ONLINE SUPPLEMENTAL INFORMATION

PRESENT ADDRESSES

Jasmin Coulombe-Huntington: Institute for Research in Immunology and Cancer, Université de Montréal, Montreal, QC H3C 3J7, Canada
Shuli Kang: Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA
Ryan R. Murray: Biomedicum Helsinki 1, University of Helsinki, Helsinki 00290, Finland
Bridget E. Begg: Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Andrew MacWilliams: Tecan US, Inc., Morrisville, NC 27560, USA
Quan Zhong: Department of Biological Sciences, Wright State University, Dayton, OH 45435, USA
Shelly A. Trigg: Biological Sciences Department, University of California, San Diego, La Jolla, CA 92093, USA
Stanley Tam: Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA
Lila Ghamsari: Genocea Biosciences, Inc., Cambridge, MA 02140, USA
Maria D. Rodriguez: Biomedical Sciences and Translational Medicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
Kourosh Salehi-Ashtiani: Division of Science and Math, and Center for Genomics and Systems Biology (CGSB), New York University Abu Dhabi, Abu Dhabi, United Arab Emirates
Benoit Charloteaux: Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liege,

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

ORF cloning

Systematic cloning of alternatively spliced ORFs (altORFs) of selected target genes and 454 GS-FLX sequencing to identify unique altORFs was carried out as described previously (Salehi-Ashtiani et al., 2008). Total RNA isolated from heart, liver, brain, testis, and placenta was obtained from Ambion (now Life Technologies). Reverse transcription (RT) was carried out using the SuperScript III kit (Invitrogen) with oligo (dT)16 primers according to the manual. The resultant cDNAs were used as templates for PCR amplification using KOD HotStart Polymerase (Novagen) and ORF-specific primers (**Table S1A**). PCR products were transferred into pDONR223 by Gateway BP reaction (Rual et al., 2004) followed by transformation into *E. coli* DH5α. Transformed *E. coli* cells were plated on LB agar containing spectinomycin for overnight growth at 37˚C, after which up to 12 colonies were isolated for each gene using a Genetix Q-Pix2T Robot.

Sequencing and annotation of isoform clones

The ORF inserts in picked colonies were amplified from *E. coli* lysates with KOD HotStart Polymerase using universal primers (M13G forward and M13G reverse) targeting the ORF-flanking regions in pDONR233 (Rual et al., 2004). The primer sequences are:

M13G forward:
5'-CCCAGTCACGACGTTGTAAAACG
M13G reverse:
5'-GTAACATCAGAGATTTTGAGACAC

The colony-PCR products were arrayed into 12 pools such that a single colony representing a single isoform from the same gene is present in each pool (Salehi-Ashtiani et al., 2008). A 1ml aliquot of pooled PCR products was purified using the MinElute PCR Purification Kit (Qiagen), and DNA concentration was measured via UV-Vis. The purified PCR products were processed using the following kits from Roche Applied Science: GS Standard DNA Library Preparation kit, GS-FLX Standard emPCR kit (Shotgun), GS-FLX PicoTiterPlate Kit (70×75), and GS-FLX Standard LR70 Sequencing Kit.

Raw 454 sequencing data was converted into fastq format using sff_extract (http://bioinf.comav.upv.es/sff_extract/), and fastq-mcf (http://code.google.com/p/ea-utils/wiki/FastqMcf) was used to trim vector sequences and low quality bases. The processed reads were aligned to the hg19 human reference genome (Genome Reference Consortium GRCh37) using a spliced aligner, GMAP (Wu and Watanabe, 2005) (version 2011-03-28), with the parameters "-d hg19 -B 2 -A -t 10 -f samse -H 8 -K 3000000 -L 4000000". The output was saved in SAM format (Li et al., 2009). Only reads that aligned within the genomic regions of the 1,492 target genes were kept.

To assemble the isoform sequences, each position within the locus was annotated as either 'exonic' or 'intronic', determined by the consensus quality scores (CQSs, described below) of aligned nucleotides and gaps in the alignments covering the position. The quality score of a nucleotide in a read was obtained from the fastq files

directly, and the score of a gap was calculated as the average quality score of two flanking nucleotides. The junctions were confirmed by junction-spanning reads. To control for low-quality alignments at the ends of reads, the leftmost and rightmost three nucleotides in each alignment were ignored. For a position covered by nucleotides from *m* forward reads and *n* reverse reads, the quality scores of nucleotides from forward and reverse reads were sorted in descending order, respectively. Then the CQS was calculated using the following formula:

$$CQS = \sum_{i=1}^{m} \frac{x_i}{i} + \sum_{j=1}^{n} \frac{y_j}{j}$$

where x_i is the *i*th quality score in forward reads, and y_j is the *j*th quality score in the reverse reads. The CQSs of the gaps were calculated in the same way. The default annotation of each position was "intronic". A position would be annotated as "exonic" only if the CQS of aligned nucleotides was larger than that of aligned gaps.

We kept for further exon structure annotation only the sequences of fully sequenced clones. If the read depth at any position in an isoform was less than two, the isoform was not considered fully sequenced. Primer information was integrated into the genomic alignment in order to optimize the sequence analysis of the PCR end regions and shorter terminal exons.

At each nucleotide position in the alignment, if there was at least one gap in a spanning read and at least one nucleotide aligned, then the sequence was considered ambiguous. The ambiguity may be caused by either mixed clones or by errors in sequencing, base calling, or alignment. For each ambiguous position, a binomial test was performed with the null hypothesis that the disagreement is due to background errors. To minimize the false positive rate in the identification of mixed clones, we assumed a high background error rate of 0.1 in sequencing, base calling or alignment. If the test result was significant ($P < 0.05$), the observed ambiguity could not be explained by background errors only. When the overall coverage (the total number of both aligned nucleotides and gaps) was low (< 5), the difference was always considered significant. If the results were significant for at least 30 continuous nucleotide positions, we concluded that the ambiguity was due to a mixture of isoforms and the corresponding "clone" was removed from further analysis.

Because the focus of the current work is to study PPIs influenced by splicing events rather than genomic variations such as SNPs or in-frame insertions/deletions, mismatches and short insertions or deletions of less than 30 nucleotides in the read alignments were not considered to be splicing events and were masked in our isoform exon structure annotations. Therefore, multiple clones could be considered the same isoform although the actual sequences at the nucleotide level may be slightly different due to genomic variations. In the case of multiple altORFs encoding the same isoform, only one clone was used for subsequent analysis.

All alternative ORFs with unique exon structures were Sanger-sequenced in both directions. Phred (Ewing et al., 1998) was used to extract sequences from the raw data. Reads with at least 50 nucleotides with non-zero quality scores were aligned using BLAST (bl2seq). Alignments of at least 50 nucleotides with more than 95% identity were integrated into the corresponding contigs of 454 reads by CAP3 (Huang and Madan, 1999) to generate the final consensus sequences.

All isoform structures were compared against the hORFeome and 7 public gene annotation databases: Aceview (2010 release), CCDS (downloaded Sept 2014), Gencode (version 7), hORFeome, MGC (downloaded Sept 2014), RefSeq (downloaded May 2011), and UCSC (downloaded Sept 2014) (Harrow et al., 2012; Karolchik et al., 2014; Pruitt et al., 2014; Pruitt et al., 2009; Temple et al., 2009; Thierry-Mieg and Thierry-Mieg, 2006; Yang et al., 2011). An isoform was considered known if, over its length, it had the exact same junctions as an annotated transcript in any database. Schematic diagrams of isoform exon-intron structures and ORF sequences are available at http://isoform.dfci.harvard.edu/.

RNA abundance

For human brain, heart, liver, and testis, RNA-Seq files from the Illumina Body Map 2.0 project (GEO accession: GSE30611) were downloaded in fastq format. For placenta, RNA-Seq files were downloaded from the NCBI's sequence read archive (SRA) http://www.ncbi.nlm.nih.gov/sra (ERR315336) and converted into fastq format. RefSeq annotated human transcript sequences were downloaded from www.ncbi.nlm.nih.gov/refseq in FASTA format and included 'NM' (protein coding) and 'NR' (non-coding) transcript entries (date January 18th, 2012; 41,899 entries). In the present study RefSeq transcript sequences corresponding to genes for which there were multiple isoform clones (one reference ORF and one or more altORFs) were removed and replaced with the isoform clone sequences (reference ORF and altORF sequences), thus creating a customized transcript FASTA file containing only cloned reference ORFs and altORFs for RNA-Seq Expectation Maximization (RSEM) analysis (Li and Dewey, 2011). For each of the four adult tissues, the "rsem-calculate-expression" command was run using ~80

million 50 bp RNA-Seq reads and the modified RefSeq annotated transcript file for the input. For placenta, ~33 million, 101 bp RNA-Seq reads were used as the input. Isoform-level estimates of transcriptional abundance are reported as transcripts per million (TPM). For each gene, the major isoform was determined by identifying the most abundant isoform (i.e. isoform with the highest TPM) and the corresponding annotation (reference ORF or altORF). The 95% credibility interval (CI) for the TPM value of the major isoform was compared with the TPM values of all other isoforms of that same gene. The putative major isoform was denoted "the major isoform", if the lower bound of the 95% CI was higher than the TPM of all other isoforms, or "the likely major isoform", if there was overlap in the 95% CI with one or more minor isoforms.

Binary interaction mapping and validation

Y2H screening: Haploid *S. cerevisiae* strains Y8930 (*MATα*) and Y8800 (*MATa*) were used for Y2H as described previously (Dreze et al., 2010; Rolland et al., 2014). The altORFs were transferred from entry clones by Gateway LR reaction into pDEST-DB and subsequently introduced into Y8930 (*MATα*) as described previously (Dreze et al., 2010). The Human ORFeome v5.1 (hORFeome) collection used as prey in the Y2H assay against isoform baits was previously transferred into pDEST-AD-CYH2 and introduced into Y8800 (*MATa*) (Rolland et al., 2014).

Before Y2H screening, auto-activation of the *Gal1-HIS3* reporter gene in each DB-X strain was detected by growing the DB-X strains on solid SC media lacking leucine and histidine and containing 1mM 3-amino-1,2,4-triazole (SC-Leu-His + 1mM 3AT) at 30°C for 3 days. During both the screening and pairwise testing steps, latent or *de novo* auto-activation by any DB-X was also identified by growing yeast cells on solid SC-Leu-His + 1mM 3AT + 1μg/ml cycloheximide agar plates as described (Dreze et al., 2010). Any diploid strains showing growth on SC-Leu-His + 1mM 3AT + 1μg/ml cycloheximide were considered to carry DB-X auto-activators and were removed from consideration.

Individual DB-X yeast strains (haploid Y8930 (*MATα)* containing altORFs in the pDEST-DB vector) were screened against ~15,000 ORFs (from 13,000 genes) arrayed in mini-libraries containing 188 individual AD-Y strains (haploid Y8800 *MATa* yeast containing hORFs in the pDEST-AD-CYH2 vector). The DB-X strains were mated with AD-Y strains on YEPD plates overnight at 30°C and then replica plated onto solid SC media lacking leucine, tryptophan, and histidine and containing 1mM 3AT (SC-Leu-Trp-His + 1mM 3AT). Replica plates were incubated at 30°C for 3 days. Yeast colonies were picked into 96-well plates containing SC-Leu-Trp liquid medium and grown for 2 days at 30°C before being spotted onto SC-Leu-Trp agar plates. These plates were incubated for 2 days at 30°C and replica-plated onto (1) SC-Leu-Trp-His + 1mM 3AT to test the activity of the *HIS3* reporter gene and (2) SC-Leu-His + 1μg/ml cycloheximide to identify *de novo* auto-activation of the *HIS3* reporter gene by DB-X alone. Phenotypes were scored after 3 days of growth at 30°C. Colonies were picked into 96-well plates containing SC-Leu-Trp media and cultured at 30°C for 24 to 48 hours. Additional candidate PPIs for the reference ORFs were obtained from the HI-II-14 screen (Rolland et al., 2014).

For colonies growing on selective media plates containing SC-Leu-Trp-His + 1mM 3AT but not on plates containing SC-Leu-His + 1mM 3AT +1μg/ml cycloheximide, the DB-X and AD-Y were amplified using colony-PCR of DB-X and AD-Y followed by stitching-PCR to fuse BD-X and AD-Y through a linker region (Yu et al., 2011). Stitched DB-X and AD-Y PCR products arranged as "bait tail-linker-prey tail" were sequenced using the Roche 454FLX next-generation sequencing technology.

Yeast lysates and PCR were performed as described previously (Yu et al., 2011). Briefly, 5μl of yeast cultures were transferred from SC-Leu-Trp plates to 96-well PCR plates containing 20μl lysis buffer (2.5mg/ml Zymolyase 20T (Seikagaku Corporation) in 0.1M sodium phosphate buffer (pH7.4)). The plates were incubated for 1 to 2 hours at 37°C followed by 5 minutes at 95°C. Yeast lysates were then diluted to 100μl with ddH_2O, from which 2μl was used as a template in a 30μl PCR reaction with HiFi Taq polymerase (Invitrogen). Two sets of PCR reactions were performed to separately amplify the DB-X and AD-Y ORFs. In order to "stitch" the DB-X and AD-Y colony-PCR products through a linker region, the DB primer was paired with a DB vector primer containing a DB-stitching linker tail, and the AD primer was paired with an AD vector primer containing sequence complementary to the DB stitching linker tail.

The primers used in primary colony-PCR were:

DB primer:

5'-GGCTTCAGTGGAGACTGATATGCCTC

DB-Stitching primer:

5'-CTCTCAGCTCGGCGGTATCCCCATCAAACCACTTTGTACAAGAAAGTTGG

AD primer:

5'-CGCGTTTGGAATCACTACAGGG

AD-Stitching primer:

5'-GGATACCGCCGAGCTGAGAGCCATCAAACCACTTTGTACAAGAAAGTTGG

Equal volumes of yeast colony lysate-PCR products of DB-X and AD-Y were mixed and diluted 50 fold from which 2µl was used as a template for stitching PCR with KOD HotStart polymerase (Novagen) using the DB primer and AD primer. The tails of ORF-X and ORF-Y were "stitched" through a linker of 82 bases.

Systematic pairwise testing: To obtain the dataset with the highest possible quality, the growth phenotype of all candidate Y2H interaction pairs from the primary screen was verified by pairwise testing (Dreze et al., 2010; Rual et al., 2005). Pairwise testing was carried out in a matrix format in order to: (1) decompose the gene-level interactions obtained from stitched ISTs into isoform level interactions; (2) systematically test all isoforms of the same gene against all possible interaction partners of any isoform so that interaction profiles of isoforms from the same gene are comparable; and (3) exclude possible growth events on selection media due to physiological adaption or genetic mutation of yeast cells during the screening. Pairwise Y2H tests were performed in triplicate for each gene to generate the complete isoform-interaction partner matrix (all isoforms against all interaction partners of any isoform of that gene), with isoform clones as DBs and hORFeome interaction partners as ADs. Diploid yeast that grew on SC-Leu-Trp-His +1mM 3AT plates but not on SC-Leu-His + 1mM 3AT +1µg/ml cycloheximide plates in at least two out of three colony growth tests were considered positive pairs. Positive pairs were tested a fourth time using the same mating and scoring method, and final positives were picked for colony-PCR followed by Sanger sequencing. Only Sanger sequencing-confirmed pairs were considered to be verified PPIs. Pairs with auto-activation, a no growth phenotype, or that failed sequencing were scored as "NA". Schematic diagrams of isoform PPIs are available at: http://isoform.dfci.harvard.edu.

Detection of protein isoform expression in yeast cells using Western blotting: Yeast cultures of 50ml were grown to mid-log phase and harvested by centrifugation when cultures reached an A_{600} of 1.0. Pellets were frozen, resuspended in 0.3ml RNP lysis buffer (0.1M HEPES pH 7.4, 100mM NaCl, 0.1% NP-40, 0.1mM PMSF, 1mg/ml each of leupeptin, pepstatin, and aproptinin) and lysed by vortexing at 30Hz for 5 minutes in the presence of 450-600µm acid-washed glass beads. Cell debris was pelleted by centrifugation, and cleared lysates collected into a fresh tube. Protein concentrations were determined by Bradford assay (BioRad 500-0006). Gel electrophoresis was performed by running 50µg total protein lysates on a NuPAGE 4-12% Bis-Tris Mini Gel (Life Technologies NP0323) and blotted overnight onto PVDF membrane (Life Technologies LC2005). Isoform Gal4-DB-fusion proteins were detected with anti-Gal4-DB antibody (Santa Cruz Biotechnology, SC-577). Anti-G6PDH antibody (Sigma #A9521) was used as a loading control.

Protein complementation assay (PCA): A subset of positive pairs and negative pairs were selected for validation using PCA (Braun et al., 2009). The corresponding ORFs of the verified protein pairs to be tested were transferred by Gateway LR reaction (Invitrogen) into the pF1N and pF2C vectors, with proteins detected as bait and prey fused to the N-terminus of the F1 and C-terminus of the F2, respectively. After transformation into *E. coli* and selection of transformants in liquid terrific broth medium containing the appropriate antibiotic selection markers, plasmid DNA was extracted and purified using Qiagen 96 Turbo kits (Qiagen) on a BioRobot 8000 (Qiagen**).** The two plasmids carrying F1N-X and Y-F2C (where X = bait and Y = prey) were co-transfected into 293T cells using Lipofectamine 2000 (Invitrogen). 30,000 cells were initially seeded in each well of 96-well plates. Transfection was carried out the next day, and fluorescence of positive cells was detected using flow cytometry analysis on the third day.

The $\log_2$ of the p2 event (cells with YFP fluorescence) over p1 event (total cells gated) was the final raw reporter value for each protein pair. The threshold was set such that any pair scoring above that threshold is considered "positive", and the complement of that set is considered "negative". The recovery rate measured as positive pairs over tested pairs can be viewed as a function of the score threshold.

Calculating Jaccard distances for pairs of isoform interaction profiles

In order to quantify the distinctness of the interaction profiles of pairs of isoforms, we determined the Jaccard distance for any two isoforms encoded by a common gene. The Jaccard distance is defined as the fraction of unshared interaction partners over the union of all interaction partners. We considered only pairs of isoforms where both possess one or more interaction partners and we only considered interactions that were verified as either positive or negative for each of the two isoforms.

Defining isoform-specific regions (ISRs) associated with interaction specific interactions

For each isoform-specific interaction partner, we asked whether the capability of each isoform to interact with a partner correlates with the presence or absence of an ISR. To find isoform-specific regions (ISRs), we searched for

contiguous sequence regions that were present in only one or a subset of isoforms by sliding a ten amino acid window across all isoforms encoded by a gene and determining to which set of isoforms the window matches perfectly. This identified both ISRs, defined as the widest merged window that maps uniquely to one or the same subset of isoforms, and constitutive regions, defined as sequences found in all isoforms of a gene. We filtered our dataset for all isoform-specific interactions where the interacting partner protein interacted with some but not all isoforms of the same gene. For each of these isoform-specific interaction partners, we searched for cases in which the presence of an ISR (of at least 40 amino acid residues in length) in an isoform or subset of isoforms was perfectly correlated, either negatively or positively, with the protein interaction being examined.

For those genes where only two isoforms were interrogated, every isoform-specific interaction must necessarily be correlated with an ISR. Therefore, to assess whether an ISR occurs more frequently than expected by chance, we only examined genes with three or more isoforms (n = 266) where it is possible *a priori* to observe a perfect or imperfect correlation. In 61% of the cases analyzed, we found that isoform-specific interactions could be explained by a single ISR. This is significantly higher than expected by chance (52%) based on a control dataset assembled by randomly shuffling the isoform/partner protein interactions within each gene 10,000 times (1.2–fold, one-sided $P = 6.0 \times 10^{-4}$). This provides the first systematic experimental evidence for a statistically significant link between the presence of isoform-specific sequences and the ability to mediate particular PPIs.

If a region was perfectly positively correlated with the interaction, the region was deemed "interaction promoting". If a region was perfectly negatively correlated with the interaction, the region was deemed "interaction inhibiting". To determine whether isoform-specific interactions were more likely to be associated with a potential promoting or inhibiting region than expected by chance, for cases where a partner interacted with three or more isoforms from the same gene (ISGs), we compared the number of isoform-specific interactions that could be explained by an ISR in our dataset with a randomized control. For the control, the isoforms from the same gene were shuffled 10,000 times, and the number of isoform-specific interactions that was perfectly correlated to an ISR was calculated for each shuffling to create an expected distribution from which the p value was calculated.

Identification of linear-motif binding domains (LMBDs) and linear motifs

We scanned ISRs for linear motifs from the ELM database, excluding matches shorter than 4 amino acid residues or found at very high frequency (>5% of all identified motifs) (Dinkel et al., 2012). When counting motifs in a region, only matches to linear motifs of different linear-motif binding domains (LMBDs) were allowed to overlap. For each interaction partner in our dataset, we determined the linear motif density in the longest ISR associated with that partner. Each distinct ISR was considered only once, regardless of the number of partner associations, and ISRs that were both promoting and inhibiting with different partners were assigned to both categories.

To quantify the enrichment of LMBDs in isoform-specific interaction partners, Pfam-A domains (Finn et al., 2014) were mapped to all interaction partners using HMMER 3.0 (e-value=10^{-2}) (Finn et al., 2011), and each partner was classified as either containing an LMBD, as annotated in the ELM (Dinkel et al., 2012) or DILIMOT (Neduva and Russell, 2006) databases, or not. Interaction partners were then assigned either as exhibiting isoform-specific interactions associated with a promoting ISR or not.

Identification of splice-mediated disruption of potential domain-domain interactions (DDIs)

Pfam-A domains (Finn et al., 2014) were mapped to all isoforms and interaction partners using Hmmer 3.0 (e-value = 10^{-5}) (Finn et al., 2011), and isoform-partner pairs encoding a predicted DDI from iPfam (Finn et al., 2005), 3Did (Mosca et al., 2014), or Domine (Yellaboina et al., 2011) were identified. We searched our dataset for isoform-partner pairs containing a predicted DDI and, where possible, determined how often that interaction was lost upon disruption of the domain in another isoform of the same gene. As a control, we started with the same isoform-partner pairs and determined how often an interaction was lost when another isoform of the same gene was shorter by at least 50 amino acids, thus controlling for the observation that shorter isoforms tended to participate in fewer interactions in this dataset.

Structural analysis of isoform-specific interactions

To obtain structural information for the unveiled interactions, we mapped Entrez Gene IDs to Uniprot accession numbers using the Uniprot ID mapping tool. Protein-protein interactions were submitted to Interactome3D (May 2015) (Mosca et al., 2013). When more than one structure was provided, we selected that with a 'rank major' of 1, i.e. we maximized the sequence coverage and prioritized experimental structures.

To map unique interactions between proteins of two genes (i.e. without considering different isoforms) onto three-dimensional structures, we defined the interaction interface as the set of residues that had a heavy atom at a distance < 6 Å from the binding partner for each binary complex. To map isoform sequences onto structures, we

performed a local pairwise alignment between the structure sequence and the corresponding isoform and identified the interface residues.

Interactome network analysis of isoform interaction partners.
To compare the features of two isoforms from the same gene, we did pairwise comparison of isoform interaction partners for their network distance, co-expression, and disease subnetwork association. Starting with all partner proteins interacting with one or more isoforms, we identified pairs of partners belonging to the following three groups (**Figure 5A**): i) "single protein", in which the two partner proteins interact with the same protein isoform; ii) "alternative isoforms", in which each partner protein of the pair interacts with one or more isoforms of the same gene with which the other protein does not; and iii) "products of different genes", in which two partner proteins do not interact with any isoforms encoded by the same gene (control). Some pairs of proteins interacting with multiple isoforms from one or more genes may appear in both categories (i) and (ii).

The mean shortest path distance in HI-II-14 (Rolland et al., 2014) between any two proteins that interact with the same single protein, interact with alternative isoforms, or interact with proteins encoded by separate genes was calculated. Paths traversing proteins derived from the same gene as the isoforms were ignored; protein pairs with no connecting path were also ignored.

Using all 75-base-pair runs from the Illumina Body Map 2.0 16-tissue RNA-Seq dataset (Illumina BodyMap 2.0), which we retrieved from the NCBI's Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra, study: ERP000546, runs:ERR030888-ERR030903), and the Bowtie alignment tool (Langmead and Salzberg, 2012) with default settings, we mapped reads to all hORFeome clone sequences and calculated the $\log_2$ read count for each gene for each tissue. We then normalized expression values for each gene to that of the upper-quartile most highly expressed gene for each tissue, as described by Bullard and colleagues (Bullard et al., 2010), and calculated the Pearson correlation on all pairs of interaction partners after filtering out genes with a maximal expression below $1/32^{nd}$ of the upper-quartile gene expression, -5 in normalized $\log_2$ space. The fraction of pairs co-expressed (i.e. having a positive Pearson correlation coefficient greater than 0.15) was calculated for each of the three groups of pairwise proteins described above.

Disease subnetworks were created for each disease by mapping the set of disease associated genes from GeneCards (Safran et al., 2010) onto an independently-mapped human interactome (Rolland et al., 2014) and retrieving the disease genes and their first degree PPI neighbors. Genes having an isoform screened in this study were omitted from the subnetworks. The mean of the Jaccard index of disease subnetwork co-occurrence for all protein pairs within each class was then calculated.

Tissue-specificity of isoform interaction partners.
To estimate the fraction of tissue-specific of interaction partners, we measured the range of normalized $\log_2$ expression levels in the Illumina Body Map 2.0 16-tissue RNA-Seq dataset (Illumina BodyMap 2.0) and considered genes with a range greater than 7 as tissue-specific. Using the range of expression levels ensures the analysis is sensitive to differences in a single tissue. We compared the tissue-specificity of partners affected by change-over interaction differences to other interaction partners.

Yeast-based functional complementation assays.
To further investigate the functionality of the isoforms with different protein interaction profiles, we exploited the yeast-based cross-species complementation assays to measure their ability to rescue phenotypic defects of a loss-of-function mutation in a cognate yeast gene. Among the 138 genes for which the isoforms have different protein interaction profiles, 8 genes showed yeast/human complementation relationships in a recent study (Kachroo et al., 2015). The reference ORFs, altORFs, and a GFP ORF were transferred into pHYCDest-LEU2 (CEN/ARS-based, ADH1 promoter, and LEU2 marker) by Gateway LR reactions followed by transformation into NEB5α competent *E. coli* cells (New England Biolabs) and selection for ampicillin resistance. After confirmation of ORF identity by Sanger sequencing, plasmids expressing the reference ORFs, altORFs, and GFP were further transformed into the corresponding yeast temperature sensitive (TS) mutants. For yeast TS mutants transformed with expression vectors, cells were grown to saturation in 96-well cell culture plates at room temperature. Each culture was then adjusted to an OD600 of 1.0 and serially diluted to 5^{-1}, 5^{-2}, 5^{-3}, 5^{-4}, and 5^{-5}. These cultures (5µl of each) were then spotted on SC–LEU plates as appropriate to maintain the plasmid and incubated at either 24ºC, 36ºC, or 38ºC. Plates were imaged after two or three days depending on the growth. Results were interpreted by comparing the growth difference between the yeast strains expressing different protein isoforms and the corresponding control strain expressing the GFP gene. Two independent cultures were grown and assayed for each strain.

Analysis recapitulated with known isoforms and non-NMD isoforms

To control for the possibility that some of our cloned altORFs may not encode stable protein isoforms, we repeated the bioinformatic analyses related to enrichment of linear motifs, domain-domain disruptions, isoform partner network properties, and isoform partner tissue specificity with the following subsets of the isoforms: (1) isoforms for which all splice-sites are represented in at least one of seven public gene annotation databases (Aceview, CCDS, Gencode, hORFeome, MGC, RefSeq, and UCSC), labeled "with known splice sites", (2) isoforms which are represented in their full length in at least one of the seven databases, labeled "with known full length", and (3) isoforms which are predicted not to undergo nonsense-mediated decay, labeled "with no predicted NMD targets". Isoforms with a premature stop codon more than 55 nucleotides away from the last splicing junction are considered NMD targets.

Disordered regions in ISRs.

We have applied the VSL2 disorder predictor (Peng et al., 2006) to all four categories of isoform pair PPI profiles from **Figure 6A**. First, the disorder predictions were run on all full-length isoforms from these datasets. Second, the disordered fragments within isoform-specific regions (ISRs) of each isoform pair were analyzed using various lengths cutoffs. After filtering extremely short ISRs (<10 aa), the longest consecutive disordered region (VSL2 score ≥ 0.5) in the ISRs of each isoform pair has been identified. Finally, the percentage of isoform pairs with disordered ISRs longer than certain length threshold was plotted for each type of isoform pair.

SUPPLEMENTAL REFERENCES

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175-185.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.* (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222-230.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29-37.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., *et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760-1774.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. *9*, 868-877.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., *et al.* (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. *42*, D764-770.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Mosca, R., Ceol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nat. Methods *10*, 47-53.

Mosca, R., Ceol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. *42*, D374-379.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., *et al.* (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756-763.

Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J., *et al.* (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. *19*, 1316-1323.

Thierry-Mieg, D., and Thierry-Mieg, J. (2006). AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol *7 Suppl 1*, S12 11-14.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859-1875.

Yang, X., Boehm, J.S., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T., Green, T.M., *et al.* (2011). A public genome-scale lentiviral expression library of human ORFs. Nat. Methods *8*, 659-661.

Yellaboina, S., Tasneem, A., Zaykin, D.V., Raghavachari, B., and Jothi, R. (2011). DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Res. *39*, D730-735.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X., *et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat. Methods *8*, 478-480.