Supplementary Materials

Non-canonical open reading frames encode functional proteins essential for cancer cell survival

John R. Prensner[1,2,3], Oana M. Enache[1], Victor Luria[4], Karsten Krug[1], Karl R. Clauser[1], Joshua M. Dempster[1], Amir Karger[5], Li Wang[1], Karolina Stumbraite[1], Vickie M. Wang[1], Ginevra Botta[1], Nicholas J. Lyons[1], Amy Goodale[1], Zohra Kalani[1], Briana Fritchman[1], Adam Brown[1], Douglas Alan[1], Thomas Green[1], Xiaoping Yang[1], Jacob D. Jaffe[1,8], Jennifer A. Roth[1], Federica Piccioni[1,9], Marc W. Kirschner[4], Zhe Ji[6,7], David E. Root[1], Todd R. Golub[1,2,3*]

Author affiliations:
[1]Broad Institute of Harvard and MIT, Cambridge, MA, 02142, USA.
[2]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02215
[3]Division of Pediatric Hematology/Oncology, Boston Children's Hospital, Boston, MA, 02115
[4]Department of Systems Biology, Harvard Medical School, Boston, MA, 02115, USA
[5]IT-Research Computing, Harvard Medical School, Boston, MA, USA, 02115
[6]Department of Pharmacology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611
[7]Department of Biomedical Engineering, McCormick School of Engineering, Northwestern University, Evanston, IL 60628
*Corresponding author

[8]Present address: Inzen Therapeutics, Cambridge, MA, 02139, USA
[9]Present address: Merck Research Laboratories, Boston, MA, 02115, USA

Key words: Cancer, gene dependency, translatome, CRISPR, unannotated genes

Address correspondence to:
Todd R. Golub, MD

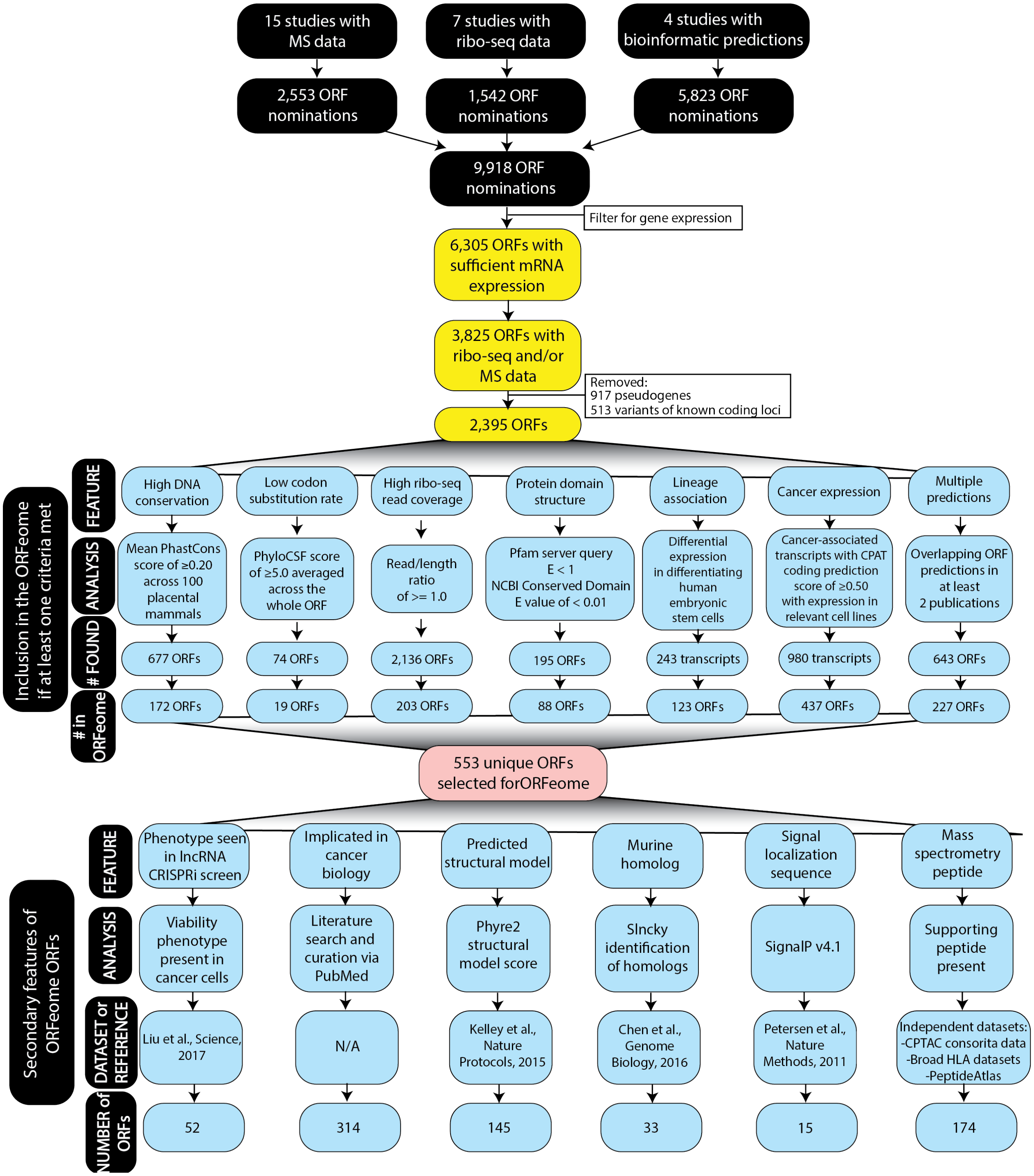
Chief Scientific Officer
Broad Institute of Harvard and MIT

Room 4013
415 Main Street
Cambridge, MA, 02142
Email: golub@broadinstitute.org
Phone: 617-714-7050

Table of contents

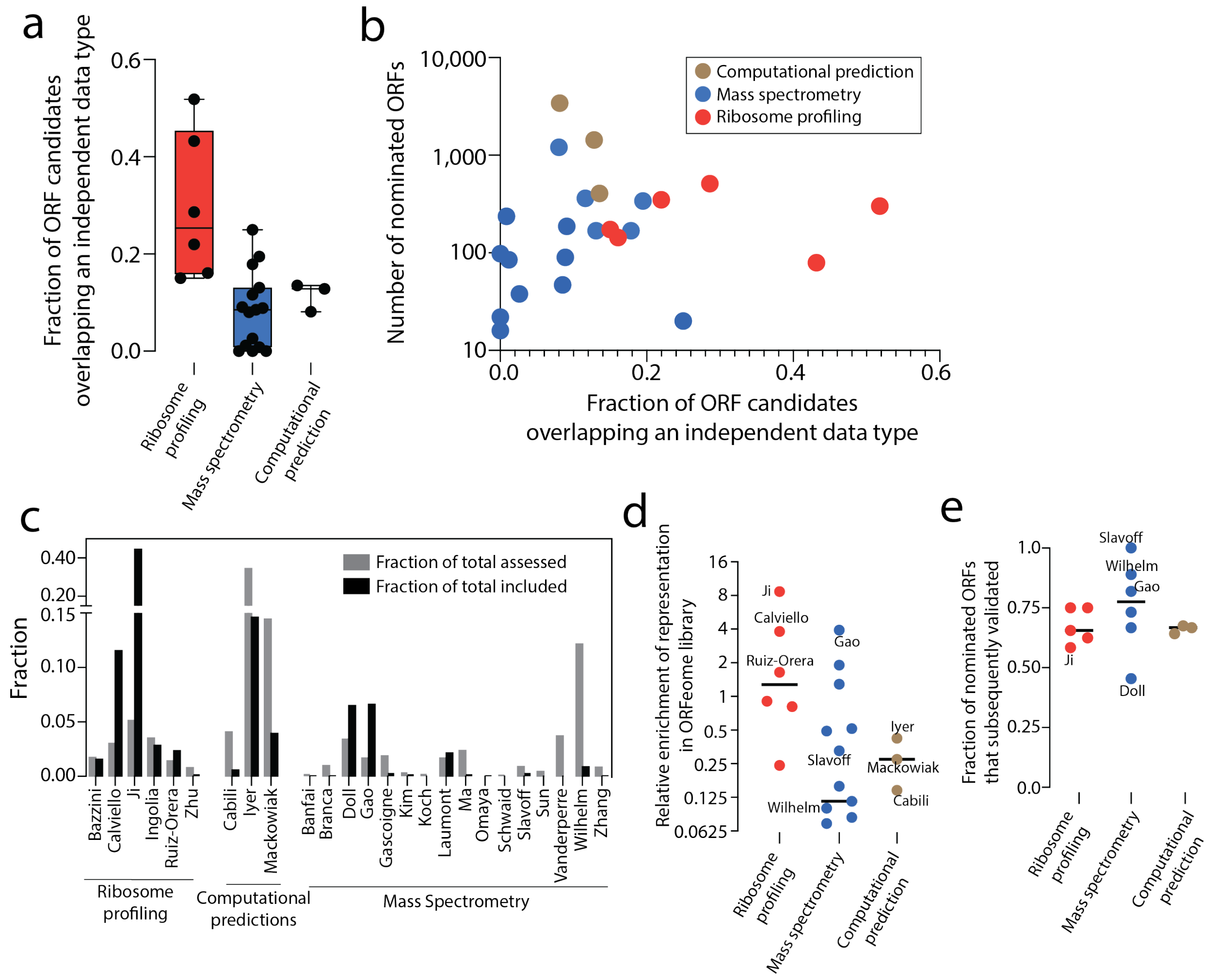
Supplementary figures

Supplementary discussion

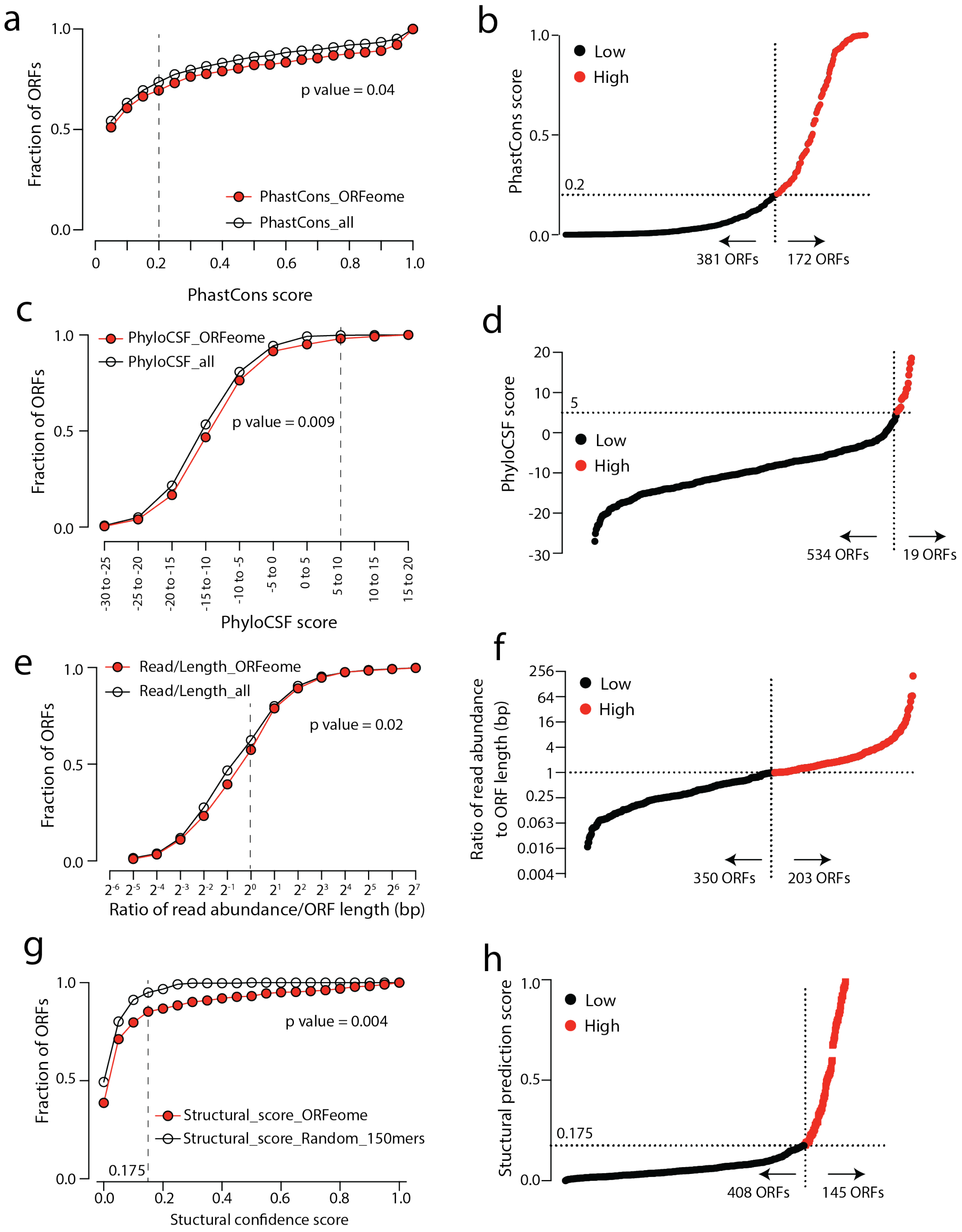

15 studies with MS data
7 studies with ribo-seq data
4 studies with bioinformatic predictions
2,553 ORF nominations
1,542 ORF nominations
5,823 ORF nominations
9,918 ORF nominations
Filter for gene expression
6,305 ORFs with sufficient mRNA expression
3,825 ORFs with ribo-seq and/or MS data
Removed:
917 pseudogenes
513 variants of known coding loci
2,395 ORFs
Inclusion in the ORFeome if at least one criteria met
FEATURE
ANALYSIS
# FOUND
# in ORFeome
High DNA conservation
Low codon substitution rate
High ribo-seq read coverage
Protein domain structure
Lineage association
Cancer expression
Multiple predictions
Mean PhastCons score of ≥0.20 across 100 placental mammals
PhyloCSF score of ≥5.0 averaged across the whole ORF
Read/length ratio of >= 1.0
Pfam server query E < 1
NCBI Conserved Domain E value of < 0.01
Differential expression in differentiating human embryonic stem cells
Cancer-associated transcripts with CPAT coding prediction score of ≥0.50 with expression in relevant cell lines
Overlapping ORF predictions in at least 2 publications
677 ORFs
74 ORFs
2,136 ORFs
195 ORFs
243 transcripts
980 transcripts
643 ORFs
172 ORFs
19 ORFs
203 ORFs
88 ORFs
123 ORFs
437 ORFs
227 ORFs
553 unique ORFs selected forORFeome
Secondary features of ORFeome ORFs
FEATURE
ANALYSIS
DATASET or REFERENCE
NUMBER of ORFs
Phenotype seen in lncRNA CRISPRi screen
Implicated in cancer biology
Predicted structural model
Murine homolog
Signal localization sequence
Mass spectrometry peptide
Viability phenotype present in cancer cells
Literature search and curation via PubMed
Phyre2 structural model score
Slncky identification of homologs
SignalP v4.1
Supporting peptide present
Liu et al., Science, 2017
N/A
Kelley et al., Nature Protocols, 2015
Chen et al., Genome Biology, 2016
Petersen et al., Nature Methods, 2011
Independent datasets:
-CPTAC consorita data
-Broad HLA datasets
-PeptideAtlas
52
314
145
33
15
174

Supplementary Figure 1: A flowchart for ORF selection.

Manual curation of ~9900 ORF loci from the indicated dataset sources were then filtered using the indicated biological attributes and selection criteria. After selection, the 553 ORFs were then evaluated by additional metrics as shown. Please see the Methods for additional details on selection criteria.

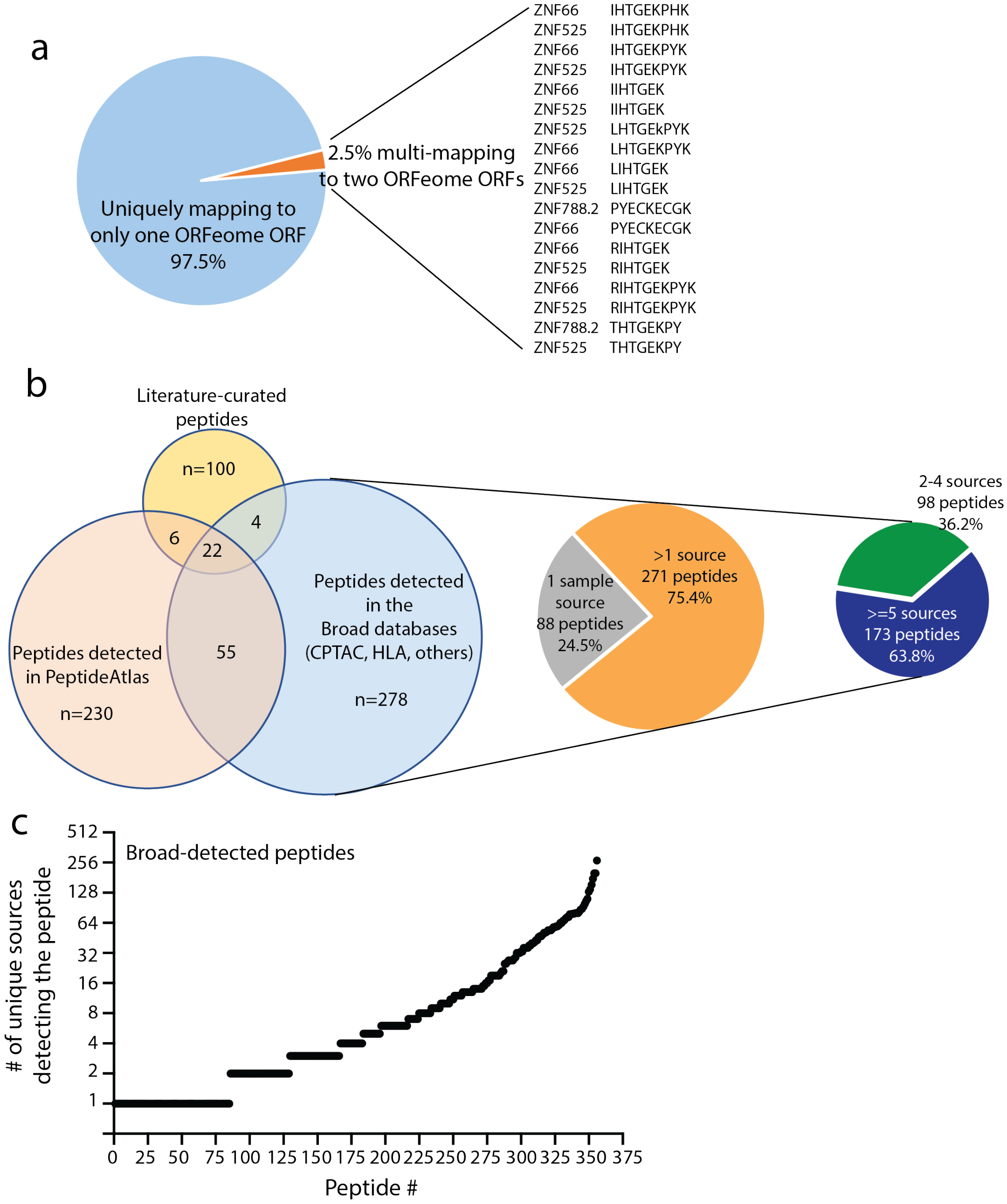


Supplementary Figure 2: Overlap and representation of datasets in the ORFeome library.
a) The fraction of ORF nominations in a given dataset that are also represented in an independent dataset. Each dot represents a literature source for data. Box plots represent median and interquartile ranges with whiskers indicating minimum and maximum values. **b)** A scatter plot showing the number of nominated ORFs in a given study compared to the fraction of ORF candidates overlapping an independent dataset. Datasets are color-coded as indicated. **c)** A barplot showing the fraction of ORFs in the total 9,918 set that are represented in each dataset, as well as the fraction of the 553 ORFeome candidates represented in each dataset. **d)** The relative enrichment of representation in the ORFeome library for each dataset. Each dot represents a dataset and the line shows the median value in the indicated group. Enrichment is calculated as (ORF_fraction_of_total_assessed / ORF_fraction_of_total_included) for each dataset. **e)** For each dataset, the fraction of tested ORFs that subsequently validated by either V5-tag cDNA translation or independent peptide identification in a unique mass spectrometry dataset. The line indicates the median for each group. Only datasets contributing >5 candidates to the ORFeome library are included.

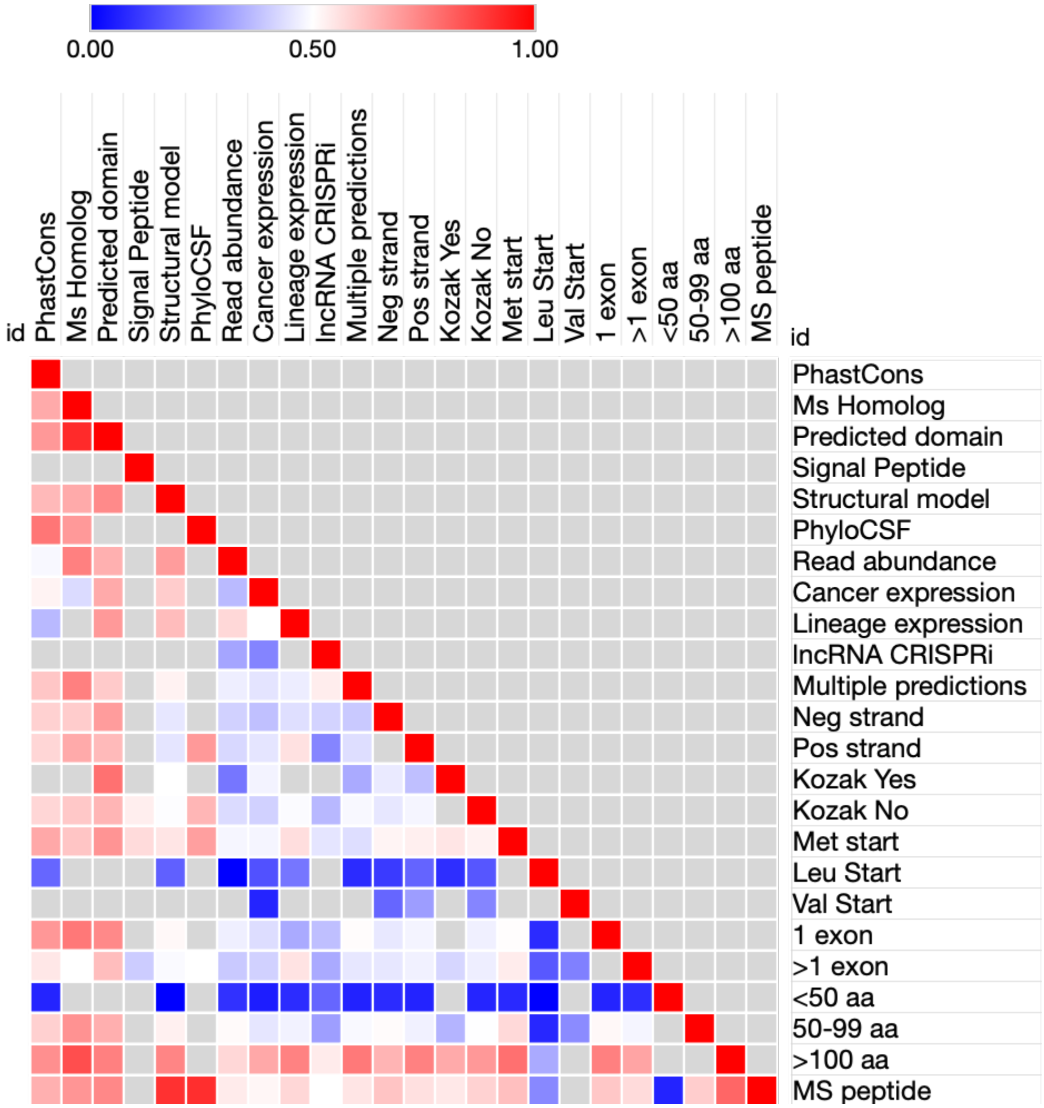


Supplementary Figure 3: Thresholds defined for ORF feature analysis.

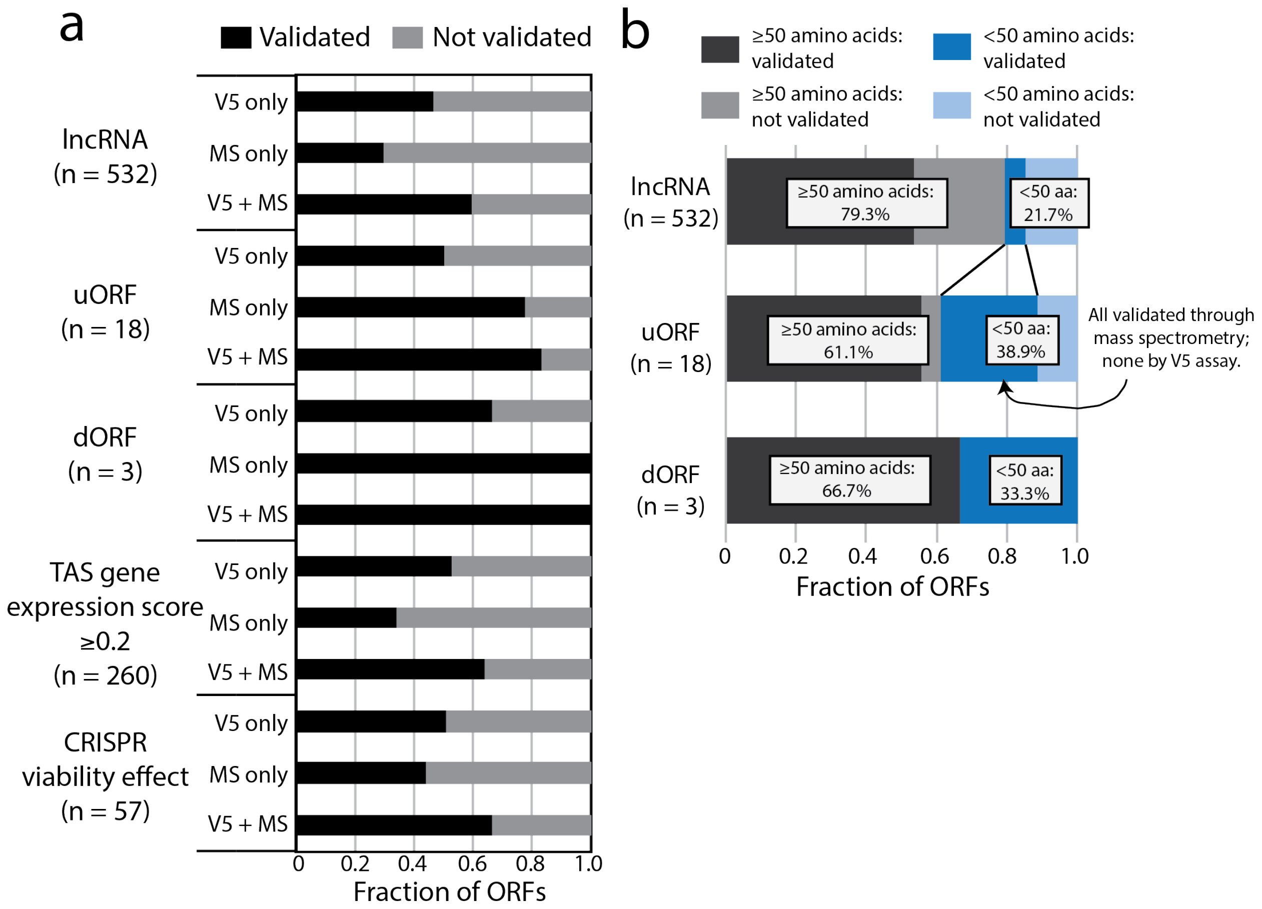
a) The cumulative fraction of selected ORFeome ORFs (n=553) or all manually inspected ORF candidates (n=2,395) scoring for the indicated PhastCons values. **b)** Raw PhastCons scores for the 553 ORFs in the ORFeome, with the indicated threshold used for analysis. **c)** The cumulative fraction of selected ORFeome ORFs (n=553) or all manually inspected ORF candidates (n=2,395) scoring for the indicated PhyloCSF values. **d)** Raw PhyloCSF scores for the 553 ORFs in the ORFeome, with the indicated threshold used for analysis. **e)** The cumulative fraction of selected ORFeome ORFs (n=553) or all manually inspected ORF candidates (n=2,395) scoring for the indicated read/length abundance ratio. **f)** Raw read/length ratios for the 553 ORFs in the ORFeome, with the indicated threshold used for analysis. **g)** The cumulative fraction of selected ORFeome ORFs (n=553) or randomly generated 150mer amino acid sequences (n=500) scoring for the indicated structural confidence score. **h)** Raw structural confidence scores for the 553 ORFs in the ORFeome, with the indicated threshold used for analysis. All p values in this figure were calculated by a two-sided Kolmogorov-Smirnov test.



Supplementary Figure 4: Most detected peptides have multiple sources identifying them.
a) A pie chart showing the percentage of identified tryptic peptide sequences that map to a single ORF or multiple ORFs, with multi-mapping peptides detailed on the right. **b)** *Left*, a Venn diagram demonstrating the numbers of peptides found in literature datasets, PeptideAtlas, and Broad datasets. *Middle*, a pie chart showing the fraction of trypic peptides in the Broad datasets for which more than one source reports the peptide. *Right*, among the peptides with more than one Broad source, the majority have at least 5 sources identifying the peptide. **c)** A scatter plot that shows the specific number of identifying sources per peptide. Data plotted represents only peptide spectrum processed at the Broad Institute (e.g. CPTAC), excluding public peptide repositories (e.g. PeptideAtlas).

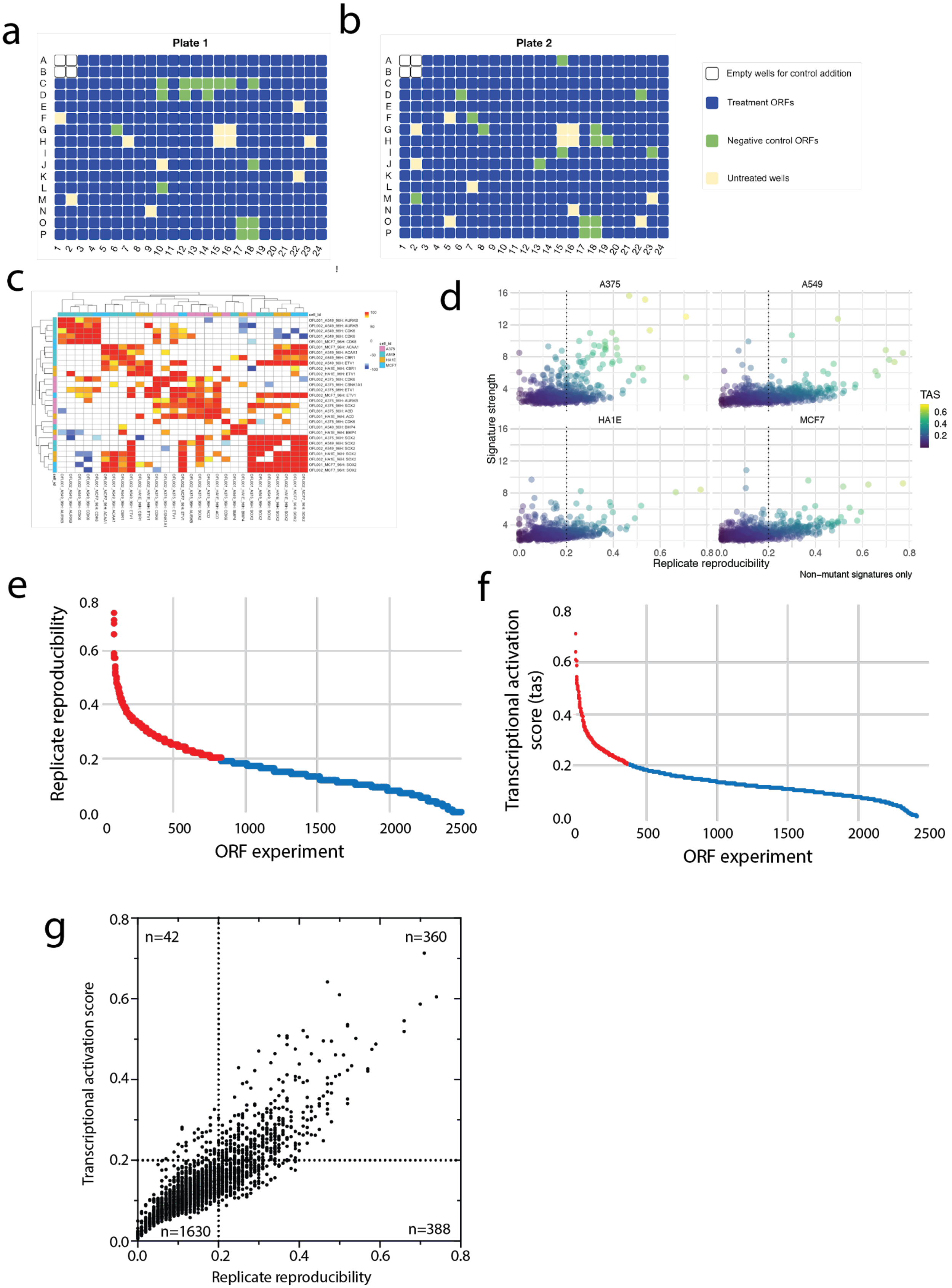


Supplementary Figure 5: Pairwise analysis of ORF features and V5 translation in experimental assays. For each pair of criteria, the fraction of ORFs with those two features that validated by ORFeome V5 detection is plotted.

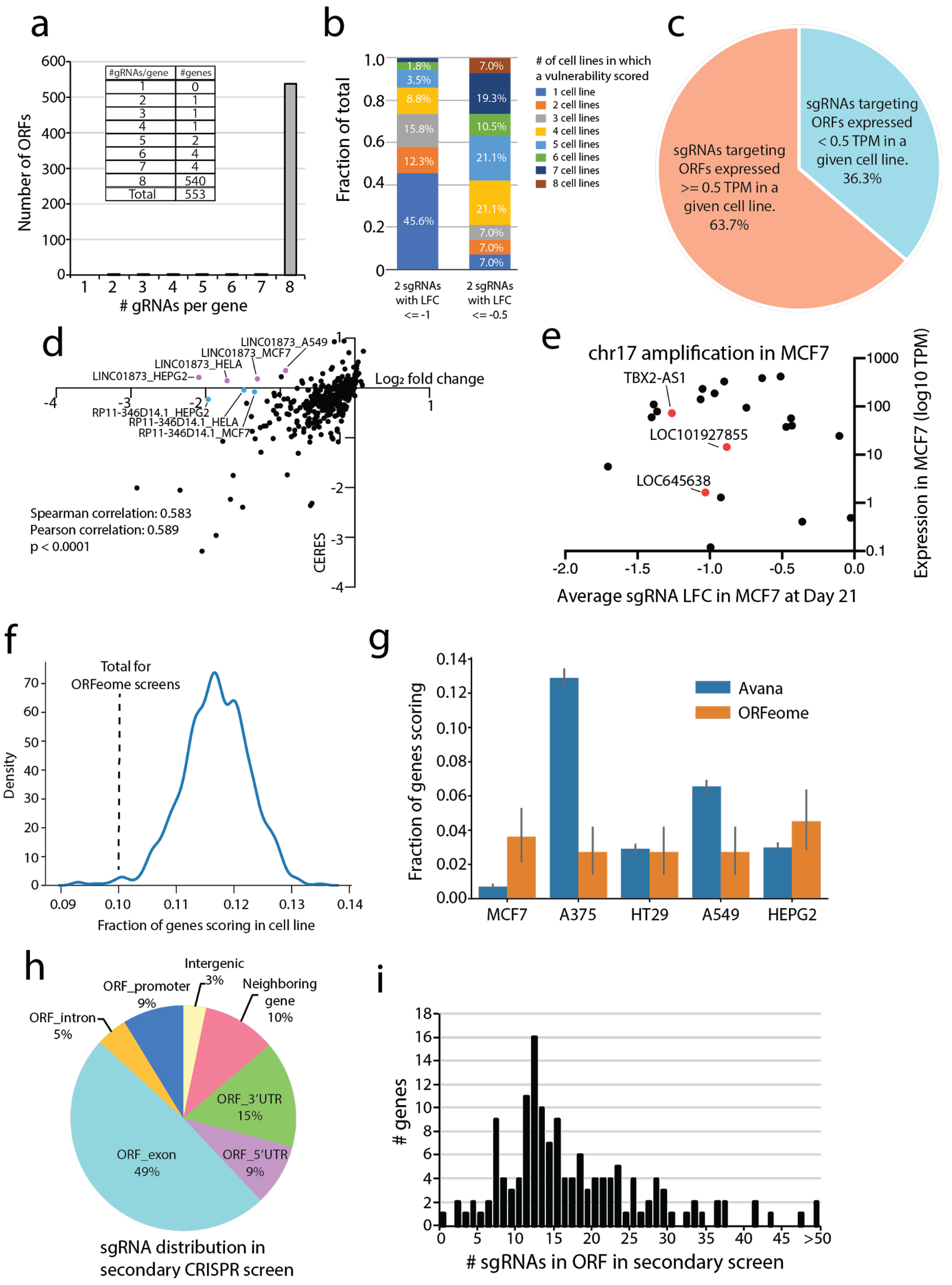


Supplementary Figure 6: Stratification of validated ORFs by nomination type and cellular phenotype.

a) A bar plot showing the fraction of ORFs within each indicated group that validated by V5 *in vitro* translation assay, endogenous mass spectrometry peptides, or the summation of these. **b)** A stacked bar plot showing the fraction of ORFs within each indicated group that validated when stratified by ORF size. ORF size was stratified into ORFs that were >= 50 amino acids in length, or <50 amino acids in length.

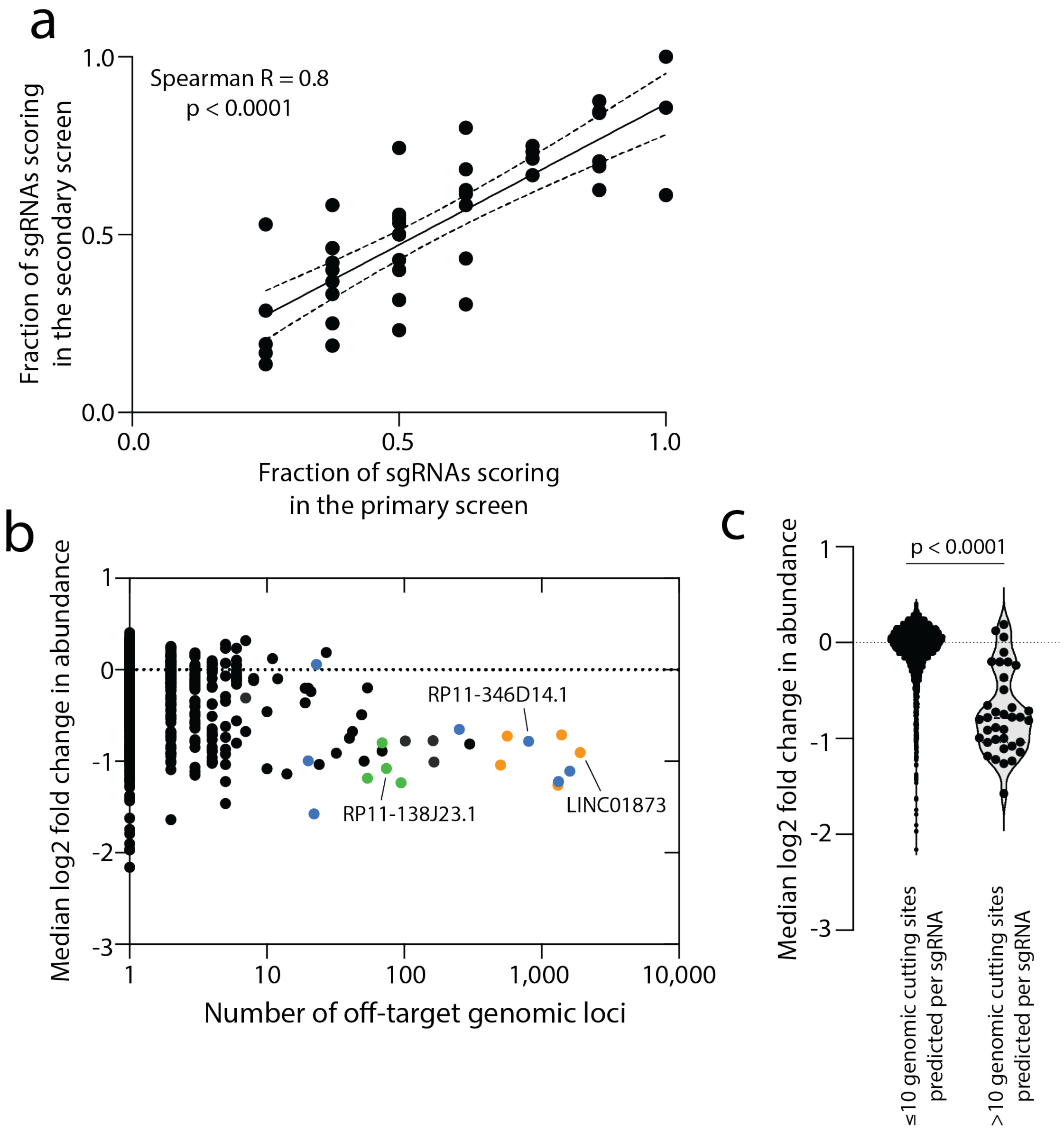

a
Plate 1
A B C D E F G H I J K L M N O P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
b
Plate 2
A B C D E F G H I J K L M N O P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
□ Empty wells for control addition
🟦 Treatment ORFs
🟩 Negative control ORFs
🟨 Untreated wells
c
d
A375
A549
HA1E
MCF7
Signature strength
16
12
8
4
0.0
0.2
0.4
0.6
0.8
Replicate reproducibility
TAS
0.6
0.4
0.2
Non−mutant signatures only
e
Replicate reproducibility
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
ORF experiment
f
Transcriptional activation score (tas)
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
ORF experiment
g
Transcriptional activation score
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
Replicate reproducibility
n=42
n=360
n=1630
n=388

Supplementary Figure 7: ORF gene expression data on the L1000 platform.
a) A L1000 perturbational plate layout showing locations of treatment ORFs, non-human proteins, untreated wells, and technical positive control ORFs. **b)** A second L1000 perturbational plate layout showing locations of treatment ORFs, non-human proteins, untreated wells, and technical positive control ORFs. **c)** Level 5 L1000 data processing (“MODZ” score) and clustering of L1000 signatures for positive control ORFs with a TAS score of >= 0.2. Color red in cells denotes a connectivity score of 95 percentile or greater (similar signatures); blue denotes <= -95 percentile (dissimilar signatures). **d)** Scatter plots of L1000 data for experimental ORFs. The Y axis represents signature strength and the X axis represents reproducibility, the two metrics used to calculate the TAS score. Each TAS score is indicated by the color code of each individual ORF. Each data point represents one ORF. **e)** The distribution of replicate reproducibility scores across all L1000 experiments. Red denotes signatures >= 0.2, which indicated that a signature was present. Blue denotes signatures < 0.2, which denotes that a signature was not detected. **f)** The distribution of transcriptional activation scores (TAS) across all L1000 experiments. Red denotes signatures >= 0.2, which indicated that a signature was present. Blue denotes signatures < 0.2, which denotes that a signature was not detected. **g)** Intersection of replicate reproducibility and TAS scores shows a high degree of correlation. 360 signatures were considered positive for both replicate reproducibility and high TAS score.

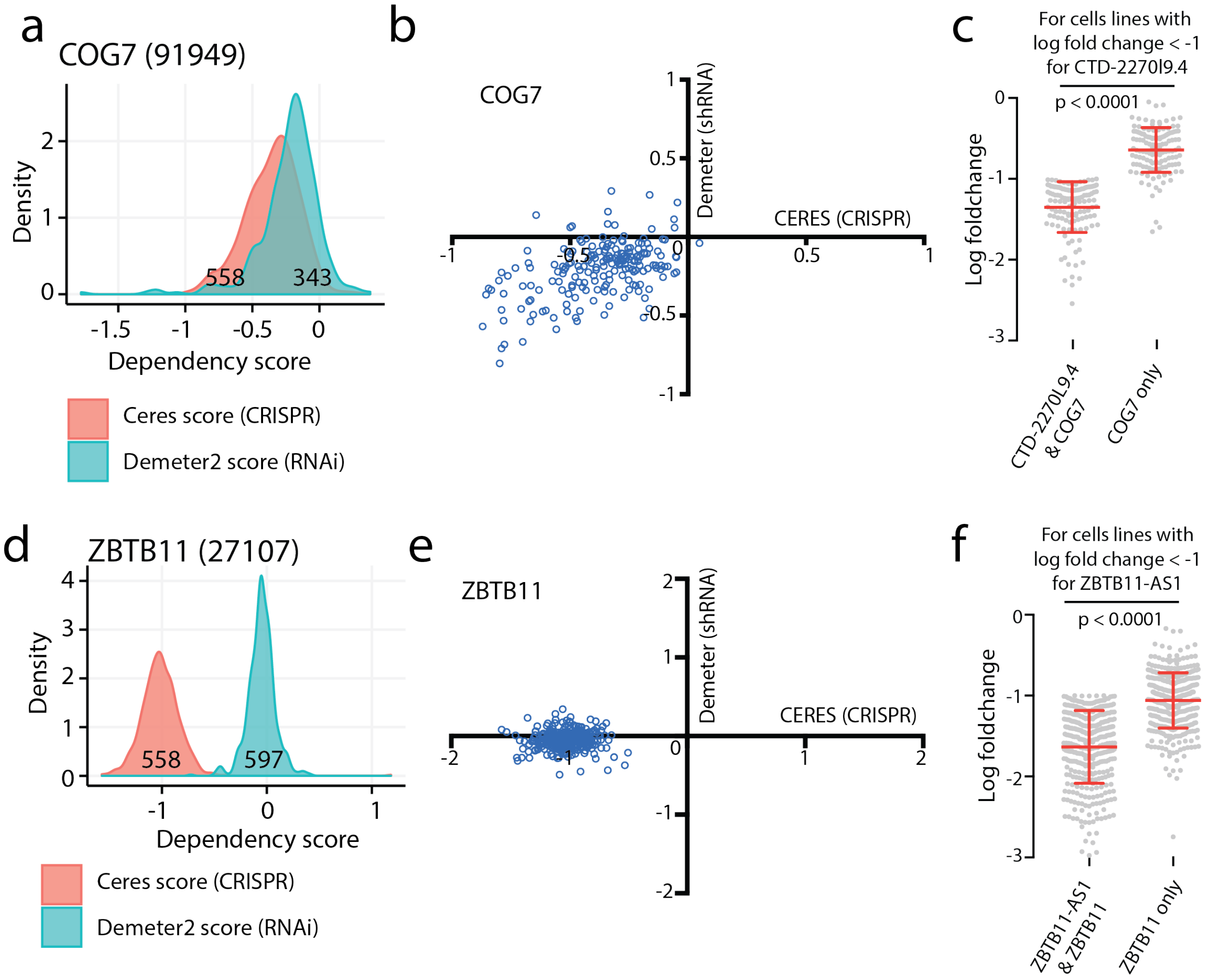


Supplementary Figure 8: CRISPR screens for new ORFs.

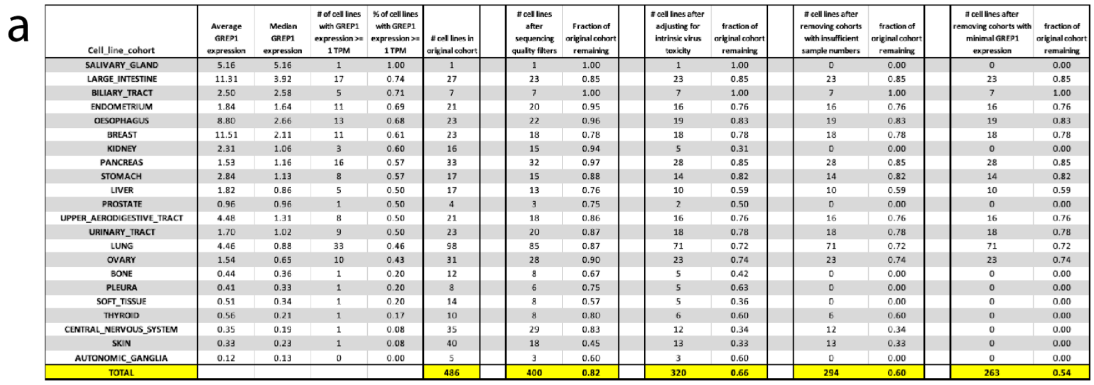
a) A barplot and inset table showing the number of sgRNAs per ORF in the primary CRISPR screen. **b)** Frequency distribution of putative CRISPR hits using a viability threshold of log fold change of <= -1 or <= -0.5 in the primary CRISPR screen. **c)** The percentage of nominated CRISPR hits which had minimal detectable expression or expressed above the threshold of >= 0.5 TPM. **d)** The correlation between log fold change values (n=385 independent measurements) for nominated CRISPR hits and the CERES score for each gene, which integrates copy number data for each cell line. Spearman and Pearson correlations are shown with a two-sided Spearman's p value shown. **e)** An example of the chr17q23 amplification locus in MCF7 cells. CRISPR knockout of genes (n=22 independent experiments) in the locus result in nonspecific cell death due to excessive genomic cutting, regardless of gene expression level. Three putative ORFs were located in this genomic region, indicated with red dots in the figure. **f)** A histogram showing the fraction of genes that would score as a vulnerability gene from a randomly selected set of 500 annotated genes from cell lines in the Cancer Dependency Map. The ORFeome CRISPR screen result is indicated. **g)** The rate of genes scoring as viability genes in the canonical Avana gene library and the ORFeome sgRNA library for the five cell lines shared between both screens. Barplots represent mean and error bars represent standard deviation. **h)** The distribution of sgRNAs across various genome regions in the secondary CRISPR screen. **i)** A histogram showing the number of sgRNAs per ORF in the secondary CRISPR screen.



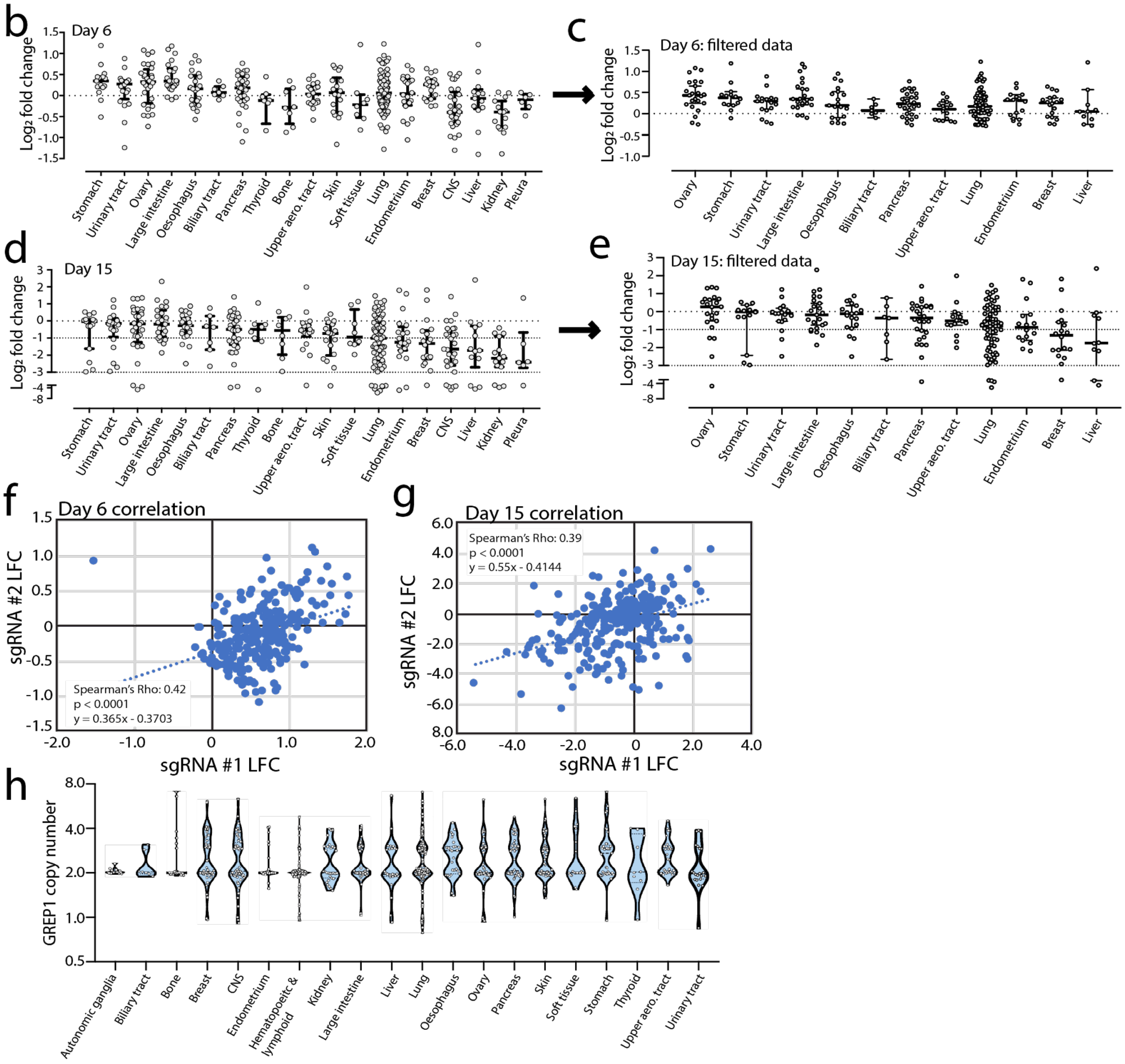
Supplementary Figure 9: Specificity and off-target effects of primary CRISPR screen.
a) A comparison of the fraction of sgRNAs that demonstrated a viability phenotype in the primary screen and secondary screen for genes (n=37) that had >5 sgRNAs in both screens. Significance is by a two-sided Spearman's Rho test. b) The number of off-target genomic effects of each sgRNA (n=4391 independent experiments) compared to the fold change of sgRNA representation at the Day 21 timepoint after lentiviral infection in the CRISPR screen. Three ORFs with off-target sgRNAs are highlighted. c) A violin plot showing the median log2 fold change in sgRNA abundance at Day 21 in the primary CRISPR screen for sgRNAs with =< 10 genomic cutting sites (n= 4355 independent experiments) predicted of >10 genomic cutting sites predicted (n=36 independent experiments). Genomic cutting sites predicted by the Cas-OFFinder algorith. P value by a two-tailed Student's t test.



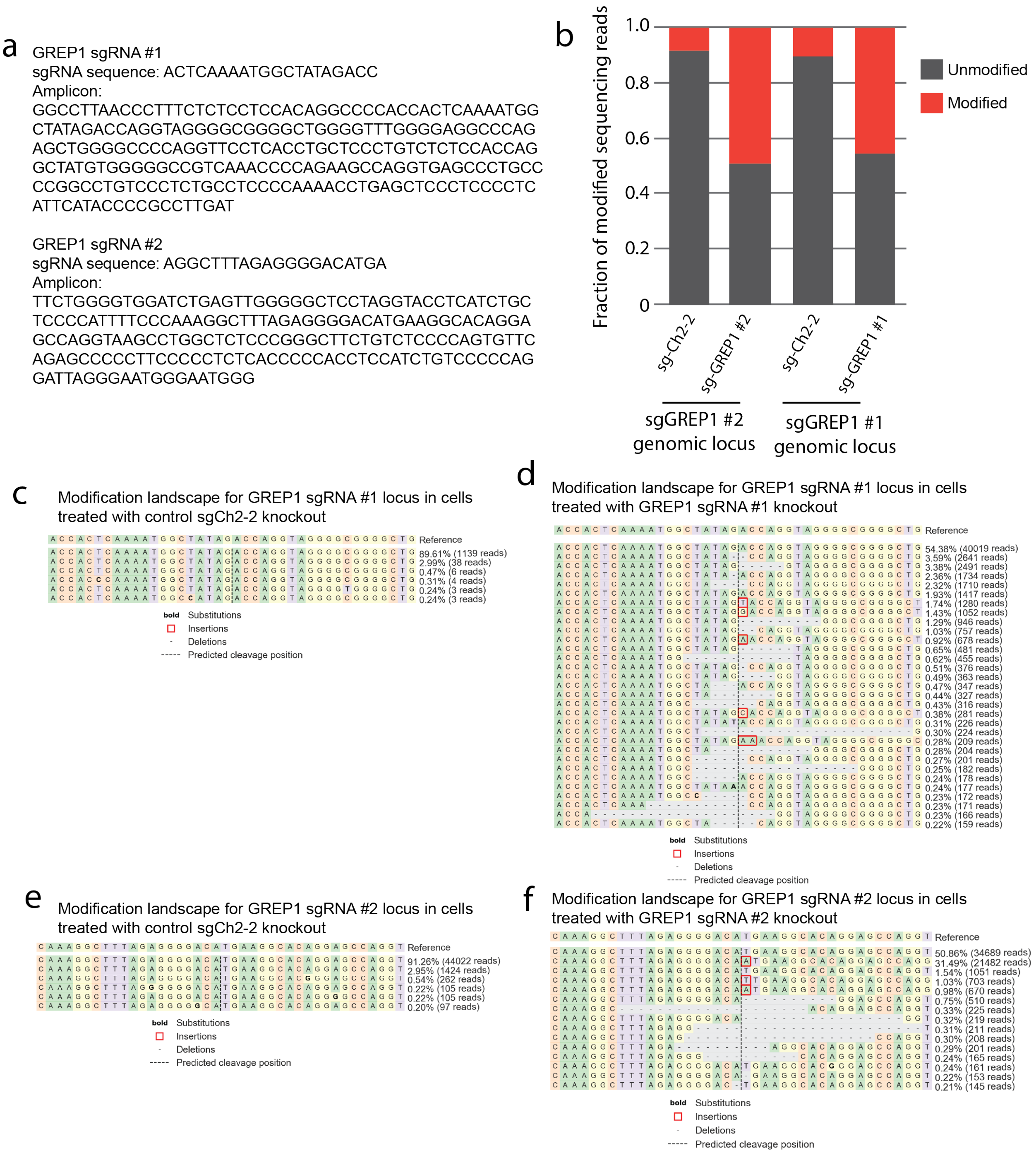
Supplementary Figure 10: Discordant RNAi and CRISPR data for two overlapping ORFs.
a) The dependency profile for COG7 using RNAi or CRISPR data. **b)** A scatter plot comparing the magnitude of dependency phenotype for individual cell lines in RNAi or CRISPR data. **c)** A comparison of the log fold change in cell abundance using the average LFC of the two sgRNAs targeting CTD-2270L9.4 and COG7, compared to two sgRNAs targeting COG7 alone. Only cell lines with a viability phenotype in the CTD-2770L9.4 targeting sgRNAs are shown. N=132 independent cell lines. P value by a two-tailed Student's t test. **d)** The dependency profile for ZBTB11 in RNAi or CRISPR data. **e)** A scatter plot comparing the magnitude of dependency phenotype for individual cell lines in RNAi or CRISPR data. **f)** A comparison of the log fold change in cell abundance using the average LFC of the two sgRNAs targeting ZBTB11 and ZBTB11-AS1, compared to two sgRNAs targeting ZBTB11 alone. Only cell lines with a viability phenotype in the ZBTB11-AS1 targeting sgRNAs are shown. N=384 independent cell lines. P value by a two-tailed Student's t test.



Cell_line_cohort	Average GREP1 expression	Median GREP1 expression	# of cell lines with GREP1 expression >= 1 TPM	% of cell lines with GREP1 expression >= 1 TPM	# cell lines in original cohort		# cell lines after sequencing quality filters	Fraction of original cohort remaining		# cell lines after adjusting for intrinsic virus toxicity	fraction of original cohort remaining		# cell lines after removing cohorts with insufficient sample numbers	fraction of original cohort remaining		# cell lines after removing cohorts with minimal GREP1 expression	fraction of original cohort remaining
SALIVARY_GLAND	5.16	5.16	1	1.00	1		1	1.00		1	1.00		0	0.00		0	0.00
LARGE_INTESTINE	11.31	3.92	17	0.74	27		23	0.85		23	0.85		23	0.85		23	0.85
BILIARY_TRACT	2.50	2.58	5	0.71	7		7	1.00		7	1.00		7	1.00		7	1.00
ENDOMETRIUM	1.84	1.64	11	0.69	21		20	0.95		16	0.76		16	0.76		16	0.76
OESOPHAGUS	8.80	2.66	13	0.68	23		22	0.96		19	0.83		19	0.83		19	0.83
BREAST	11.51	2.11	11	0.61	23		18	0.78		18	0.78		18	0.78		18	0.78
KIDNEY	2.31	1.06	3	0.60	16		15	0.94		5	0.31		0	0.00		0	0.00
PANCREAS	1.53	1.16	16	0.57	33		32	0.97		28	0.85		28	0.85		28	0.85
STOMACH	2.84	1.13	8	0.57	17		15	0.88		14	0.82		14	0.82		14	0.82
LIVER	1.82	0.86	5	0.50	17		13	0.76		10	0.59		10	0.59		10	0.59
PROSTATE	0.96	0.96	1	0.50	4		3	0.75		2	0.50		0	0.00		0	0.00
UPPER_AERODIGESTIVE_TRACT	4.48	1.31	8	0.50	21		18	0.86		16	0.76		16	0.76		16	0.76
URINARY_TRACT	1.70	1.02	9	0.50	23		20	0.87		18	0.78		18	0.78		18	0.78
LUNG	4.46	0.88	33	0.46	98		85	0.87		71	0.72		71	0.72		71	0.72
OVARY	1.54	0.65	10	0.43	31		28	0.90		23	0.74		23	0.74		23	0.74
BONE	0.44	0.36	1	0.20	12		8	0.67		5	0.42		0	0.00		0	0.00
PLEURA	0.41	0.33	1	0.20	8		6	0.75		5	0.63		0	0.00		0	0.00
SOFT_TISSUE	0.51	0.34	1	0.20	14		8	0.57		5	0.36		0	0.00		0	0.00
THYROID	0.56	0.21	1	0.17	10		8	0.80		6	0.60		6	0.60		0	0.00
CENTRAL_NERVOUS_SYSTEM	0.35	0.19	1	0.08	35		29	0.83		12	0.34		12	0.34		0	0.00
SKIN	0.33	0.23	1	0.08	40		18	0.45		13	0.33		13	0.33		0	0.00
AUTONOMIC_GANGLIA	0.12	0.13	0	0.00	5		3	0.60		3	0.60		0	0.00		0	0.00
TOTAL					486		400	0.82		320	0.66		294	0.60		263	0.54



Supplementary Figure 11: Pooled GREP1 knockout across cell lines.
a) A table summarizing all input cell lines in the pool and filters applied to the data for final analysis. **b)** All raw cell line viability data at Day +6 prior to data filtering. N=400 independent cell lines, distributed among the indicated cancer types. Each dot represents one cell line. Lines represent median +/- interquartile range. **c)** Cell line viability data at Day +6 after data filtering. N=263 independent cell lines, distributed among the indicated cancer types. Each dot represents one cell line. Lines represent median +/- interquartile range. **d)** All raw cell line viability data at Day +15 prior to data filtering. N=400 independent cell lines, distributed among the indicated cancer types. Each dot represents one cell line. Lines represent median +/- interquartile range. **e)** Cell line viability data at Day +15 after data filtering. N=263 independent cell lines, distributed among the indicated cancer types. Each dot represents one cell line. Lines represent median +/- interquartile range. **f)** Correlation of GREP1 sgRNAs at Day +6 using filtered data. N=263 independent cell lines. P value for the two-sided Spearman's rho is shown. **g)** Spearman's correlation of GREP1 sgRNAs at Day +15 using filtered data. N=263 independent cell lines. P value for the two-sided Spearman's rho is shown. **h)** GREP1 locus copy number profile across cell line tumor types using Cancer Cell Line Encyclopedia data. No cell lineage harbors high-level amplifications. N=731 independent cell lines, distributed among the indicated cancer types. Each dot represents one cell line.



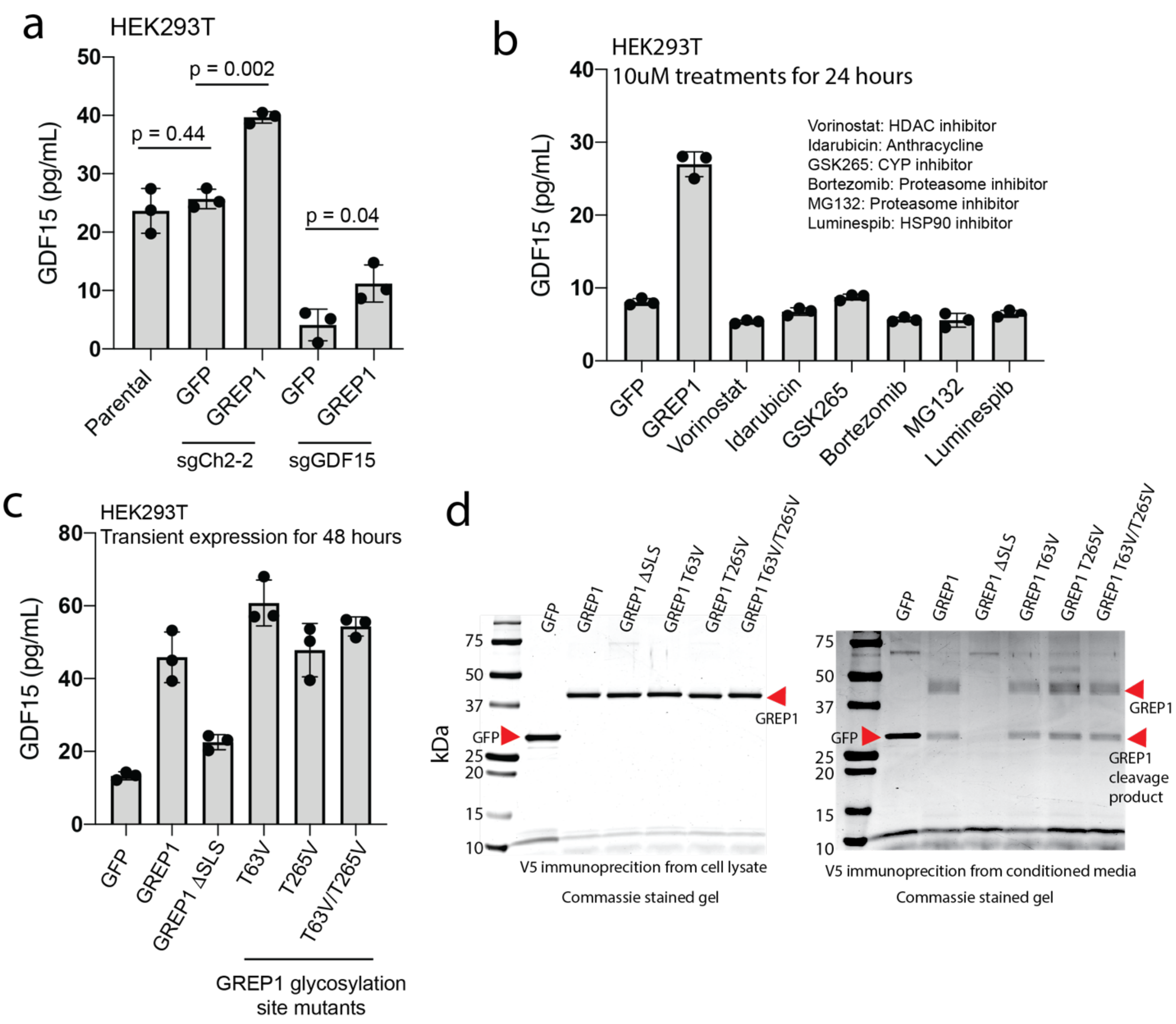
GREP1 sgRNA #1
sgRNA sequence: ACTCAAAATGGCTATAGACC
Amplicon:
GGCCTTAACCCTTTCTCTCCTCCACAGGCCCCACCACTCAAAATGG
CTATAGACCAGGTAGGGGCGGGGCTGGGGTTTGGGGAGGCCCAG
AGCTGGGGCCCCAGGTTCCTCACCTGCTCCCTGTCTCTCCACCAG
GCTATGTGGGGGCCGTCAAACCCCAGAAGCCAGGTGAGCCCTGCC
CCGGCCTGTCCCTCTGCCTCCCCAAAACCTGAGCTCCCTCCCCTC
ATTCATACCCCGCCTTGAT

GREP1 sgRNA #2
sgRNA sequence: AGGCTTTAGAGGGGACATGA
Amplicon:
TTCTGGGGTGGATCTGAGTTGGGGGCTCCTAGGTACCTCATCTGC
TCCCCATTTTCCCAAAGGCTTTAGAGGGGACATGAAGGCACAGGA
GCCAGGTAAGCCTGGCTCTCCCGGGCTTCTGTCTCCCCAGTGTTC
AGAGCCCCCTTCCCCCTCTCACCCCCACCTCCATCTGTCCCCCAG
GATTAGGGAATGGGAATGGG

Supplementary Figure 12: Genome modifications observed with GREP1 sgRNAs.

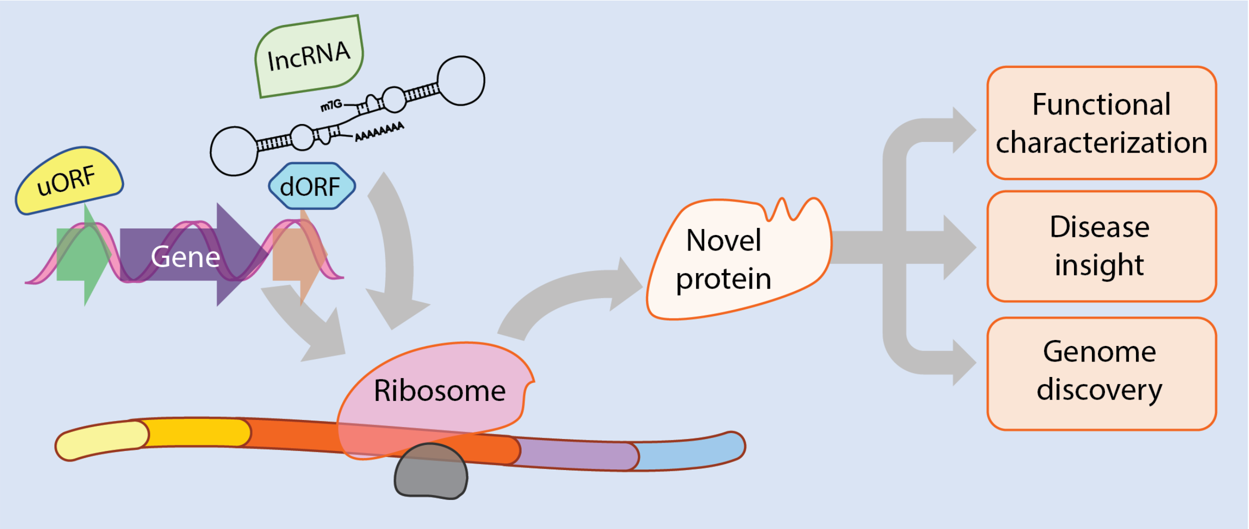
a) GREP1 sgRNA sequences and genomic amplicon subjected to sequencing for modifications. **b)** The fraction of modified reads from the GREP1 sgRNA amplicons in cells treated with either control sgCh2-2 knockout of GREP1 knockout. **c)** The landscape of genomic amplicon modifications for the sgGREP1 #1 locus in cells subjected to control sgCh2-2 knockout. **d)** The landscape of genomic amplicon modifications for the sgGREP1 #1 locus in cells subjected to

GREP1 knockout with sgGREP1 #1. **e)** The landscape of genomic amplicon modifications for the sgGREP1 #2 locus in cells subjected to control sgCh2-2 knockout. **f)** The landscape of genomic amplicon modifications for the sgGREP1 #2 locus in cells subjected to GREP1 knockout with sgGREP1 #2.



Supplementary Figure 13: Increased GDF15 secretion is specific to GREP1 overexpression and not changed by mutation of glycosylation sites.

a) Expression of GREP1 specifically increases GDF15 abundance with non-specific genome cutting and specific genome cutting. Genomic cutting with the sgCh2-2 chromosome 2 locus does not elevate GDF15 abundance in conjunction with GFP overexpression, but genomic cutting with sgCh2-2 along with GREP1 overexpression increases GDF15 abundance. Genomic knockout of GDF15 is partially rescued by GREP1 overexpression. N=3 technical replicates and N=2 independent biological replicates. Barplots represent mean +/- standard deviation. **b)** GDF15 abundance in cell culture media 24 hours after ectopic expression of GFP or GREP1, or treatment with a toxic dose (10uM) of the indicated pharmacologic inhibitors. Pharmacologic inhibitors do not elevate GDF15 levels. N=3 technical replicates and N=2 independent biological replicates. Barplots represent mean +/- standard deviation. **c)** Ectopic expression of GREP1 glycosylation mutants result in equivalent GDF15 accumulation compared to wild type GREP1. N=3 technical replicates and N=2 independent biological replicates. Barplots represent mean +/- standard deviation. **d)** *Left*, Commassie stained gel demonstrating protein expression of GREP1 constructs in the whole cell lysate for the experiment shown in **c**. *Right*, Commassie stained gel demonstrated abundance of GREP1 in the conditioned media for the experiment shown in **c**. P values in this figure determined by a two-tailed Student's T test.



Supplementary Figure 14: A graphical model

Supplementary discussion

Historical perspectives on the human genome annotation

The human genome is now generally felt to have ~19,029 protein-coding genes (*Homo sapiens* CCDS release 22 as of October 10, 2019). The single largest gene discovery project was the Human Genome Project (HGP). The RefSeq database included approximately ~10,000 genes prior to publication of the HGP[1,2], which had doubled from the 4,270 genes in the July 1995 GenBank Release 89.9[3]. Many of these genes were known from positional cloning and other techniques.

The initial HGP in 2001 postulated 30,000 - 40,000 human genes. By itself, this was a dramatic reduction in the ~50,000 - ~100,000 anticipated genes[4–6]. However, by the revision of the HGP in 2004, this number had been decreased to 20,000 - 25,000[7]. It was subsequently reduced to ~19,000, with ~17,600 confidently observed by mass spectrometry[8]. This number has been the current estimate for the past 10 years and the number used as the basis of all exome sequencing studies.

Assumptions made during gene discovery

In the HGP, mRNAs were queried for the presence of an open reading frame that was >= 100 amino acids and began with a methionine start codon. If present, this ORF was reported as a novel protein in the HGP. Such methods had basis in precedent, but were not without challenges: the established noncoding RNA Xist was initially reported to have a 894 bp ORF[9] until it was determined that this ORF was not actually coded[10].

Proteins less than 100 amino acids were included in the HGP only if they had been previously known, as such ORFs were difficult to predict due to noise from existing cDNA fragments at that time. Therefore, ORFs less than 100 amino acids were not nominated solely based on computational analyses[11–13].

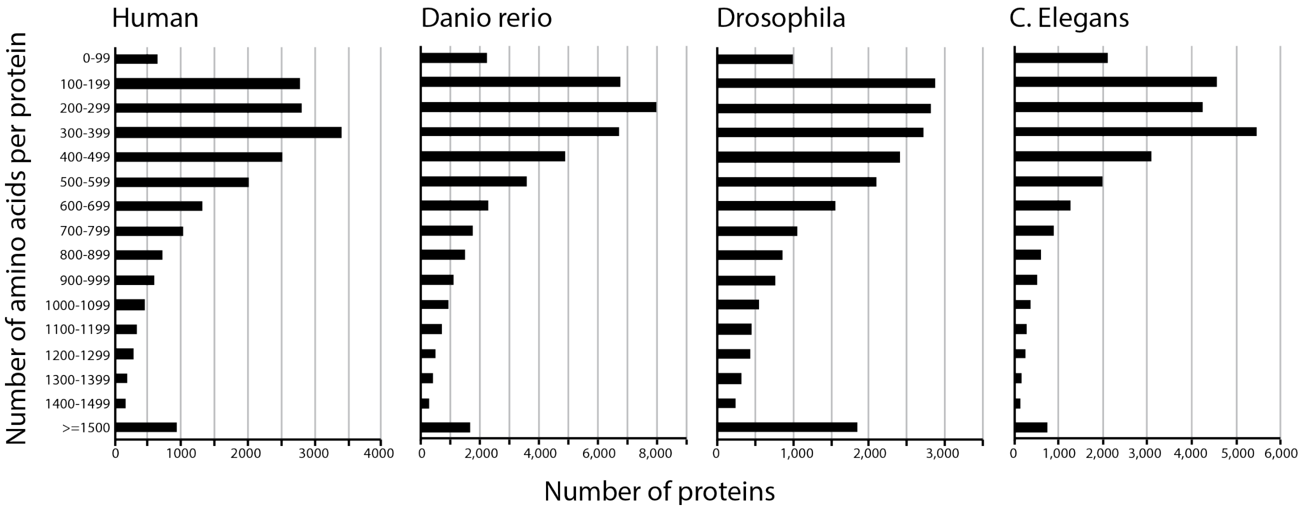
Protein size and function

There is no specific scientific rationale for why smaller proteins would be less real. An analysis by John Mattick and colleagues suggested that an ORF of >100 amino acids was approximately two standard deviations above the average random ORF size in a random 1kb segment of genome sequence[14]. This is statistic, though, is not particularly meaningful as most genes are much longer than 1000bp due to extended untranslated regions (UTRs). However, it highlights the challenge in computationally separating signal from noise.

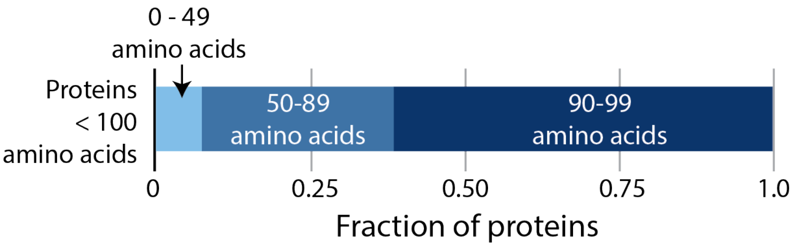
It is not clear whether there is a minimum size required for peptide/protein function. The smallest known functional unit is a zinc finger, which is an aggregation of the Cys_2-His_2 four amino acid motif. It is typically thought that a minimum of four or five such motifs are required for functional zinc finger DNA binding, thus suggesting that a peptide of 20 amino acids or greater may be eligible for this function. Secondary structure for a peptide may exist with as few as four or five amino acids[15], and enkephalins are five amino acid peptides found in the central nervous system and thought to be functional[16]. An alpha helical peptide can be stably produced with 14 amino acids[17]. There are also now known proteins less than 50 amino acids[18,19].

Skew in the size distribution of annotated proteins

Most annotated proteins are >100 amino acids in most organisms. As shown below, the fraction of the annotated proteome for humans, *C. elegans*, *D. melanogaster* and *D. rerio*.



Among human proteins <100 amino acids, 61% are 90 - 99 amino acids large, and thus proteins < 90 amino acids are very rare in annotated databases. Below these data are shown in figure formation for *H. sapiens*.



Methods to validate a putative protein

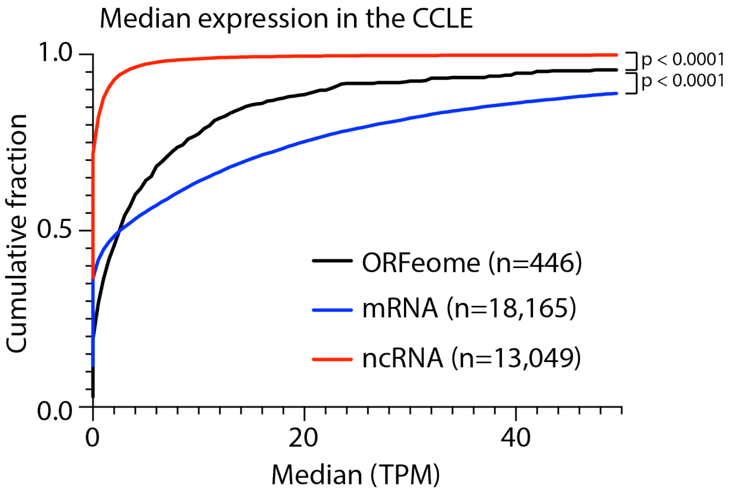
Once a potential protein is identified, there are many possible ways to demonstrate its existence. Mass spectrometry of endogenous peptides can provide evidence, though small proteins often have few trypic sites and may not perform well by mass spectrometry. Also, many unannotated proteins are likely lineage-restricted and may not be historically well represented in the mature tissues profiled by mass spectrometry.

Tag-free biochemical transcription/translation with rabbit or wheat germ lysates can be used, but these assays have a high false negative rate and are biochemical assays only. *In vitro* studies can include ribosome profiling/polysome association to see if the mRNA is bound by ribosomes, though this is not direct evidence of translation. Other *in vitro* studies are exogenously expressing an epitope-tagged plasmid construct. However, the epitope tags may destabilize small proteins, leading to protein elimination.

Other approaches include development of a new antibody for a protein for experimental use. This approach is limited as it is expensive and takes a significant amount of time. A genetic knock-in of a fusion-tagged cDNA is also possible, but again costly and time-intensive.

Expression of the ORFeome compared to other lncRNAs

It is well-established that, in general, so-called lncRNAs are more tissue-restricted and lower expressed than annotated human proteins[20,21]. To evaluate the expression level of the ORFs in our ORFeome, we were able to extract gene expression data for 13,049 ncRNA, 18,165 mRNA, and 446 of our ORFs in the Cancer Cell Line Encyclopedia dataset[22]. We found that the ORFs were significantly higher expressed than baseline ncRNAs, though less highly expressed than canonical proteins. See figure below (p values by the Kolmogorov-Smirnov test):



Features of the ORFeome amino acid sequences

For the 490 ORFeome ORFs with predicted amino acid sequences longer than 40 amino acids, we evaluated several biophysical properties, including protein sequence length, number of protein binding-sites, aggregation propensity, disorder and number of Pfam-annotated protein domains. First, the amino acid sequences of these ORFs suggest that they have a large proportion of their outer surface exposed to water (73% ± 0.4%), have a high number of predicted protein-binding sites (12.79 ± 0.2 per 100 aa) and disorder (0.98 ± 0.04 per 100 aa), and that have few Pfam-annotated protein domains (0.08 ± 0.01 per 100 aa). In contrast, average mammalian genes, including human genes, encode much longer proteins of ~500 aa that have a low amount of disorder and high aggregation propensity[23,24].

Distinguishing a predictive structural model from background signal

Predicting protein structure was performed with the PHYRE2 server[25] for the 530 ORFs that were >= 40 amino acids in length. To control for the chance of randomly predicted protein structure, we created a score to distinguish background signal. Each amino acid sequence was given a percent confidence score and an alignment coverage percentage by the PHYRE2 server. We multiplied these two numbers together to create a protein structure score. We then computationally generated a list of 500 random 150 amino acid peptides with a methionine start site, and analyzed these in the same manner. We used the distribution of these datasets to define a threshold for determining the presence of a robust structural prediction.

Updated annotation status of the ORFs in this manuscript

This project was initiated in January 2016 and therefore we employed databases available at that time. Over the past several years, these gene annotation databases have been updated, but our study was not able to accommodate changes in annotation status due to the nature of large-scale ORF and CRISPR library generation for functional genomics. Therefore, a subset of the genes included in this study are now annotated in the recent versions of GENCODE. A few of the ORFs in this study have now been functionally characterized and published in other studies as well.

We have now re-evaluated the annotation status of our ORFs in GENCODE v31. There are 61 ORFs that are now annotated as protein-coding in GENCODE v31. 43 of these 61 (70.5%) are annotated as the same ORF in GENCODE v31 as in our ORFeome. 2 of the 61 are annotated as different ORFs in the two databases. 44 of the 61 (74%) validated in our V5 western blot assay as a translated protein. The table below shows a list of ORFs that are now annotated as protein-coding, along with the current transcript name and a publication investigating that ORF, if available.

Name	**GENCODE v31 name**	**Validation percentile**	**Validated?**	**Publications**
LINC01420	NBDY	1	Yes	8,26–28
LINC00116	MTLN	0.998	Yes	27,29,30
CHTF8	DERPC	0.996	Yes	31
ASNSD1	ASDURF	0.989	Yes	8,28
RPP14_ORF1	HTD2	0.972	Yes	
RP11-429J17.8	IQANK1	0.971	Yes	
LINC00693	AC098650.1	0.967	Yes	
LOC105371267	AC007906.2	0.965	Yes	
AATK1-AS1	PVALEF	0.961	Yes	
LINC01314	CTXND1	0.958	Yes	
LOC284023	RNF227	0.956	Yes	

LINC00371	C13orf42	0.954	Yes	
LOC93622	AC093323.1	0.945	Yes	32
LOC728743	AC073111.4	0.943	Yes	
PIGBOS1	PIGBOS1	0.932	Yes	
RP11-680F20.6	VSIG10L2	0.929	Yes	
EFCAB10	EFCAB10	0.923	Yes	
G029442	LINC00514	0.916	Yes	
LOC389332	SMIM32	0.902	Yes	
LINC00176	C20orf204	0.882	Yes	
RP11-195B21.3	RP11-195B21.3	0.869	Yes	
MIEF1	AL022312.1	0.865	Yes	8,28,33
SLC35A4_ORF1	SLC34A4	0.856	Yes	8,28
ERVK3-1	ERVK3-1	0.831	Yes	
LINC00094	BRD3OS	0.822	Yes	
ZNF525	ZNF525	0.818	Yes	
LOC100133315	AP002495.1	0.817	Yes	
AP000783.1	GRAM1B	0.809	Yes	
TINCR	TINCR	0.8	Yes	
C5orf56	AC116366.3	0.798	Yes	
MKKS	AL034430.2	0.789	Yes	8,28,34
NCBP2-AS2	NCBP2AS2	0.784	Yes	
FAM83H-AS1	IQANK1	0.782	Yes	
TOPORS-AS1	SMIM27	0.78	Yes	
RP11-689K5.3	part of RASGEF1B	0.759	Yes	
LINC00961	SPAAR	0.74	Yes	35
CTD-3088G3.8	AC099489.1	0.73	Yes	

LINC00493	SMIM26	0.719	Yes	
LINC00998	SMIM30	0.684	Yes	
LINC01272	SMIM25	0.679	Yes	
G086960	PRRT1B	0.641	Yes	
ZNF738	ZNF738	0.548	Yes	
RP11-539I5.1	part of HSPA12A	0.536	Yes	
RP11-166B2.1	NPIPB2	0.516	No	
SNHG3	Part of RCC1	0.512	No	
DDIT3	AC022506.1	0.51	No	
AP000783.2	GRAM1B	0.503	No	
PNRC2	PNRC2	0.498	No	
TCONS_l2_00007040	CFAP97D2	0.465	No	
ZNF66	ZNF66	0.413	No	
FAM220A	AC009412.1	0.409	No	
LOC100507002	Part of RGS9	0.369	No	
RP11-345F18.1	EXOC1L	0.269	No	
SPTY2D1-AS1	SPTY2D1OS	0.24	No	
MMP24-AS1	MMP24OS	0.192	No	
LINC00617	TUNAR	0.13	No	
LOC105372440	AC010325.1	0.092	No	
PTP4A1	AL135905.2	0.056	No	
FTCDNL1	FTCDNL1	0.027	No	
LINC00634	LINC00634, but it is a different ORF	0.039	No	
RP11-295G20.2	AL445524.2, but it is a different ORF	0.728	Yes	

Supplementary references

1. Zoubak, S., Clay, O. & Bernardi, G. The gene distribution of the human genome. *Gene* **174**, 95–102 (1996).
2. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
3. Benson, D. A., Boguski, M., Lipman, D. J. & Ostell, J. GenBank. *Nucleic Acids Research* vol. 22 3441–3444 (1994).
4. Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nat. Genet.* **7**, 345–346 (1994).
5. Schuler, G. D. *et al.* A gene map of the human genome. *Science* **274**, 540–546 (1996).
6. Liang, F. *et al.* Gene Index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25**, 239–240 (2000).
7. Consortium, I. H. G. S. & International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* vol. 431 931–945 (2004).
8. Omenn, G. S. *et al.* Progress on Identifying and Characterizing the Human Proteome: 2019 Metrics from the HUPO Human Proteome Project. *Journal of Proteome Research* (2019) doi:10.1021/acs.jproteome.9b00434.
9. Borsani, G. *et al.* Characterization of a murine gene expressed from the inactive X chromosome. *Nature* vol. 351 325–329 (1991).
10. Brockdorff, N. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* vol. 71 515–526 (1992).

11. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences* vol. 104 19428–19433 (2007).

12. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* vol. 268 78–94 (1997).

13. Consortium, M. G. S. & Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* vol. 420 520–562 (2002).

14. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLoS Computational Biology* vol. 4 e1000176 (2008).

15. Kumar, N. & Kishore, R. Determination of an unusual secondary structural element in the immunostimulating tetrapeptide rigin in aqueous environments: insights via MD simulations, 1H NMR and CD spectroscopic studies. *Journal of Peptide Science* (2010) doi:10.1002/psc.1260.

16. Marcotte, I., Separovic, F., Auger, M. & Gagné, S. M. A Multidimensional 1H NMR Investigation of the Conformation of Methionine-Enkephalin in Fast-Tumbling Bicelles. *Biophysical Journal* vol. 86 1587–1600 (2004).

17. Imura, T. *et al.* Minimum Amino Acid Residues of an α-Helical Peptide Leading to Lipid Nanodisc Formation. *Journal of Oleo Science* vol. 63 1203–1208 (2014).

18. Anderson, D. M. *et al.* A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* vol. 160 595–606 (2015).

19. Nelson, B. R. *et al.* A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271–275 (2016).

20. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).

21. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* vol. 25 1915–1927 (2011).

22. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

23. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**, 117 (2013).

24. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature Ecology & Evolution* vol. 1 (2017).

25. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

26. D'Lima, N. G. *et al.* A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **13**, 174–180 (2017).

27. Paik, Y.-K. *et al.* Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function. *J. Proteome Res.* **17**, 4042–4050 (2018).

28. Omenn, G. S. *et al.* Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **17**, 4031–4041 (2018).

29. Makarewich, C. A. *et al.* MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-

Oxidation. *Cell Rep.* **23**, 3701–3709 (2018).

30. Chugunova, A. *et al.* LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 4940–4945 (2019).

31. Sun, M. *et al.* A Human Novel Gene DERPC Located on 16q22.1 Inhibits Prostate Tumor Cell Growth and Its Expression Is Decreased in Prostate and Renal Tumors. *Molecular Medicine* vol. 8 655–663 (2002).

32. Liu, L. *et al.* Interaction between p12CDK2AP1 and a novel unnamed protein product inhibits cell proliferation by regulating the cell cycle. *Mol. Med. Rep.* **9**, 156–162 (2014).

33. Rathore, A. *et al.* MIEF1 Microprotein Regulates Mitochondrial Translation. *Biochemistry* **57**, 5564–5575 (2018).

34. Akimoto, C. *et al.* Translational repression of the McKusick–Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1830**, 2728–2738 (2013).

35. Matsumoto, A. *et al.* mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* vol. 541 228–232 (2017).