SI Methods and Materials

Construction of the cDNA library. The library construction started with a mixture of Poly(A)$^+$ mRNAs from human brain, heart, spleen, thymus, and muscle (Stratagene) that were purified by 3 rounds of oligo(dT) as we previously reported (reference 12). A degenerate primer TTN_6 was used as the reverse primer for the synthesis of first-strand cDNA. 5-Me dCTP w□as used for both first and second strands synthesis to protect ORF regions from subsequent restriction digestion. After the generation of blunt ends with T4 DNA polymerase, the directional *EcoR* I/*Hind*III linker (GCTTGAATTCAAGC, Novagen) was ligated to both ends of the dsDNA that allowed the introduction of different left and right arms by restriction followed by ligation. The left arm contains a T7 promoter and a deletion mutant of the TMV 5'-UTR for efficient *in vitro* transcription and translation, respectively. The right arm contains a short sequence for hybridizing with the puromycin-containing oligo linker. Sequences that code for an Avi-tag in the left arm and a FLAG/$His_{\times 6}$ tag in the right arm were also included to facilitate the immobilization and purification of the mRNA-displayed proteome library. The ligated dsDNA was PCR amplified using two primers complementary to the consensus T7 promoter and 3' linker-hybridization region at the 5' and 3' ends, respectively. The PCR product was fractionated to generate a cDNA library with length distribution in the range of 0.5 – 2 kb.

cDNA library thus generated was preselected by one round of mRNA-display to isolate in-frame sequences (reference 1). To remove mRNA secondary structures that may interfere with the selection step, the fusion molecules were converted to DNA/RNA hybrids by reverse transcription. The resulting mRNA-displayed proteins were successively purified, based on the affinity tags at the N- and C-termini. Since only in-frame transcripts have a right C-terminal affinity tag and internally initiated transcripts lack the N-terminal tag, only sequences with both terminal affinity tags, and thus contiguous ORFs, were enriched. This allowed the removal of sequences that contained frame-shifts or untranslated regions. The final cDNA library codes for proteins in the range of 100 to 600 amino acids. Most of the sequences were in the range of 80-200 aa from the ORF (or 140-260 aa with consensus sequences), presumably because the longer

protein fragments were not displayed effectively or that these longer clones were low representation in the library.

Subtraction of abundant sequences from the selected round 3 pool. After three rounds of selection, the selected sequences were PCR amplified and inserted into a TOPO cloning vector for sequence analysis. The selected sequences were diverse and most of them were found averagely twice. Several sequences had much more copies, particularly those originated from SCG3, ANKRD1, C10ORF6, and MYH1. To facilitate the identification of more unique genes based on colony pick-up, these abundant sequences were removed from the selected pool. First, the selected pool was incubated with 50 mM NaOH to digest the mRNA strand on the cDNA/RNA-protein fusion. The resulting first strand cDNAs were recovered and annealed with a mixture of 18 oligonucleotides that were biotinylated at the 5' ends. The mixture was passed through a streptavidin-agarose column to capture the abundant sequences. The flowthrough was used for PCR cloning and sequencing.

Each biotinylated oligonucleotide was complementary to the shortest common region of an abundant sequence. To design these oligos, the shortest common region of the fragments originated from the same gene was mapped (as in Supplemental Figure 3B) and the corresponding nucleotide sequence for its first strand cDNA determined. This sequence was then folded as a single-strand DNA using the mfold program (http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/dna-form1.cgi). Based on the predicted secondary structure, one or two 20-nucleotide regions that were not involved in a stable secondary structure were chosen as the target region(s). The biotinylated oligonucleotide was basically the complementary sequence of this target region. SI Table 5 lists those abundant sequences that were subtracted from the selected round 3 pool using the biotinylated oligonucleotides listed in SI Table 6. Also shown in SI Table 5 is the number of copies found before and after one round of such subtraction procedure. Averagely, the abundance of the subtracted sequences was decreased from 7- to 12-fold.

Construction of Recombinant Genes and Protein Overexpression

Full-length genes of interest were RT-PCR amplified from a cDNA library using gene-specific primers with CACC at the 5' end of the forward primer. The PCR product was gel-purified and subcloned into a pcDNA3.1D/V5-His TOPO expression vector (Invitrogen, Carlsbad, CA). Human HEK 293T cells were maintained in DMEM

supplemented with 10% FBS/2 mM L-glutamine/1,000 units/ml penicillin/1 mg/ml streptomycin. Ten milliliters of cells were transfected with 24 μg of plasmid and 24 μl of Lipofectamine 2000. Approximately 24 h after transfection, cells were lysed and the protein concentrations measured.

1. Shen X, Valencia CA, Szostak J, Dong B, Liu R (2005) *Proc Natl Acad Sci USA* 102:5969-5974.