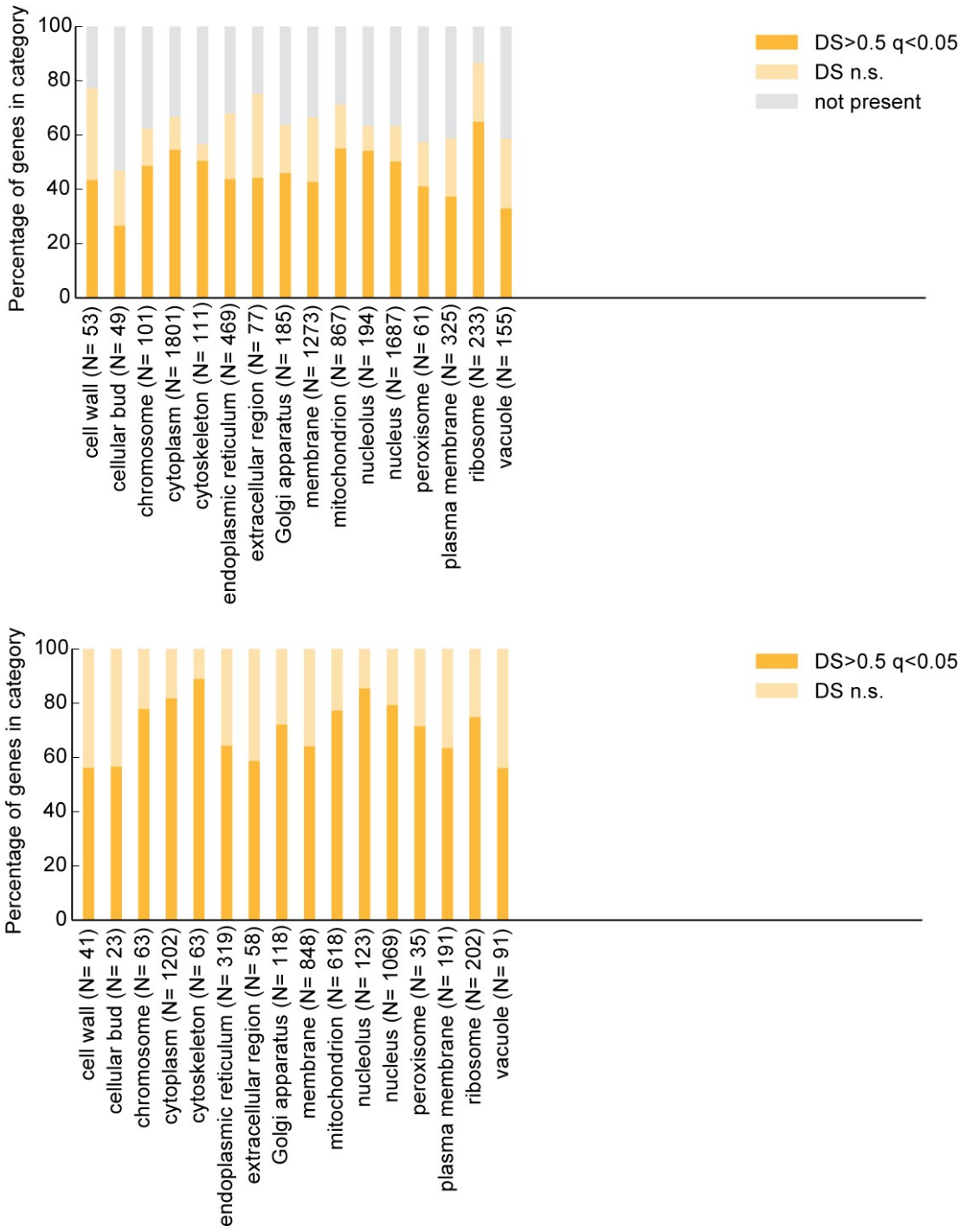


Figure S17. *in vivo* mRNA display yeast library proteins span cellular compartments.

Percentage of *in vivo* mRNA display yeast library proteins with significant Display Scores per GO term compartment category.

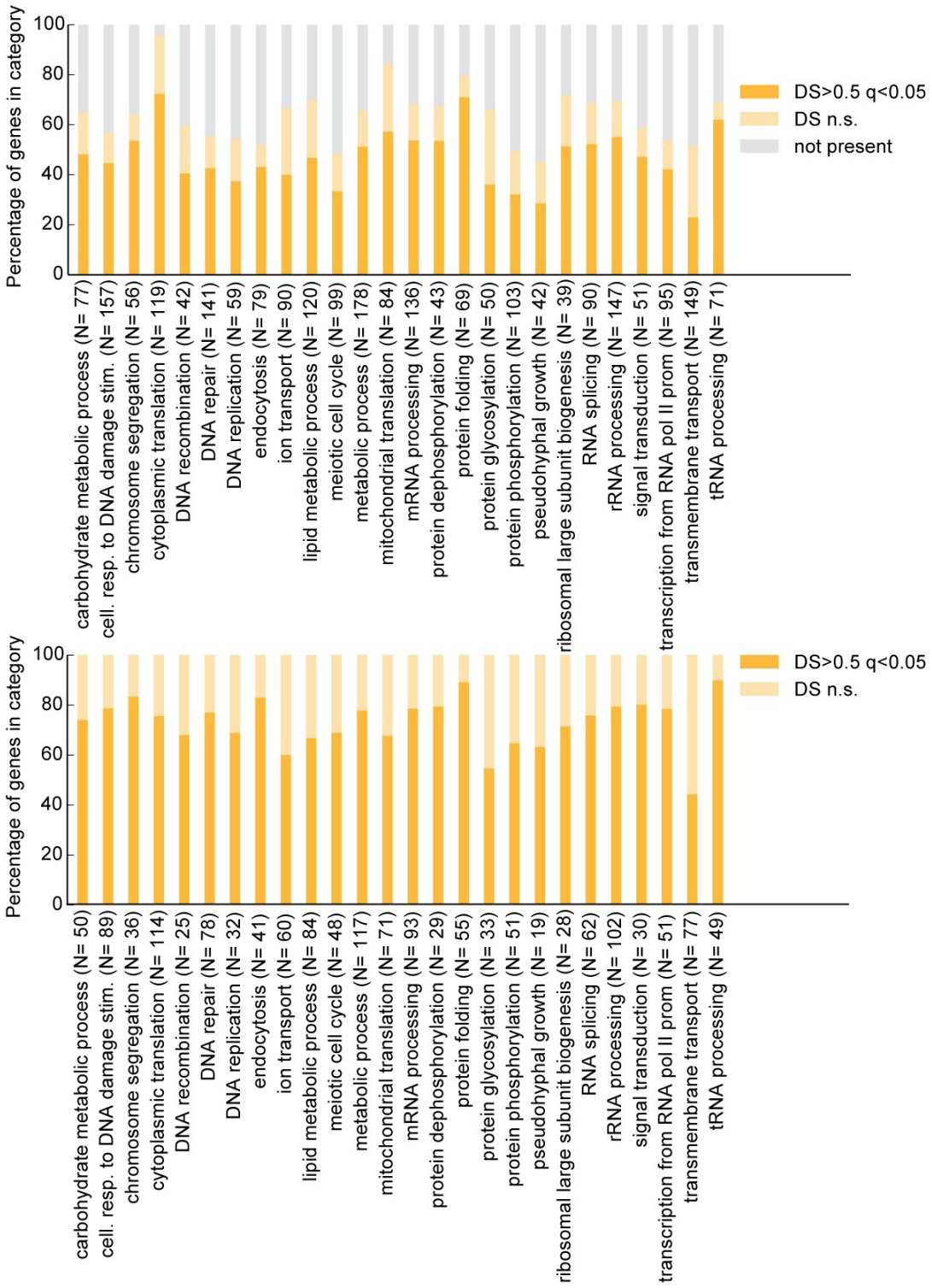


Figure S18. *in vivo* mRNA display yeast library proteins span biological processes.

Percentage of in vivo mRNA display proteins with significant Display Scores per GO term biological process category.

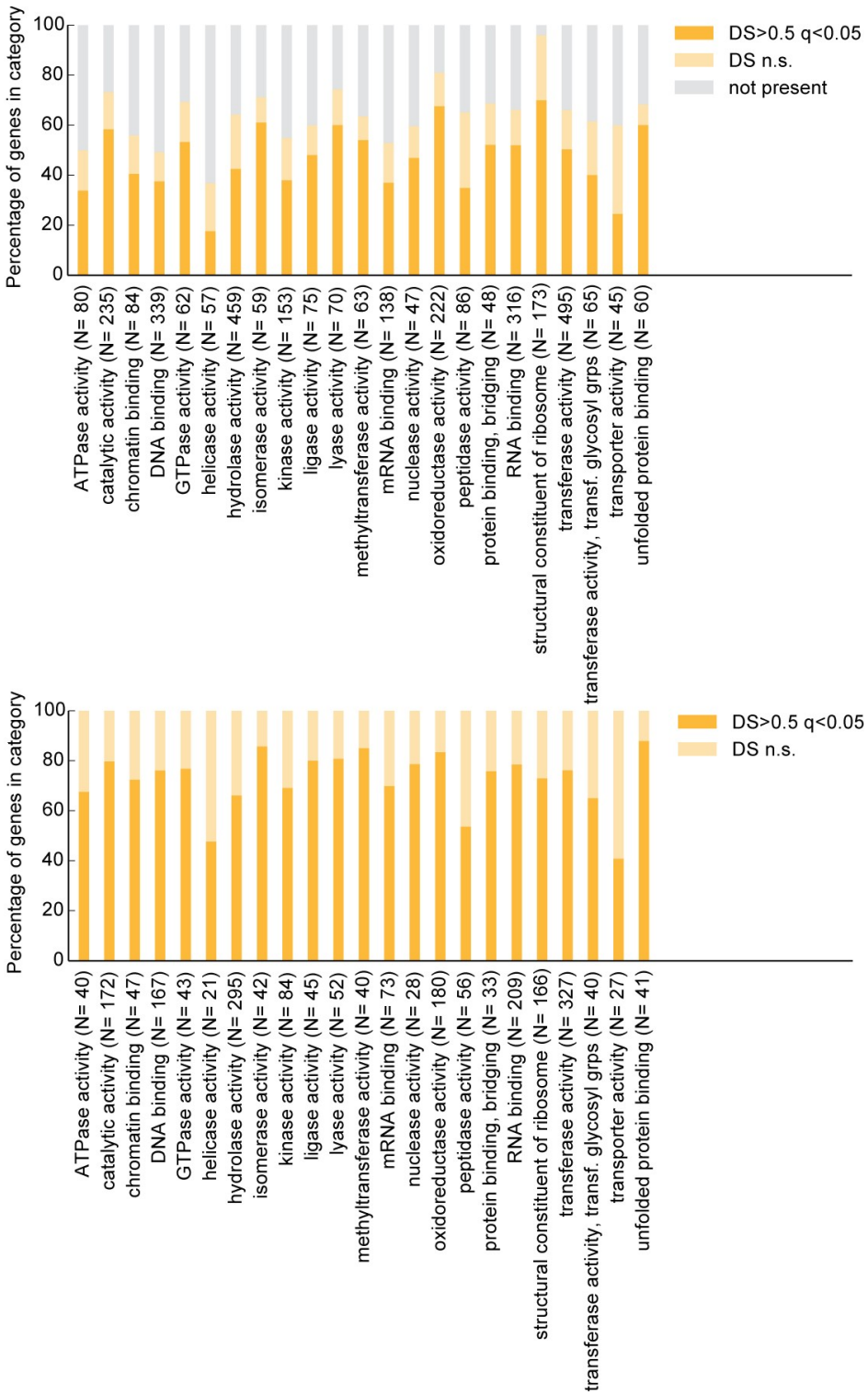


Figure S19. *in vivo* mRNA display yeast library proteins span molecular functions.

Percentage of in vivo mRNA display proteins with significant Display Scores per GO term molecular function category.

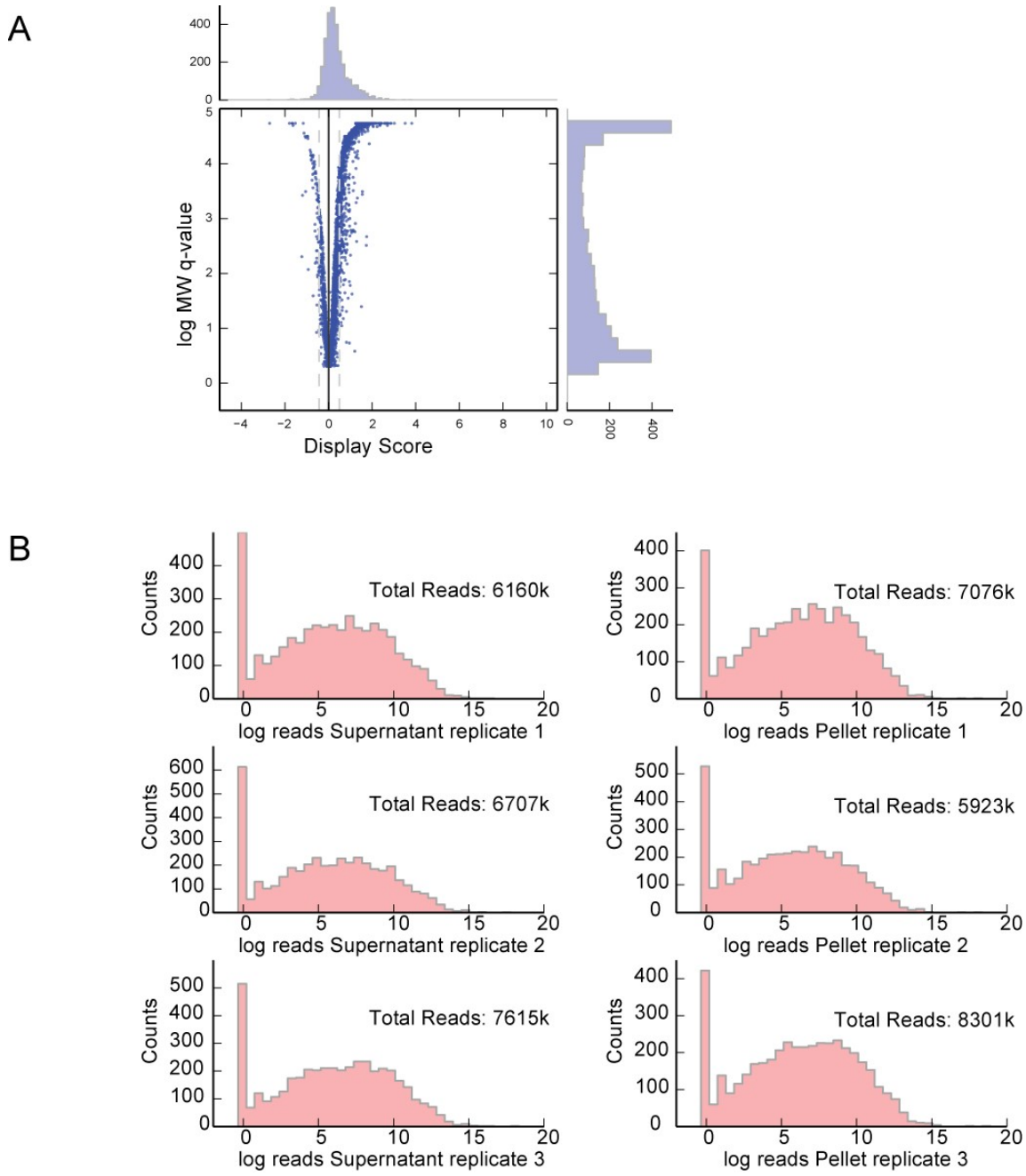


Figure S20. Crude Mitochondrial Isolation Volcano Plot and Read Distributions.

(A) Volcano plot for the crude mitochondrial purification replicates. Average display score (x-axis) against q-values (p-values were calculated with respect to the non-specific functional controls and Benjamini-Hochberg corrected; see Methods). (B) Distribution of $\log_2$(reads+1) for every sample of the crude mitochondrial subfractionation (Left: Supernatant; Right: Crude Mitochondrial Pellets). Total number of reads indicated in thousands for each sample.

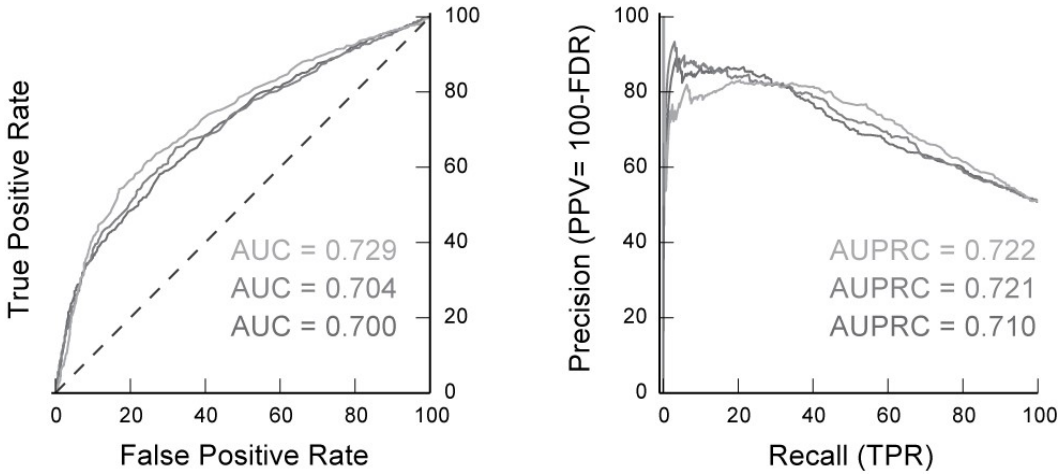


Figure S21. Receiver Operating Characteristic and Precision Recall curves for the crude mitochondrial isolation in Fig. 3.

Members of the library were classified according to their respective Display Score and compared to the GO Term compartment categories. ROC and PR curves for individual replicates are shown.

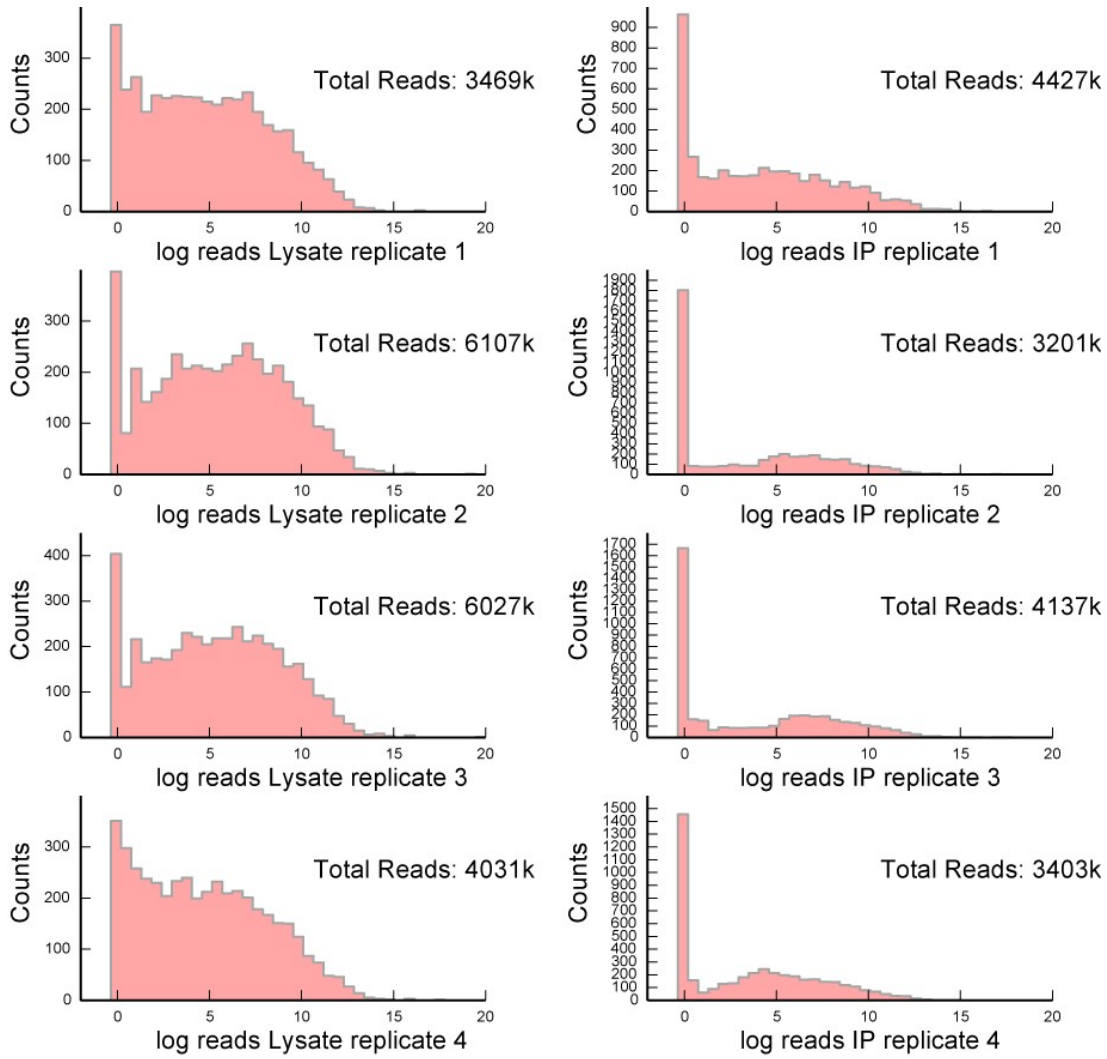


Figure S22. Distribution of reads for *in vivo* mRNA display yeast library SAM2 purification.

Distribution of log (reads+1) for every SAM2 purification in Fig.3C. Total number of reads noted in thousands for each sample.

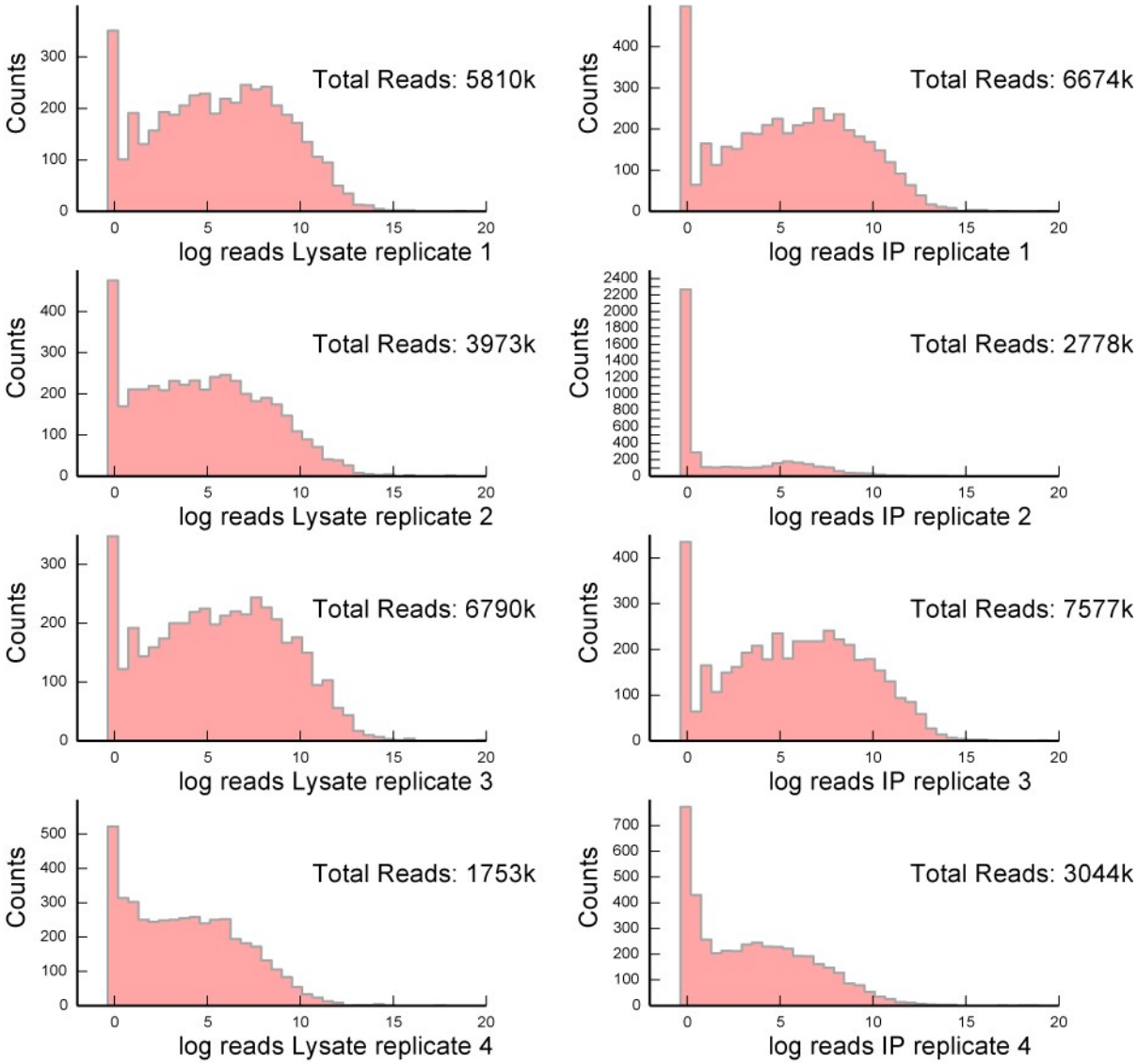


Figure S23. Distribution of reads for *in vivo* mRNA display yeast library ARC40 purification.

Distribution of log (reads+1) for every ARC40 purification in Fig.4B. Total number of reads noted in thousands for each sample.

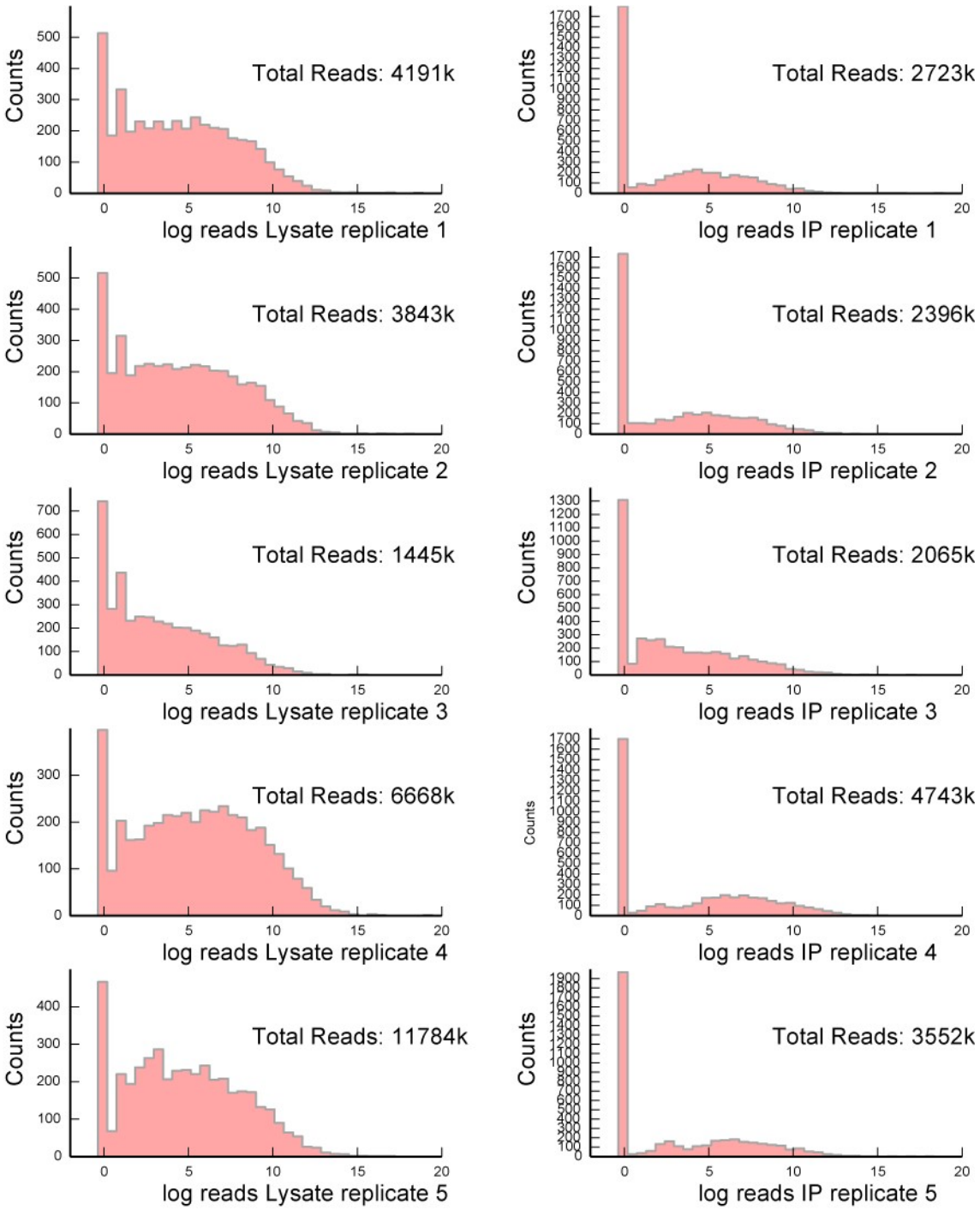


Figure S24. Distribution of reads for *in vivo* mRNA display yeast library negative control purifications.

Distribution of log (reads+1) for every control purification in Fig.4C. Total number of reads noted in thousands for each sample.

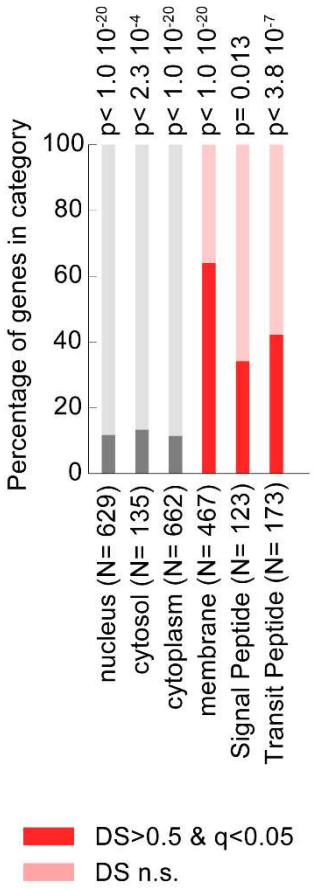


Figure S25. Percentage of *in vivo* mRNA display proteins with significant Display Scores for proteins with signal and transit peptides.

Shown in red are proteins that contain a N-terminal Signal peptide (UniProt annotation), or a Transit peptide (UniProt annotation), or membrane proteins (GO term: 16020). Cytoplasmic and nuclear fractions are reported in grey for reference. Membrane proteins are enriched at an overall higher percentage than proteins carrying peptides responsible for transport, which are usually cleaved from the mature protein and could interfere with the function of the MS2 N-terminal fusion (hypergeometric test for p-values).

1. S. Hocine, P. Raymond, D. Zenklusen, J. A. Chao, R. H. Singer, Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nat. Methods* **10**, 119–121 (2013).

2. J.-D. Pédelacq, S. Cabantous, T. Tran, T. C. Terwilliger, G. S. Waldo, Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79–88 (2006).

3. D. M. Gelperin, *et al.*, Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes Dev.* **19**, 2816–2826 (2005).

4. U. Gueldener, J. Heinisch, G. J. Koehler, D. Voss, J. H. Hegemann, A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. *Nucleic Acids Res.* **30**, e23 (2002).

5. R. Daniel Gietz, R. A. Woods, "Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method" in *Methods in Enzymology*, (Elsevier, 2002), pp. 87–96.

6. W.-K. Huh, *et al.*, Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).

7. R. Elsaesser, J. Paysan, Liquid gel amplification of complex plasmid libraries. *BioTechniques* **37**, 200–202 (2004).

8. M. A. Freeberg, *et al.*, Pervasive and dynamic protein binding sites of the mRNA transcriptome in Saccharomyces cerevisiae. *Genome Biol.* **14**, R13 (2013).

9. E. P. Szymanski, O. Kerscher, Budding yeast protein extraction and purification for the study of function, interactions, and post-translational modifications. *J. Vis. Exp. JoVE*, e50921 (2013).

10. A. K. Hottes, S. Tavazoie, "Microarray-Based Genetic Footprinting Strategy to Identify Strain Improvement Genes after Competitive Selection of Transposon Libraries" in *Strain Engineering*, J. A. Williams, Ed. (Humana Press, 2011), pp. 83–97.

11. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

12. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

13. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).