nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-00806-2

Noncanonical 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] and Todd R. Golub[1,2,3] ✉

Although genomic analyses predict many noncanonical open reading frames (ORFs) in the human genome, it is unclear whether they encode biologically active proteins. Here we experimentally interrogated 553 candidates selected from noncanonical ORF datasets. Of these, 57 induced viability defects when knocked out in human cancer cell lines. Following ectopic expression, 257 showed evidence of protein expression and 401 induced gene expression changes. Clustered regularly interspaced short palindromic repeat (CRISPR) tiling and start codon mutagenesis indicated that their biological effects required translation as opposed to RNA-mediated effects. We found that one of these ORFs, *G029442*—renamed *g*lycine-rich extracellular protein-1 (GREP1)—encodes a secreted protein highly expressed in breast cancer, and its knockout in 263 cancer cell lines showed preferential essentiality in breast cancer-derived lines. The secretome of GREP1-expressing cells has an increased abundance of the oncogenic cytokine GDF15, and GDF15 supplementation mitigated the growth-inhibitory effect of *GREP1* knockout. Our experiments suggest that noncanonical ORFs can express biologically active proteins that are potential therapeutic targets.

Early analyses of the human genome sequence suggested the existence of 100,000 or more protein-coding genes, but further scrutiny revealed that the majority of those candidate genes were more likely to be producing noncoding RNAs, fragmented complementary DNA clones or RNAs expressed at inconsequential levels[1–3]. The current Human Proteome Project NeXtProt database recognizes ~17,600 protein-coding genes confirmed by mass spectrometry and ~2,100 unconfirmed protein-coding genes[4]. Nevertheless, a growing body of evidence utilizing high-throughput profiling of ribosome-associated RNAs suggests that additional, noncanonical translation exists in genes currently annotated as noncoding RNAs or pseudogenes, as well as 5′ and 3′ untranslated regions (UTRs) of protein-coding genes[5–8]. Nevertheless, it is unclear whether such translation reflects proteins overlooked during the construction of reference genome databases[9–12], leaky ribosome scanning or confounded computational predictions[13–15], since stringent conservation-based analyses have added only a small number of new proteins to the human genome[13]. Indeed, systematic experimental evidence interrogating whether such predicted proteins are in fact stably translated and biologically functional is lacking.

To address this, we curated a list of 553 high-priority ORFs nominated in long noncoding RNAs and regions upstream and downstream of known protein-coding genes (uORFs and dORFs, respectively). These were selected based on integrative analyses of published predictions of ORF translation, with additional analyses to eliminate pseudogenes and ORFs representing variants of known protein-coding regions[5,6,14,16–33] (Supplementary Table 1, Supplementary Figs. 1 and 2 and Methods). At least two independent studies identified 227/553 (41%) as translated. Overall, mass spectrometry and computational predictions contributed fewer candidates compared to ribosome-profiling datasets (Supplementary Fig. 2). We annotated the 553 ORFs according to 12 metrics including evolutionary conservation, expression and structural features (Supplementary Tables 2–13, Supplementary Fig. 3 and Methods). Out of 553 selected ORFs, 450 (81%) scored highly for at least two metrics in support of relevance (Supplementary Fig. 1 and Supplementary Table 2).

We next asked whether systematic functional studies could test the predicted translation of these ORFs (Fig. 1a). The capacity for ORFs to produce a stably translated protein was assessed by three independent methods. First, we queried independent, publicly available mass spectrometry databases (Methods) and observed 707 distinct tryptic peptides supporting 174 of 553 ORFs (31%). Many tryptic peptides were reproducibly detected in numerous independent samples and datasets, for a total of 6,724 peptides identified (Supplementary Fig. 4 and Supplementary Tables 14 and 15). Next, we designed a cDNA expression library of the 553 ORFs containing a V5 epitope tag and developed a scalable assay for individual protein evaluation by anti-V5 detection (Fig. 1b and Extended Data Fig. 1a–d). A total of 257 ORFs (46%) yielded a V5-tagged protein detectable by in-cell visualization (Fig. 1c–e, Extended Data Fig. 1e–g and Supplementary Table 16). ORFs nominated through ribosome

[1]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [2]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [3]Division of Pediatric Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA. [4]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [5]IT-Research Computing, Harvard Medical School, Boston, MA, USA. [6]Department of Pharmacology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. [7]Department of Biomedical Engineering, McCormick School of Engineering, Northwestern University, Evanston, IL, USA. [8]Present address: Inzen Therapeutics, Cambridge, MA, USA. [9]Present address: Merck Research Laboratories, Boston, MA, USA. ✉e-mail: golub@broadinstitute.org

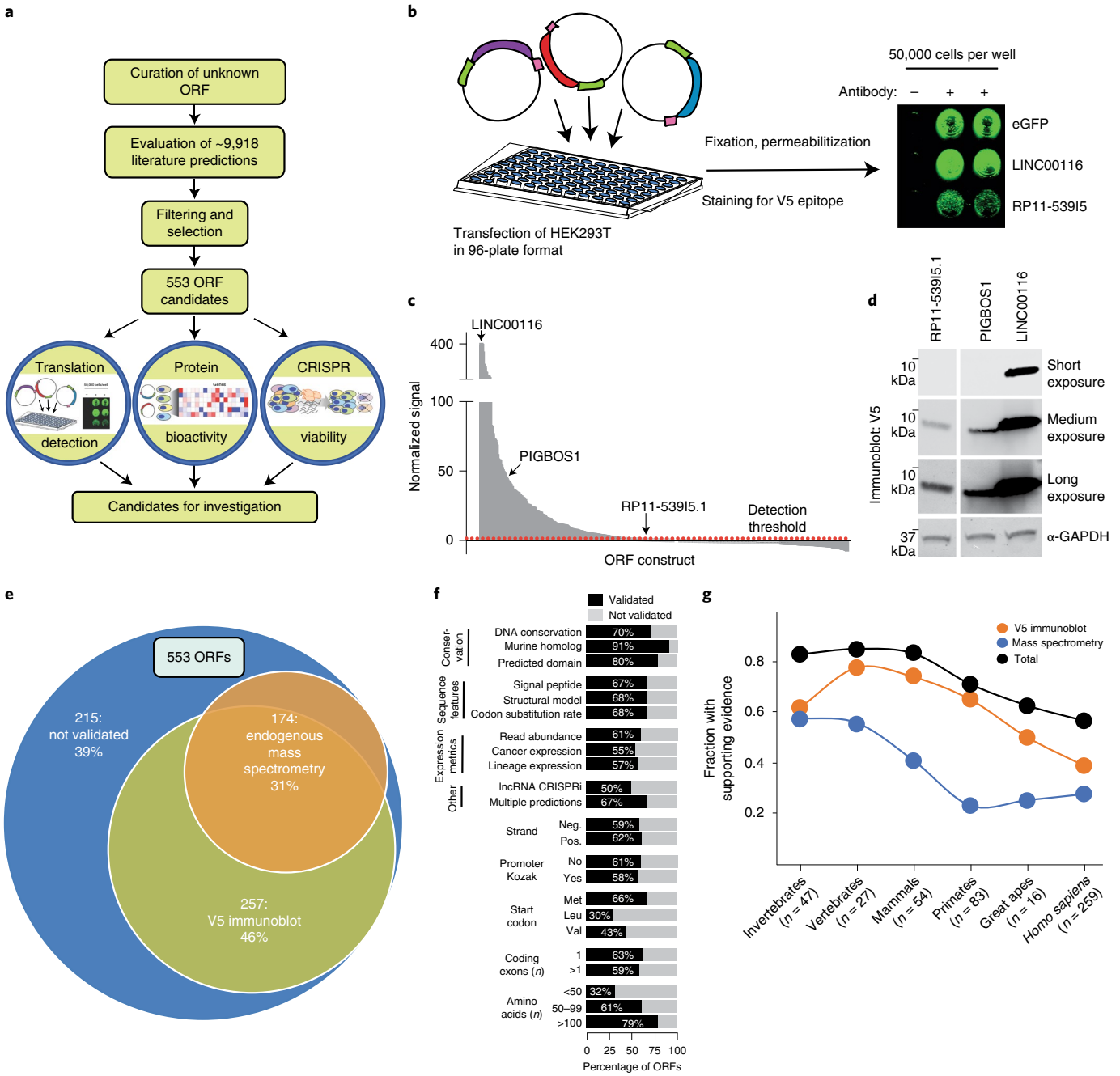


Fig. 1 | Identification of translated unannotated or unstudied ORFs. a, Schematic overview of the research project. **b**, Experimental setup for in vitro detection of protein translation by transfection of V5-tagged cDNAs into HEK293T cells followed by in-cell immunoblotting. **c**, In-cell immunoblot signal for each ORF; values are the average of three replicates. **d**, Correlates for three ORFs (shown in **c**) identified by in-cell immunoblotting; results were repeated in three independent experiments. **e**, Overview of biological support for translation of a subset of ORFs. **f**, Subgroup analyses of ORF biological features demonstrating ORF fractions supported by ectopic V5 translation assays, mass spectrometry or both. **g**, Fractions of ORFs supported by evidence of translation across major epochs in evolutionary time. Evidence of translation shown as the fraction of ORFs with V5 immunoblot signal, endogenous mass spectrometry peptides and the summation of both. CRISPRi, CRISPR interference.

profiling, mass spectrometry and bioinformatic approaches validated at similar rates (Supplementary Fig. 2). Lastly, we detected a protein for ten of 30 ORFs tested by in vitro transcription and translation (Extended Data Fig. 1h). Taken together, experimental evidence of protein translation was obtained for 334/553 (60%) of the ORFs. Translatability was associated with evolutionary conservation (Fig. 1f), with ancient ORFs being more likely to be translated compared to evolutionarily recent ORFs as determined by phylostratigraphy ($P < 0.001$, two-way analysis of variance (ANOVA); Fig. 1g and Supplementary Table 17). Pairwise analysis of combinations of ORF biological features highlighted conservation, size and identification of a mass spectrometry peptide as the strongest predictors of V5-tagged ORF translation (Supplementary Fig. 5). ORFs predicted to encode proteins with <50 amino acids were less likely to yield a detectable protein (Fig. 1f), although this may be explained by the deleterious effect of fusing a 14-amino acid V5 tag to a very small protein. uORFs validated at a higher rate than lncRNA-derived ORFs, largely due to more frequent mass spectrometry evidence for

small uORFs of <50 amino acids, though this may be confounded by the small sample size of uORFs ($n = 18$) (Supplementary Fig. 6).

Since the majority of noncanonical ORFs show evidence of translatability, we next asked whether such translation is associated with biological activity. To address this, we expressed the 553 ORFs in each of four cell lines (MCF7, A549, A375 and HA1E) and then performed RNA expression analysis using the L1000 platform[34] (Fig. 2a), which monitors the expression of 978 messenger RNAs. Ectopic expression of 401 ORFs (73%) yielded a reproducible gene expression consequence, of which 237 induced a high transcriptional activation score (TAS) indicating marked cellular changes[34] (Fig. 2b, Supplementary Fig. 7 and Supplementary Table 18). In comparison, 81% of 2,283 canonical protein-coding genes yielded a gene expression consequence in this assay, indicating that the frequency of biological activity of known genes and unannotated ORFs is similar (Fig. 2b). To exclude the possibility that the observed transcriptional signature was due simply to overexpression of RNA, we mutated translational start sites and repeated the L1000 profiling. In 48 of 51 (94%) cases, the perturbational response was lost when translation was prevented, indicating that the biological effect was indeed mediated by a protein rather than a noncoding RNA (Fig. 2c–f, Extended Data Fig. 2 and Supplementary Tables 19 and 20).

The transcriptional responses observed following ORF expression could conceivably be a consequence of overexpression of the transgene. To address the functional relevance of endogenous expression of these ORFs, we performed CRISPR/Cas9 loss-of-function viability screens in eight cancer cell lines using a guide RNA library targeting the 553 ORFs (Fig. 3a, Supplementary Fig. 8a and Supplementary Table 21). Knockout of 57 of the 553 ORFs (10%) demonstrated growth-inhibitory effect (Fig. 3b,c and Supplementary Tables 22 and 23). Of these, 31 (54%) impaired survival of all eight cell lines whereas 26 (46%) displayed selective dependency (Supplementary Fig. 8b–e).

To compare these data to knockout of canonical proteins, we analyzed the Cancer Dependency Map (www.depmap.org) for the viability effects of 553 randomly selected genes. Among canonical proteins, 17% demonstrated a viability effect in eight randomly chosen cell lines compared to approximately 10% for noncanonical ORFs (Fig. 3d and Supplementary Fig. 8f,g), indicating that the frequency of dependencies between known genes and noncanonical ORFs is approximately in the same order of magnitude. These results were validated in both a secondary CRISPR screen of 147 ORFs (Fig. 3e, Supplementary Fig. 8h,i and Supplementary Tables 24–28) and individually performed CRISPR assays for selected ORFs (Extended Data Fig. 3 and Supplementary Table 29). Analyses for off-target effects of sgRNAs suggested that only five of the 57 CRISPR hits (RP11-138J23.1, RP11-346D14.1, LINC01873, LINC01184 and RP11-277L2.3) were likely to have been confounded by sgRNA cutting at unintended genomic loci (Supplementary Fig. 9 and Supplementary Tables 30 and 31).

Because the viability effects from knockout of noncanonical ORFs could be explained by loss of a regulatory region in the genome rather than the protein itself, we subjected 41 ORFs to dense tiling of sgRNAs across the genomic locus of each ORF. Only 7/41 (17%) genomic regions demonstrated nonspecific viability loss suggestive of a regulatory region of the genome. For 18/41 ORFs (44%), the viability effect mapped exclusively to predicted coding exons or the coding region, as well as adjacent nucleotides in the transcript, which may reflect sites of translational regulation or sgRNAs generating indels that also impact the ORF (Fig. 3f, Extended Data Fig. 4 and Supplementary Table 32). Further, there were 4/41 (10%) ORFs where the viability effect mapped exclusively to the predicted coding region but where a nonoverlapping, neighboring gene also demonstrated a viability effect following knockout (Extended Data Fig. 4a,h).

Interestingly, in several cases a new ORF overlapped with an annotated protein-coding gene but it is the new ORF that best explained the knockout phenotype (Fig. 3g). As examples, we observed that ORFs arising from *CTD-2270L9.4* and *ZBTB11-AS1*, which overlap coding exons of *COG7* and *ZBTB11*, respectively, demonstrated markedly more dramatic viability phenotypes using sgRNAs that target the new ORF compared to adjacent sgRNAs that target only the known, parent ORF (Fig. 3g,h and Extended Data Fig. 4b). These findings were supported by Cancer Dependency Map data in which sgRNAs targeting both new and known ORFs had a more pronounced phenotype than those targeting only the known ORF (Supplementary Fig. 10). For *ZBTB11-AS1*, we validated the specificity of this phenotype through exclusive small interfering RNA knockdown of the *ZBTB11-AS1* transcript, which was rescued by ectopic expression of a siRNA-resistant *ZBTB11-AS1* ORF cDNA but not by a mutant *ZBTB11-AS1* cDNA removing the start codon (Extended Data Fig. 5). Taken together, we conclude that a surprisingly high proportion of noncanonical ORFs exhibit a viability phenotype following knockout and that previous CRISPR vulnerability screens may be confounded by cryptic, new ORFs arising from the same genomic locus.

We next noted that 13 ORFs scored highly in all three high-throughput assays, supporting translation, bioactivity and CRISPR vulnerability (Fig. 4a) and suggesting that they may have particularly important biological roles. Among these, we especially focused on *G029442* (*LA16c-380H5.3* in GENCODE) because its knockout resulted in selective cancer cell killing (one of eight cell lines killed), and it is highly expressed in several human cancer types (Fig. 4b and Extended Data Fig. 6). We subsequently renamed this gene *glycine-rich extracellular protein-1* (*GREP1*) for reasons elucidated below.

To systematically explore the importance of *GREP1* in cancer cell viability, we infected a pool of 486 barcoded human cancer cell lines with a single lentivirus harboring both Cas9 and a guide RNA targeting *GREP1* (Fig. 4c and Methods). Because lentiviral infection rates vary across cell lines, we focused our analysis on the 263 cell lines yielding the highest-quality data (Supplementary Fig. 11a–g, Supplementary Tables 33 and 34 and Methods). *GREP1* knockout resulted in preferential loss of viability in certain cell lineages, most notably breast cancer (Fig. 4d). We validated these pooled screening results with knockout and rescue experiments for *GREP1* in breast and nonbreast cell lines, which confirmed a striking breast cancer viability phenotype that correlated with *GREP1* mRNA expression (Fig. 4e,f, Extended Data Fig. 7a,b and and Supplementary Fig. 11h). Sequencing of the *GREP1* sgRNA genomic loci demonstrated an array of insertions, deletions and substitutions at the expected genomic position, confirming sgRNA target specificity (Supplementary Fig. 12). Finally, *GREP1* expression was higher in human breast cancers compared to normal breast tissue ($P = 1.4 \times 10^{-10}$) (Extended Data Fig. 6c) and was associated with decreased patient survival in breast, but not in colon, cancer patients (Extended Data Fig. 7c,d). Together, these data implicate *GREP1* as a previously unrecognized, prognostic breast cancer vulnerability gene.

To explore the function of GREP1, we noted the presence of a signal localization sequence for extracellular secretion as well as sites of glycosylation documented by mass spectrometry (Fig. 4g and Supplementary Table 35). We confirmed that ectopic expression of a GREP1 fusion protein with a C terminus V5 epitope tag, but not an N-terminal truncation mutant lacking the signal localization sequence, was indeed both secreted and cleaved into a smaller product (Fig. 4h,i, Extended Data Fig. 7e,f and Supplementary Table 36). Analyses of the GREP1 amino acid sequence revealed a conserved, glycine-rich and intrinsically disordered protein (Extended Data Fig. 8a–c), characteristics that resemble certain members of the extracellular matrix[35]. As expected, immunoprecipitation of ectopically expressed GREP1 from cell culture media followed by mass spectrometry revealed strong enrichment for extracellular matrix proteins, including fibronectin and collagen (Extended Data Fig. 8d–k and Supplementary Table 37).

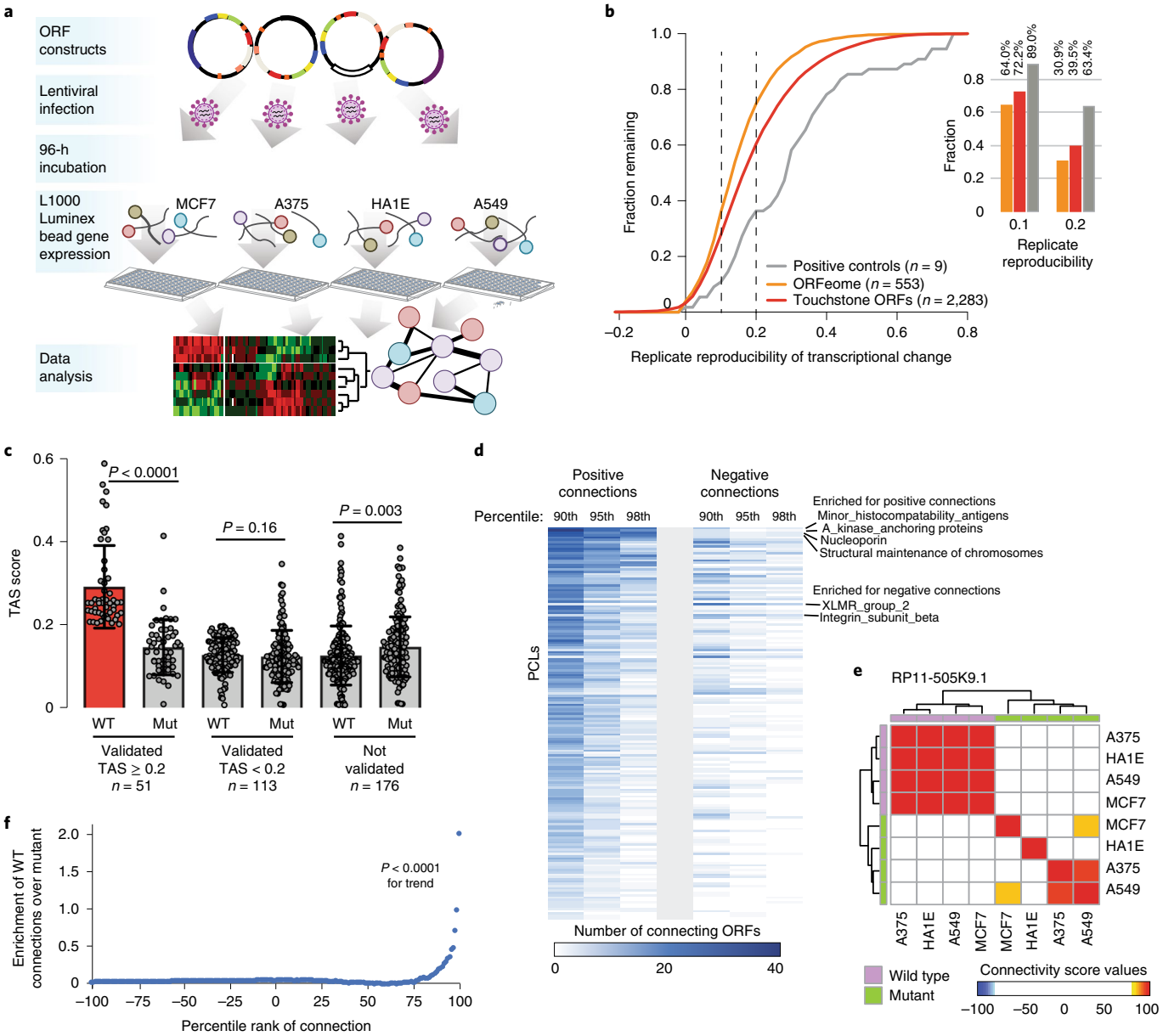


Fig. 2 | Defining bioactive ORFs through gene expression profiling. a, Schematic showing the experimental setup. Briefly, ORFs were individually transduced into four cell lines and expression was profiled 96 h after infection using the L1000 platform. **b**, Fraction of ORFs resulting in transcriptional perturbation when overexpressed in four cell lines (A375, MCF7, HA1E and A549) compared to all profiled known genes and assay-positive controls. Inset: barplot enumerating the percentage of ORFs in each group (top *x* axis) with a transcriptional signature above the indicated reproducibility threshold (bottom *x* axis). **c**, Barplot showing the strength of transcriptional perturbation following expression of the indicated groups of wild-type (WT) or mutant ORF (Mut) constructs. Numbers of pairs of wild-type or mutant ORF data are indicated (*n*). *P* values were calculated by two-sided Wilcoxon test; error bars represent s.d. **d**, Heatmap showing the number of ORFs demonstrating positive or negative connections with individual perturbational classes (PCLs) at the indicated percentile rank. **e**, An example of RP11-505K9.1 showing the high concordance of connectivity signatures when the wild-type ORF is expressed compared to the ORF with mutated translational start sites. **f**, Bland–Altman analysis demonstrating enrichment of high-ranking connectivity values following expression of wild-type ORFs compared to mutant (both $n = 19{,}012$). *P* value was calculated by two-sided Wilcoxon test.

To establish the cellular consequence of GREP1 expression, we examined the impact of *GREP1* knockout and overexpression on other secreted proteins by testing a panel of 102 secreted proteins using antibody arrays across three cell lines (Fig. 4j). The metabolic cytokine GDF15 (refs. [36,37]) demonstrated the most dramatic change, with *GREP1* knockout resulting in decreased GDF15 levels and GREP1 overexpression resulting in increased GDF15 levels (Fig. 4k,l and Extended Data Fig. 9a,b). Inducing nonspecific cellular stress through pharmacological treatment with toxic compounds did not increase GDF15 levels, indicating specificity (Supplementary Fig. 13a,b). In addition, impairment of GREP1 secretion through deletion of the signal localization sequence, but not mutation of the GREP1 glycosylation sites, prevented increase in GDF15 secretion (Supplementary Fig. 13c,d). In human cancers, expression of *GREP1* and *GDF15* was correlated across multiple tumor types in the The Cancer Genome Atlas (TCGA) database (Extended Data Fig. 9c,d). To determine whether GDF15 is functionally important in the requirement by cancer cells of GREP1 for

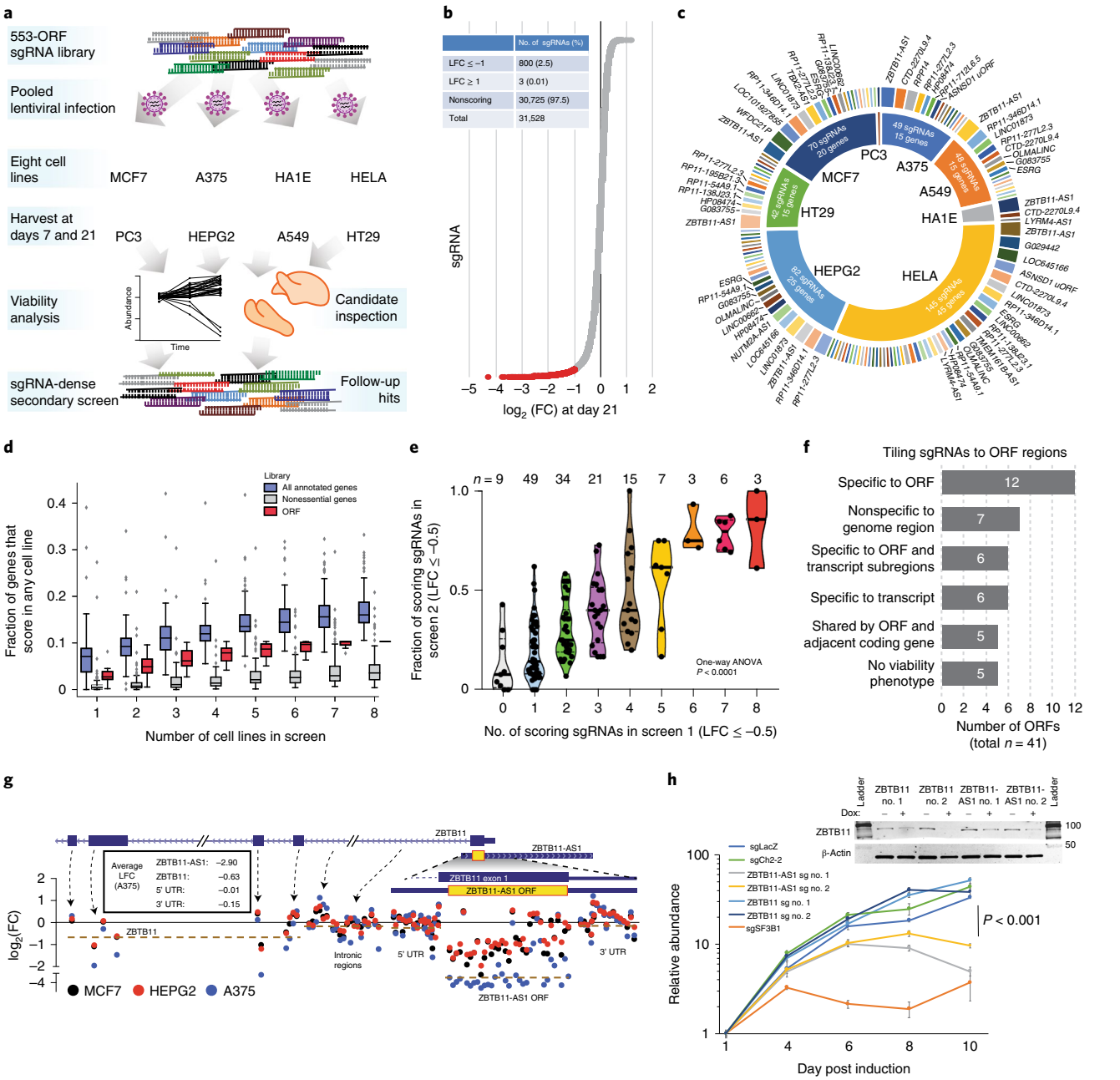
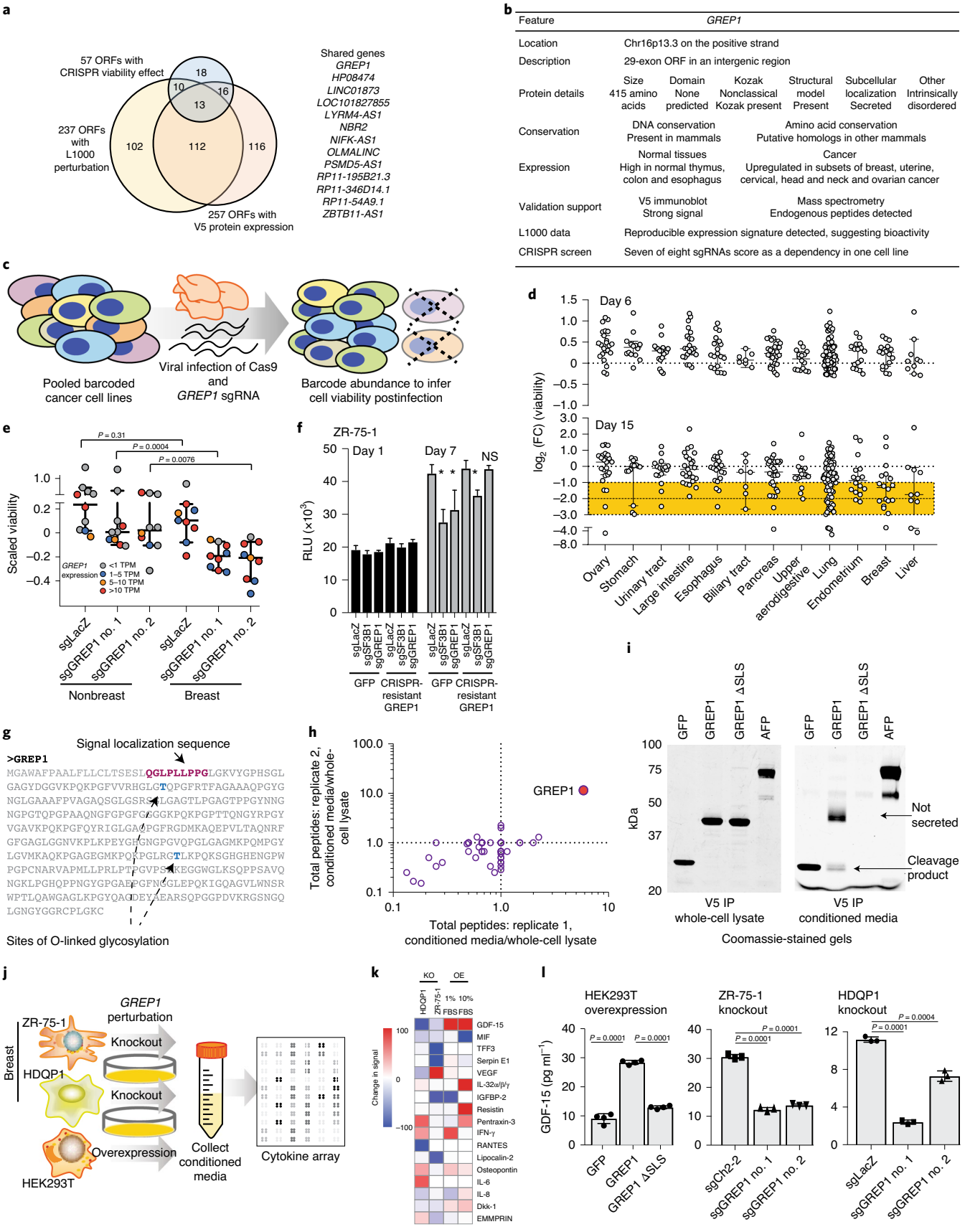


Fig. 3 | CRISPR screening to identify unknown ORFs implicated in cancer cell viability. a, Schematic showing the experimental design, including a primary screen in eight cancer cell lines and a secondary screen in three. **b**, Distribution of sgRNA depletion at day 21 following lentiviral infection in the CRISPR screen across eight cell lines; 2.5% of sgRNAs were identified as depleted in a particular cell line with $\log_2$(FC) ≤ −1. **c**, Distribution of nominated ORFs. For each cell line (inner circle), the numbers of sgRNAs with $\log_2$(FC) ≤ −1 (middle circle) and nominated genes (outer circle) are shown. The outer circle shows the ORFs nominated in that cell line, ranked by the number of supporting sgRNAs. The thickness of the outer circle boxes reflects the number of sgRNAs supporting nomination of corresponding ORFs. Only ORFs nominated with ≥2 sgRNAs are shown. **d**, Boxplot showing the fractions of annotated genes, new ORF genes and RNA interference-defined nonessential genes scoring as vulnerability genes in the indicated number of cell lines. Each data point represents a unique cell line. The cell lines for ORF genes represent those used in this study. For annotated genes, randomly selected cell lines from the Cancer Dependency Map were used. Box plots represent median with interquartile ranges (25–75%); whiskers extend to the last data point up to 1.5× interquartile distance from the box, with individual data points shown beyond this range. **e**, Correlation between the number of sgRNAs producing a viability phenotype for a given ORF in the primary screen and the fraction of sgRNAs producing a viability phenotype in the secondary screen. The number of ORFs included in each group is indicated. *P* value was calculated by one-way ANOVA. **f**, Barplot showing the number of ORFs with each category of viability phenotype in the tiling sgRNA CRISPR screens. **g**, An example of *ZBTB11* and *ZBTB11-AS1* for tiling CRISPR data, showing enhanced cell killing when the *ZBTB11-AS1* ORF is knocked out. Each data point represents one sgRNA. Data points are color coded for the indicated cell lines. **h**, Individual CRISPR knockout experiments in a doxycycline (DOX)-inducible Cas9 HeLa cell line using two sgRNAs either targeting *ZBTB11* exclusively or targeting both *ZBTB11-AS1* ORF and *ZBTB11*. The line plot shows cell viability measured by cellular ATP following induction of Cas9 activity with 2 μg ml^{-1} doxycycline. sgLacZ and sgCh2-2 are noncutting and cutting negative controls, respectively, and sgSF3B1 is a pan-lethal positive control. *n* = 6 technical replicates for each data point, with two independent experiments performed. Inset: immunoblot showing ZBTB11 protein abundance following induction of Cas9. *P* value was calculated by two-tailed Student's *t*-test. Error bars represent s.d. LFC, $\log_2$(FC).



survival, we tested the effect of *GREP1* knockout in the presence and absence of recombinant GDF15. Remarkably, supplementation of recombinant human GDF15 rescued the loss of viability caused by *GREP1* loss of function (Extended Data Fig. 9e,f). The fact that GDF15 only partially rescues *GREP1* knockout in some cell lines suggests that there may be additional mechanisms downstream of

Fig. 4 | Characterization of *GREP1* as a cancer dependency gene in breast cancer. a, Nomination of candidate ORFs with evidence for protein translation, gene expression effect and CRISPR phenotype. **b**, Table summarizing characteristics of the *GREP1* gene. **c**, Schematic showing overview of pooled CRISPR screening. **d**, $\log_2$(FC) abundance of cancer cell lines at days 6 and 15 following pooled CRISPR screening. Cell lineages are ranked based on median $\log_2$(FC) at day 15. Each data point represents a unique cell line. **e**, Individual CRISPR validation experiments for *GREP1* in a panel of nonbreast ($n=10$) and breast ($n=9$) cell lines. Data are scaled so that 0 reflects the sgCh2-2 negative cutting control and −1 reflects the degree of viability loss from the sg*SF3B1* positive control. Data were obtained 7 days after lentiviral infection. *P* values were calculated by two-tailed Mann–Whitney test. **f**, Rescue of the CRISPR phenotype with overexpression of a CRISPR-resistant GREP1 construct and not GFP. $^*P<0.05$. For GFP cells: sgLacZ versus sgSF3B1, $P=0.0005$; sgLacZ versus sgGREP1, $P=0.013$; for GREP1 cells: sgLacZ versus sgSF3B1, $P=0.0005$; sgLacZ versus sgGREP1, $P=0.08$. *P* values were calculated by two-tailed Student's *t*-test. $n=4$ technical replicates per data point, with two independent experiments performed. **g**, GREP1 amino acid sequence with the signal localization sequence and sites of glycosylation indicated. **h**, Immunoprecipitation followed by mass spectrometry of HEK293T-conditioned media and whole-cell lysate following ectopic expression of a pool of V5-tagged ORFs. The *x* and *y* axes represent the total number of mass spectrometry peptides detected in two independent experiments (replicate nos. 1 and 2). **i**, Expression of V5-tagged GREP1 or a truncated GREP1 lacking the N-terminal signal localization sequence in HEK293T cells. Cell lysates or conditioned media were subjected to V5 immunoprecipitation (IP) and protein was then visualized by Coomassie staining. Two independent biological experiments were performed. **j**, Experimental design for profiling of secreted cytokine following *GREP1* perturbation. **k**, Heatmap showing individual cell line changes in cytokine abundance following *GREP1* perturbation. Cytokines are ranked according to the average of the absolute value of signal change for each cell line. **l**, Validation of *GDF15* modulation following *GREP1* perturbation by ELISA in the indicated cell lines. $n=4$ (HEK293T) or $n=3$ (ZR-75-1, HDQP1) technical replicates per sample, with either two (HDQP1) or three (ZR-75-1, HEK293T) independent experiments performed. *P* values were calculated by two-tailed Student's *t*-test. All error bars represent s.d. NS, not significant; OE, overexpression; KO, knockout; RLU, relative light unit.

GREP1 that regulate cell survival (Extended Data Fig. 9g). While GDF15 has previously been implicated in a number of cancer phenotypes, including chemoresistance[38,39], immune evasion[40], cellular survival and invasiveness[41,42], its regulation by GREP1, which itself is a cancer dependency, is entirely new.

Despite the fact that the human genome was sequenced 18 years ago, the precise number of protein-coding genes in it remains a point of controversy. Our sampling of >550 uncharacterized ORFs provides experimental evidence that a substantial proportion of such ORFs may be functional (Supplementary Fig. 14). Importantly, we establish that approximately 10% of the ORFs in our dataset are required for the survival of cancer cells, a rate only about half that observed for known, canonical proteins. Although our dataset represents a curated list of ORFs rather than a random sampling of all possible ORFs, these experiments suggest that further investigations of unannotated ORFs in cancer and other disease states will probably yield new insights. Since computational estimates of such ORFs now exceed 5,000 (ref. [43]), our data suggest that a substantial number of those predicted ORFs may indeed encode functional proteins.

Consistent with this conclusion, a recent report by Chen et al. similarly suggests functional roles for a substantial fraction of noncanonical ORFs[44]. While a head-to-head comparison of the two datasets is difficult because they utilize different cell lines for functional analyses, the Chen et al. dataset identifies the existence of additional functional long intergenic noncoding RNA-derived ORFs beyond those identified in our dataset (Supplementary Table 38). This result suggests that the functional ORFs discovered in our study do not represent the entirety of those encoded by the human genome—more functional ORFs probably remain to be discovered. Of note, whereas our study focused primarily on lncRNA-derived ORFs, Chen et al. also expand upon the potential functional importance of a subset of uORFs[44]. While the precise number of noncanonical ORFs encoded by the human genome remains to be determined, our work suggests that future systematic interrogation of noncanonical proteins is likely to yield a rich source of previously unrecognized proteins with key roles in development and disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-00806-2.

Received: 18 February 2020; Accepted: 16 December 2020;
Published online: 28 January 2021

References

1. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25**, 232–234 (2000).
2. Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nat. Genet.* **7**, 345–346 (1994).
3. Liang, F. et al. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25**, 239–240 (2000).
4. 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).
5. Ingolia, N. T. et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8**, 1365–1379 (2014).
6. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).
7. Pertea, M. et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
8. van Heesch, S. et al. The translational landscape of the human heart. *Cell* **178**, 242–260 (2019).
9. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
10. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4**, e1000176 (2008).
11. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
12. Mouse Genome Sequencing Consortium Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
13. Mudge, J. M. et al. Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res.* **29**, 2073–2087 (2019).
14. Banfai, B. et al. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**, 1646–1657 (2012).
15. Jungreis, I. et al. Nearly all new protein-coding predictions in the CHESS database are not protein-coding. Preprint at *bioRxiv* https://doi.org/10.1101/360602 (2018).
16. Bazzini, A. A. et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).
17. Branca, R. M. et al. HiRIEF LC–MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11**, 59–62 (2014).
18. Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
19. Calviello, L. et al. Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).

20. Gao, X. et al. Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods* **12**, 147–153 (2015).
21. Gascoigne, D. K. et al. Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* **28**, 3042–3050 (2012).
22. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
23. Kim, M. S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
24. Koch, A. et al. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* **14**, 2688–2698 (2014).
25. Ma, J. et al. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.* **13**, 1757–1765 (2014).
26. Mackowiak, S. D. et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* **16**, 179 (2015).
27. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).
28. Schwaid, A. G. et al. Chemoproteomic discovery of cysteine-containing human short open reading frames. *J. Am. Chem. Soc.* **135**, 16750–16753 (2013).
29. Slavoff, S. A. et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59–64 (2013).
30. Sun, H. et al. Integration of mass spectrometry and RNA-seq data to confirm human ab initio predicted genes and lncRNAs. *Proteomics* **14**, 2760–2768 (2014).
31. Zhang, C. et al. Systematic analysis of missing proteins provides clues to help define all of the protein-coding genes on human chromosome 1. *J. Proteome Res.* **13**, 114–125 (2014).
32. Vanderperre, B. et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS ONE* **8**, e70698 (2013).
33. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
34. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
35. Nassa, M. et al. Analysis of human collagen sequences. *Bioinformation* **8**, 26–33 (2012).
36. Breit, S. N., Tsai, V. W. & Brown, D. A. Targeting obesity and cachexia: Identification of the GFRAL receptor-MIC-1/GDF15 pathway. *Trends Mol. Med.* **23**, 1065–1067 (2017).
37. Mullican, S. E. & Rangwala, S. M. Uniting GDF15 and GFRAL: therapeutic opportunities in obesity and beyond. *Trends Endocrinol. Metab.* **29**, 560–570 (2018).
38. Baroni, M. et al. Distinct response to GDF15 knockdown in pediatric and adult glioblastoma cell lines. *J. Neurooncol.* **139**, 51–60 (2018).
39. Huang, C. Y. et al. Molecular alterations in prostate carcinomas that associate with in vivo exposure to chemotherapy: identification of a cytoprotective mechanism involving growth differentiation factor 15. *Clin. Cancer Res.* **13**, 5825–5833 (2007).
40. Ratnam, N. M. et al. NF-kappaB regulates GDF-15 to suppress macrophage surveillance during early tumor development. *J. Clin. Invest.* **127**, 3796–3809 (2017).
41. Corre, J. et al. Bioactivity and prognostic significance of growth differentiation factor GDF15 secreted by bone marrow mesenchymal stem cells in multiple myeloma. *Cancer Res.* **72**, 1395–1406 (2012).
42. Peake, B. F., Eze, S. M., Yang, L., Castellino, R. C. & Nahta, R. Growth differentiation factor 15 mediates epithelial mesenchymal transition and invasion of breast cancers through IGF-1R-FoxM1 signaling. *Oncotarget* **8**, 94393–94406 (2017).
43. Martinez, T. F. et al. Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468 (2020).
44. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).

Methods

Data statement. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines and reagents. All parental cell lines were obtained from the American Type Culture Collection (ATCC). Cas9-derived cell lines were obtained from the Broad Institute. Cell lines were maintained using standard media and conditions. In brief, cell lines derived from ZR-75-1, HCC1806, HCC1954, HCC202, T47D, HT-29, HCC15, KYSE410, KYSE510, SNU503, SW837, HCT116, AU565, CAMA-1 and MDA-MB-231 were maintained in RPMI 1640 (Invitrogen) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin in a 5% CO_2 cell culture incubator at 37 °C. Cell lines derived from HDQP1, BT-474, JIMT1, A375, A549, MIAPACA2, MCF7, HEK293T and MDA-MB-468 were maintained in DMEM supplemented with 10% FBS (Invitrogen) and 1% penicillin/streptomycin (Invitrogen) in a 5% CO_2 cell culture incubator. Green fluorescent protein (GFP)-positive Cas9-derived cell lines were drug selected using 2 μg ml^{-1} blasticidin.

For stable knockout cell lines, ZR-75-1 Cas9- and HDQP1 Cas9-expressing cells were infected with lentivirus for the indicated sgRNAs that had been cloned into LentiGuide-Puro plasmid (plasmid 52963, Addgene) with 4 μg ml^{-1} polybrene. Sixteen hours after transduction, cells were selected with cell culture medium containing 2 μg ml^{-1} puromycin. Cells were maintained in puromycin-containing medium for 72 h before transitioning back to standard culture medium. Stable GREP1-overexpressing cell lines were generated in ZR-75-1 and CAMA-1 cells by infection with a sgRNA-resistant *GREP1* cDNA construct and selecting with 350 μg ml^{-1} hygromycin for 96 h, before transitioning back to standard culture medium.

RNA isolation, cDNA synthesis and quantitative PCR experiments. Total RNA was isolated using Qiazol and an miRNeasy Kit (Qiagen) with DNase I digestion according to the manufacturer's instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies). cDNA was synthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen). Quantitative real-time PCR (qPCR) was performed using Power SYBR Green Mastermix (Applied Biosystems) on a Thermo QStudio FLX Real-Time PCR System (Thermo Fisher Scientific). *GAPDH* was used as the housekeeping control gene. The relative quantity of the target gene was completed for each sample using the ΔΔCt method by comparison of the mean cycle threshold (Ct) of the gene to the average Ct of the geometric mean of the indicated housekeeping genes. The primer sequences are listed below:

GREP1 3′ UTR-forward: AGCCTCCAAATGGCTATGGAC
GREP1 3′ UTR-reverse: CTCGAGGCCACCATTAAAAC
GREP1 ORF-forward: CTGGATATCCGGCTGGAGATG
GREP1 ORF-reverse: ATTGCTGCCTCTCTTCACGTC
GAPDH-forward: TGCACCACCAACTGCTTAGC
GAPDH-reverse: GGCATGGACTGTGGTCATGAG
Beta-actin forward: AAGGCCAACCGCGAGAAG
Beta-actin reverse: ACAGCCTGGATAGCAACGTACA
Fibronectin forward: GAGAAAATGGCCAGATGATGA
Fibronectin reverse: AATGGCACCGAGATATTCCTT
Emilin2 forward: AACAAAGTGCTGGTGAACGAC
Emilin2 reverse: CTCTCCTGTACCCAGCGGTAT
ZBTB11-AS1 forward: CCGTTTTTACGTTTGAGACTCC
ZBTB11-AS1 reverse: ATGTAAATGGGCTGTCTCTGGT
ZBTB11 forward: GGAACGGGTGTGTGAAAAAT
ZBTB11 reverse: CAGCCCAAGCTACTCCACAT
HP08474 forward: GTGTAAAGAGGTCCTGGGACAG
HP08474 reverse: GCACTCCAGTCTAGACGACACA
RP11-54A9.1 forward: TTGGTGAGATGTTCCTTGAGC
RP11-54A9.1 reverse: CTCCACTTCACTGTCGGTCTC
G083755 forward: ATCCCATCTGAGTGCTTACCAA
G083755 reverse: CATGCATAATCTCCTTCCCTGC
OLMALINC forward: AGGAACATCTTGCCAATTTCA
OLMALINC reverse: TGTGGATCTTCAGTTGCTTCA
CTD-2270L9.4 forward: AGTCGTTGGCCGTTACCATA
CTD-2270L9.4 reverse: CTTCCCAGGCTCAAGCAAT
ASNSD1 uORF forward: ACAATTCGACCCCACACAAG
ASNSD1 uORF reverse: GGTTAGAAAGTTCATCCACCACA
RP11-277L2.3 forward: CTACGTGGGGCTGGAAATAC
RP11-277L2.3 reverse: CCCTTCCCAGTTCTCTGACC
GREP1_sgRNA1_amplicon_CRISPRSeq_Forward: GGCCTTAACCCTTTCTCTCCT
GREP1_sgRNA1_amplicon_CRISPRSeq_Reverse: ATCAAGGCGGGGTATGAATG
GREP1_sgRNA2_amplicon_CRISPRSeq_Forward: TTCTGGGGTGGATCTGAGTT
GREP1_sgRNA2_amplicon_CRISPRSeq_Reverse: CCCATTCCCATTCCCTAATC

Selection process for candidate ORFs. Candidate ORFs were collated via manual curation from 25 published studies and one in-house analysis of ribosomal profiling data (Z. Ji, personal communication). Published studies are listed in Supplementary Table 1. Data types included were 14 studies with mass spectrometry data, six with ribosomal profiling data, four with computational ORF predictions and one with both mass spectrometry and ribosomal profiling data. In total, there were 9,918 candidate ORFs among which 4,433 unique Ensembl transcripts were represented.

We integrated the ORF nominations with mRNA expression data across the Cancer Cell Line Encyclopedia (CCLE). There were 6,305 ORFs arising from transcripts with expression of at least one transcript per million (TPM), with at least one cell line having >10 TPM. Because candidates nominated only from computational predictions were unlikely to have any ribosome profiling or mass spectrometry correlate (Supplementary Figs. 1 and 2), we considered only the 3,825 candidates that had either literature peptide support in mass spectrometry, ribosome profiling data or both. Among this list, there were 917 annotated pseudogenes and 513 variants of known coding proteins (including N-terminal extension ORFs, ORFs of known proteins with new predicted exons and alternative ORFs located entirely within the genomic nucleotides of an annotated protein); these were removed from consideration. For the remaining 2,395 ORFs arising from a putative noncoding RNA, we recomputed PhastCons scores, ribosome read abundance, PhyloCSF scores and protein domain scores as indicated below (Supplementary Tables 2 and 3). A total of 553 high-priority ORFs were manually curated as candidates according to the criteria described below. See Supplementary Fig. 1 for an overview.

ORFeome library inclusion criteria. To be selected for the ORFeome library, an ORF had to exhibit at least one of the characteristics detailed below (Supplementary Tables 3 and 4); among 2,395 ORFs, 669 exhibited two or more of these features. Following manual inspection to eliminate overlapping candidates (for example, isoforms or variants of the same ORF), we selected the longest ORF on each transcript for 353 of the 669 examples (53%). Of the 1,726 ORFs exhibiting only one feature, we eliminated overlapping candidates and manually inspected 1,018 examples to select 200 ORFs for inclusion in the ORFeome library. Details of these features are now described.

DNA conservation. An ORF was considered to have high DNA conservation if the average PhastCons score (v.hg19_20110909) for 100 placental mammals was ≥0.20 for the entire ORF; 677 ORFs meeting this metric were manually inspected and filtered for overlapping or multiple predictions on the same mRNA. In total, 172 of the 677 ORFs (26%) were included in the ORFeome library.

Codon substitution rate. ORFs were stratified if they had a codon PhyloCSF decibans score (29-mammal alignment) of ≥5.0 averaged across the whole ORF; 74 ORFs meeting this metric were manually inspected and filtered for overlapping or multiple predictions on the same mRNA. Nineteen of 74 ORFs (26%) were included in the ORFeome library.

High read coverage. Ribosome profiling read abundance data for ORFs identified by Ji et al.[6] were used, along with in-house analyses (Z. Ji, personal communication). ORFs were stratified if they had a read/length ratio of ≥1.0 in available ribosomal profiling data; 2,136 ORFs meeting this metric were manually inspected and filtered for overlapping predictions or multiple predictions on the same mRNA, and 203 ORFs (9%) were included in the ORFeome library.

Protein domain structure. We utilized the Pfam web server (http://pfam.xfam.org/search#tabview=tab1) to identify peptide sequences harboring a putative Pfam domain (including both Pfam-A and Pfam-B), and used the default cutoff e-value <1. In addition, ORF amino acid sequences were also input into the NCBI Conserved Domain finder (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) with default settings to identify putative domains. ORFs with domain structures scoring at an e-value confidence score of <0.01 were considered; 195 ORFs meeting these criteria were then manually inspected and filtered for overlapping predictions, and 88 of 195 ORFs (45%) were included in the ORFeome.

Multiple overlapping ORF predictions. Published ORF predictions from 25 large datasets were integrated[5,6,16,18–22,25–29,31,33] and queried for overlapping ORF predictions with at least two publications supporting their existence (Supplementary Tables 1–3). Of 643 candidates, we manually inspected and removed overlapping nominations or multiple isoforms of one gene. We included 227 of 643 ORFs (35%) in the ORFeome library.

Cancer expression. We analyzed a dataset from Iyer et al.[22] that identified 980 transcripts of unknown coding potential defined by a coding-potential assessment tool coding score of >0.5 and a statistical enrichment for human cancer tissue expression compared to benign tissue ($n = 707$) or cancer lineage expression compared to other cancer types if no benign tissue was available ($n = 273$). Of these, 437 (45%) exhibited an expression level of ≥1 TPM in one of the cell lines used for CRISPR knockout studies.

Lineage association. ORFs were searched in the NIH Roadmap Epigenome Project data[45], which transcriptionally profiled human embryonic stem cells before and after differentiation into mesenchymal stem cells, neural progenitor cells, trophoblast-like cells or mesoendoderm; 243 transcripts were nominated, of which 123 (50%) harbored an ORF nomination and were included in the ORFeome.

Upstream and downstream ORFs. We used candidates from Ji et al.[6] and considered conserved upstream and downstream ORFs between mouse and human, as defined by an interspecies alignment with an e-value of <0.0001. We evaluated ORFs with all of the following attributes: (1) conservation ratio of the number of nonsynonymous substitutions per nonsynonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) <0.5; (2) ORF length ≥25 amino acids; (3) an ATG start site; (4) a predicted higher translational efficiency compared to the annotated protein residing on the same mRNA; and (5) the ORF was nonoverlapping with the annotated ORF; 49 dORFs and 195 uORFs met these criteria, and these were then manually reviewed to select candidates included in the ORFeome.

Additional subanalyses performed on the 553 ORF candidates selected are now shown.

Murine homolog. Murine homologs were defined by the Slncky program[46].

Cancer function association. ORFs were searched in the PubMed database for associations with the word 'cancer' and screened for those studies implicating potential roles in cancer.

Predicted ORF CRISPR phenotype. Data from a CRISPR interference screen of lncRNAs were employed[47]. Of 492 lncRNA hits nominated in that study, there were 312 hits with GENCODE identifiers that could be further evaluated. Of those 312 there were 292 unique GENCODE identifiers, which were manually reviewed; 52 GENCODE identifiers overlapping ORFs in this ORFeome are indicated.

Signal peptide. All ORFeome ORFs were analyzed by SignalP v.4.1 using standard default settings[48], and mean difference was divided by the overall s.d. of ≥0.45 to nominate ORFs with a classical signal localization sequence.

Structural modeling. All ORFeome ORFs ≥40 amino acids were analyzed by Phyre2 structural domain prediction software using default settings[49]. To distinguish ORFs enriched for structural models, we generated a random amino acid sequence library of 500 random 150-mer amino acid sequences with a methionine start codon. We then derived a structural model score: (percentage ORF alignment to the structural model) × (percentage confidence of the model). A structural model score of 0.175 was used to maximally differentiate ORFeome ORFs from random amino acid sequences; 145 ORFs were classified as having a robust structural prediction score.

Overall ORF confidence score. Each criterion described above, in addition to mass spectrometry peptide evidence (see below), was given a binary score of 1 if the criterion was met by the ORF, or 0 if not met by the ORF. The ORF confidence score is the summation of these binary scores.

Identification of ORFs in proteomics datasets. A fasta database containing the amino acid sequences of the 553 ORFs was appended to a reference protein database (UCSC, RefSeq) and used to search peptide mass spectra of datasets acquired or analyzed in our laboratory. These datasets predominantly comprised studies conducted by the Clinical Proteomics Tumor Analysis Consortium (CPTAC) (Supplementary Table 14). Raw mass spectrometry data were analyzed in Spectrum Mill MS Proteomics Workbench v.6.0 (Agilent Technologies) employing search parameters specific for each project. Detailed descriptions of search parameters, including enzyme definition and specificity or the number of types of variable modifications included in the database search, can be found in the corresponding publications (Supplementary Table 14). Peptide–spectrum matches (PSMs) to the ORF database were identified by automatically parsing through database search results generated by Spectrum Mill Software using an in-house-developed R script. Only PSMs validated by target-decoy-based false discovery rate (FDR) estimation were used for subsequent analysis. To further minimize the possibility of false-positive identifications, we required a minimal Spectrum Mill PSM score of 8, which roughly translates to a minimum of eight identified fragment ions in the tandem mass spectrometry (MS/MS) spectrum. All PSMs meeting the criteria described above are listed in Supplementary Table 14.

Phylostratigraphy analysis. All ORFs with ≥40 amino acids were analyzed as described previously[50,51] using TimeTree[52] (http://www.timetree.org) to identify evolutionary strata. Using a BLASTP e-value threshold of 10^{-3} and a maximum number of 200,000 hits, we identified the phylostratum in which each ORF appeared. For clarity, we aggregated results into the following evolutionary eras: invertebrates (phylostrata 1–9, including cellular organisms through Craniata, ~540 Ma); vertebrates (phylostrata 10–17, including Vertebrata through Amniota (312 Ma)); mammals (phylostrata 18–22, including Mammalia through Euarchontoglires (95 Ma)); primates (phylostrata 23–27, including primates through Hominoidae (20 Ma)); great apes (phylostrata 28–30, including Hominidae through *Homo*); and humans (phylostratum 31, including *Homo sapiens*).

Generation of the ORFeome library. Initial prototype plasmids were generated in the pLX_307 vector backbone designed for previous ORF studies[53], obtained from the Broad Institute Genomic Perturbation Platform, by PCR amplification from cell line cDNA (HeLa, HEK293T, K562 or MCF7). PCR products were gel purified (Qiagen), cloned into the nondirectional Gateway PCR8 vector (Invitrogen) as an entry vector and shuttled to pLX_307 using LR clonase II (Invitrogen) according to the manufacturer's instructions. pLX_307 is a Gateway-compatible expression vector where *EF1a* is the promoter of the ORF and *SV40* is the puromycin resistance gene (details available at https://portals.broadinstitute.org/gpp/public/resources/protocols). Following technical optimization of the insert sequence to include a barcode sequence following the V5 tag, the final ORF construct design is as follows:

vector backbone → ORF sequence lacking stop codon → C terminus V5 sequence (GGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACG) → triple stop codon (TAGTAATGA) → P1 primer site (TCTTGTGGAAAGGACGA) → barcode sequence → AC (linker sequence) → vector backbone.

Following the ORF sequence, each construct therefore had the additional sequence:

GGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGTAGTA ATGATCTTGTGGAAAGGACGA_BARCODE_AC

The ORFeome library was then generated via insert synthesis and cloning of unique plasmid inserts consisting of unique barcodes (Supplementary Table 2), by a commercial vendor (GenScript), in arrayed barcoded tube format. Each plasmid therefore had a barcode sequence flanked by common PCR primer pair for amplification of a 233-base pair (bp) amplicon, where the sense primer was located in the ORF insert and the antisense primer in the plasmid backbone as follows:

P1 sense primer: TCTTGTGGAAAGGACGA
P2 antisense primer: TTAAAGCAGCGTATCCACATAGCGT

Generation of paired mutant ORFs. The 85 mutant constructs employing an identical plasmid insert construct, as detailed above, were utilized with the following modifications: the putative ORF start codon was mutated to GCG (encoding alanine) and all internal in-frame ATG codons (encoding methionine) were mutated to GCG to reduce the likelihood of internal initiation of translation. Constructs were generated via commercial gene synthesis (GenScript).

In-cell immunoblotting. HEK293T cells were plated at a density of 20,000 per well in a 96-well black plate format, to minimize autofluorescence. Six to eight hours after plating, cells were transiently transfected with 0.1 µg of an individual plasmid with Fugene HD reagent (Promega). After 48 h, cell culture medium was removed and cells were fixed for 20 min with 150 µl of 3.7% formaldehyde solution in 1× PBS at room temperature with no shaking. Fixing buffer was removed and cells were washed five times with 200 µl of PBS containing 0.1% Triton X-100 (Sigma-Aldrich) for permeabilization. Following this, cells were blocked with 150 µl of Odyssey Blocking Buffer (LI-COR) for 90 min at room temperature on a plate shaker. Cells were then treated with anti-V5 antibody (1:200 concentration) in Odyssey Blocking Buffer or no-antibody control wells. Cells were incubated with the primary antibody overnight at 4 °C. The next day, the primary antibody was removed and cells were washed five times with 200 µl of PBS containing 0.1% Triton X-100 as above. Then, 50 µl of secondary antibody was applied at 1:1,000 dilution and samples were incubated for 1 h with gentle shaking and protection from light. Next, wells were washed five times with 200 µl of PBS containing 0.1% Tween 20 (Sigma-Aldrich). After a final wash, plates were blotted on tissue paper to remove excess wash buffer and immediately scanned on a LI-COR Odyssey system using the 800-nm light channel to image and quantify anti-V5 abundance.

Analysis of in-cell immunoblot data. First, a preliminary dilution series was performed with decreasing amounts of transfected plasmid and decreasing numbers of HEK293T cells plated per well (Extended Data Fig. 1). This was performed for two high-expressing plasmids that were verified by immunoblot (enhanced GFP (eGFP) and LINC00116), and one low-expressing verified plasmid (RP11-539I5.1). Using eGFP and RP11-539I5.1, we defined a dynamic range for the assay (Extended Data Fig. 1) by normalizing the V5 800-nm light signal to the plate background. This defined a threshold above which signal was reproducibly detected, even in low-expressing plasmids when transfected into 1,000 plated HEK293T cells.

Then, for the full ORFeome library, all plasmids were run in biological triplicate on three unique 96-well plates for in-cell immunoblot analysis. Each plate was normalized by median centering raw 800-nm signals within each plate to minimize variance in plate background. Normalized 800-nm signals were then averaged across replicates. Plasmids with averaged signal above the previously defined threshold based on RP11-539I5.1 expression were considered to have generated a protein by V5 tag detection.

In vitro transcription/translation. Thirty ORFs were selected at random from the ORFeome library for synthesis of the ORF insert lacking a V5 tag and containing a 5′ T7 promoter sequence. This tag-free insert was then cloned into pUC57 plasmid. Linearized purified plasmid (1 mcg) was then subjected to wheat germ extract in vitro transcription/translation, employing the nonradioactive Transcend tRNA

system according to the manufacturer's instructions (Promega). From 50 µl of the reaction volume, 10 µl was then heat denatured in the presence of DTT and protein bands were detected by SDS–polyacrylamide gel electrophoresis (SDS–PAGE) using Tris-glycine 10–20% gel (Thermo Fisher Scientific).

Immunoblot analysis. Cells were lysed in RIPA lysis buffer (Sigma-Aldrich) and then allowed to homogenize on ice for 30 min. Cell debris was removed by centrifugation for 15 min at 13,200 r.p.m. and the debris pellet was discarded. HALT protease inhibitor (1×, Thermo Fisher Scientific) was added to lysate supernatants. Protein abundance was quantified by the bicinchoninic acid method using the bovine-specific albumin standard curve for normalization of protein abundance. Aliquots of each protein extract were boiled in LDS sample buffer, size fractionated by SDS–PAGE at 4 °C by Tris-glycine 10–20% gels and transferred onto nitrocellulose membranes with precast gels via the iBlot-2 system (Thermo Fisher Scientific). The membrane was then incubated at room temperature for 1–2 h in LI-COR Odyssey blocking buffer, and at 4 °C with the appropriate antibody overnight. Following incubation, the blot was washed four times with 1× TBS with 0.1% Tween 20, incubated with fluorophore-specific IRDye secondary antibodies (LI-COR) and imaged on a LI-COR Odyssey machine.

For conditioned media immunoblots, conditioned medium of GFP- or GREP1-expressing HEK293T cells was concentrated by a factor of five using 3-kDa exclusion filter tubes (Millipore), then 1× HALT protease inhibitor was added to the samples. Samples were maintained at 4 °C and not frozen, to preserve protein fidelity. Immunoblots were then performed as detailed above. Uncropped and unprocessed scans of relevant immunoblots are included in the Source Data.

Antibodies used.

Antibody	Species	Monoclonal/ polyclonal	Dilution	Catalog no.	Vendor	Conditions
V5 (D3H8Q)	Rabbit	Monoclonal	1:2,000	13202S	Cell Signaling Technology	4 °C overnight
ZBTB11	Rabbit	Polyclonal	1:1,000	A303-240A-M	Bethyl Laboratories	4 °C overnight
Beta-Actin	Mouse	Monoclonal	1:4,000	A5316	Sigma-Aldrich	4 °C overnight
Goat anti-mouse secondary	Goat	NA	1:5,000	926-32210	LI-COR	20 °C for 1h
Goat anti-rabbit secondary	Goat	NA	1:5,000	926-68021	LI-COR	20 °C for 1h
Goat anti-rabbit HRP	Goat	NA	1:10,000	7074S	Cell Signaling Technology	20 °C for 1h

NA, not applicable.

Nondenaturing immunoblot. Nondenaturing immunoblot analysis was performed using the NativePage system (Thermo Fisher Scientific). In brief, HEK293T cells were transfected with plasmid-encoding GREP1; 72 h after transfection, conditioned medium was collected and cellular debris removed via centrifugation and filtering of the medium. Protease inhibitor was added to the conditioned medium for preservation. Conditioned medium was then prepared with 4× NativePAGE sample buffer without heat, detergents or reducing agents. For comparison, conditioned medium was also prepared using 4× NativePAGE sample buffer and also 1% SDS and 10% NuPAGE sample-reducing agent (Thermo Fisher Scientific), followed by boiling at 95 °C for 5 min. Samples were then run on a NativePAGE Novex Bis-Tris gel using NativePAGE running buffer and NativePAGE 20× Cathode Buffer according to the manufacturer's instructions. Proteins were transferred to a polyvinylidene difluoride membrane after membrane activation with isopropanol using a semi-dry system of 7 V for 30 min at room temperature. After blocking for 1 h at room temperature in Odyssey Blocking Buffer, membranes were treated with rabbit anti-V5 antibody at 1:2,000 dilution (clone D3H8Q, no. 13202S, Cell Signaling Technology) overnight at 4 °C, then washed four times in 1× TBS-Tween and treated with goat anti-rabbit horseradish peroxidase (HRP) secondary antibody at 1:10,000 dilution (Cell Signaling, no. 7074S). Chemiluminescence was achieved with SuperSignal West Dura Extended Duration Substrate (Thermo Fisher Scientific), and images were developed with CareStream Kodak BioMax light film (Kodak).

Lentivirus production for L1000 experiments. Complete details of standard virus production pipelines can be found at the Broad Institute Genetic Perturbation Platform website: https://portals.broadinstitute.org/gpp/public/.

Virus was produced in arrayed 96-well plates via triple transfection of HEK293T cells with each packaging vector (100 ng), envelope plasmid (10 ng) and each respective pLX317 plasmid (100 ng). Lentiviral-containing supernatants were harvested at 32–56 h post transfection and stored in polypropylene plates at −80 °C until use.

Cell lines and lentiviral transduction for L1000 expression profiling. A549 and A375 cells were cultured in RPM1 medium supplemented with 10% FBS and 1% penicillin/streptomycin. MCF7 and HA1E cells were cultured in DMEM medium supplemented with 10% FBS and 1% penicillin/streptomycin. To perform L1000 high-throughput gene expression profiling, cells were robotically seeded (40 µl per well) into 384-well plates. Optimized seeding densities were either 250 cells per well (MCF7) or 400 cells per well (A549, A375 and HA1E). Twenty-four hours post seeding, cells were spin infected in the presence of polybrene (4 µg ml^{-1} for A549 and HA1E and 8 µg ml^{-1} for MCF7 and A375). The plates were then centrifuged for 30 min at 1,178*g* at 37 °C. The supernatant was robotically removed and replaced with fresh medium at either 3 h (A549) or 24 h post infection (A375, MCF7, HA1E), and cells were cultured for an additional 72 h till assay.

Infections were carried out in five replicates, three of which were used for the L1000 assay and two for assessment of infection efficiency. To assess infection efficiency, cells were treated with or without puromycin selection (1.5 µg ml^{-1}) 24 h post infection, and cell viability was determined using CellTiter-Glo (Promega) after 72 h of selection. For the remaining plates, medium was removed 96 h post infection and cells were lysed with the addition of TCL buffer (Qiagen). Plates were then sealed and stored at −80 °C until gene expression profiling.

L1000 experimental design. Two 384-well plates of perturbational ORFs were designed for cell treatment before L1000 profiling, each containing 352 unique ORFs, negative control ORFs, internal technical controls and untreated wells. The plate format can be found in Supplementary Fig. 7. In each plate, 346 wells were devoted to treatment ORFs and ten to ORFs targeting L1000 landmark genes, which were included for positive control purposes. These positive control wells would later be assessed for targeted gene *z*-score (≥2) and targeted gene rank (computed relative to the expression levels of that same gene across the assay plate). Control genes included were *ACAA1*, *ACD*, *AURKB*, *BMP4*, *CBR1*, *CCDC90A*, *CDK6*, *CSNK1A1*, *ETV1* and *SOX2*. Genes were selected for overall high baseline expression levels in the lines profiled and previous reproducibility in the L1000 assay. Additionally, 16 wells of negative control ORFs targeting blue fluorescent protein, eGFP or HcRed were added. Each plate also contained 12 untreated wells.

Cell lines MCF7, HA1E, A549 and A375 were chosen to represent a diversity of tissue types, and also to match CMap cell lines previously profiled extensively and that were constituents of the CMap reference database Touchstone[34].

L1000 data processing. Detailed protocols for the L1000 assay are provided at https://clue.io/sop-L1000.pdf. Each plate was profiled 96 h post infection. Antibiotic selection was not employed, and each plate was processed using the standard L1000 data-processing pipeline, which has been described elsewhere[34]. Briefly, mRNA was initially captured using 384-well oligo deoxythymidine (dT)-coated Turbocapture plates; after removal of lysate and addition of a reverse-transcription mix containing Moloney murine leukemia virus reverse transcriptase, the plate was washed and a mixture was added of both upstream and downstream probes (each containing a gene-specific sequence and a universal primer site) for each of the 978 ('Landmark') genes measured. The probes were first annealed to cDNA over a 6-h period and then ligated together to form a PCR template. After ligation, Hot Start Taq and universal primers were added to the plate, the upstream primer was biotinylated to facilitate subsequent staining with streptavidin-pycoerythrin, and the PCR amplicon was hybridized to Luminex microbeads using the complementary and probe-specific barcode on each bead. After overnight hybridization the beads were washed and stained with streptavidin-pycoerythrin, and Luminex FlexMap three-dimensional scanners were used to measure each bead independently, reporting bead color, identity and fluorescence intensity of the stain. Fluorescence intensity of staining values was then converted into median intensity values for each of the 978 measured genes using a deconvolution algorithm (resulting in 'GEX'-level data). These GEX data were then normalized relative to a set of invariant genes, and subsequently quantile normalized (resulting in 'QNORM'-level data). An inference model was applied to the QNORM data to infer gene expression changes for a total of 10,174 genes, which corresponds to the best inferred genes reported below. Next, expression values of each individual well were converted to robust *z*-scores by *z*-scoring gene expression relative to corresponding expression across the entire plate population; these *z*-scored differential expression gene signatures were finally replicate collapsed to a single differential expression vector per treatment, which we term a signature (and 'MODZ'-level data).

L1000 quality control. All samples profiled passed internal technical L1000 assay quality control measures described elsewhere[34]. Additionally, all samples included passed an internal fingerprinting algorithm that verifies the identity of cell lines on L1000 plates by comparing quantile-normalized gene expression data in each well to a ranked reference library of >1,000 cancer cell lines; samples are defined as passed if the Spearman correlation to their respective reference profile is higher than equivalent correlation values to all other reference cancer profiles. Additionally, 67% of positive control ORFs included had a replicate correlation of 0.25 or greater and an induced a *z*-score of 2 or greater in their target gene. Notably, ORFs targeting *CNSK1A1* represented the majority of poorly performing positive control ORFs. Positive control ORFs showing a high transcriptional activity score (TAS) also clustered together (Supplementary Fig. 7c).

Measures of L1000 signature bioactivity. Each perturbagen's transcriptional activity was represented using a TAS, which has been described in depth elsewhere[34]. Briefly, TAS is computed as a geometric mean of signature strength (SS—the number of landmark ($n=978$) genes in a signature with absolute z-score greater than or equal to 2) and replicate correlation (RC—the 75th quantile of all pairwise Spearman correlations between replicate-level z-score profiles):

$$\mathrm{TAS} = \sqrt{(\mathrm{SS} \times \max(\mathrm{RC}, 0))/978} \times (\mathrm{RC}, 0)$$

Signatures were considered to be bioactive if they had a TAS score of 0.2 or higher, which represents the value at which 95% of negative control wells fall below[34].

L1000 signature queries. Each MODZ-level signature profiled was queried against the other L1000 signatures in the dataset and the CMap dataset that has been published and described elsewhere[34]. Similarity values between these signatures were assessed using a percentile score derived from a normalized weighted connectivity score (WTCS). Briefly, WTCS is a similarity measure based on the weighted Kolmogorov–Smirnov enrichment statistic (ES) described previously[54] and is computed as follows for a given query gene set pair (q_{up}, q_{down}) and a reference signature r:

$$w_{q,r} = \left\{ \frac{\mathrm{ES_{up}} - \mathrm{ES_{down}}}{2} \text{ if } \mathrm{sgn}(\mathrm{ES_{up}}) \neq \mathrm{sgn}(\mathrm{ES_{down}}), 0 \text{ otherwise} \right\}$$

where $\mathrm{ES_{up}}$ is the enrichment of q_{up} in r, and $\mathrm{ES_{down}}$ is the enrichment of q_{down} in r. WTCS ranges between −1 and 1, and is positive for signatures that are positively related, negative for the converse and near zero for unrelated signatures.

WTCS is then normalized to allow for comparison of connectivity scores across cell and perturbagen types; this normalization is similar to that used in gene set enrichment analysis and accounts for differences in connectivity that may occur across such covariates. Given a vector of WTCS values from a query, normalization occurs as follows:

$$\mathrm{NCS_{c,t}} = \left\{ \frac{w_{c,t}}{\mu^{+}_{c,t}} \text{ if } \mathrm{sgn}(w_{c,t}) > 0, \frac{w_{c,t}}{\mu^{-}_{c,t}} \text{ otherwise} \right\}$$

where $\mathrm{NCS}_{c,t}$, $w_{c,t}$, $\mu+_{c,t}$ and $\mu-_{c,t}$ are, respectively, the normalized connectivity scores, raw WTCS and signed means (the mean of positive and negative values evaluated separately) of the WTCS values within the subset of signatures corresponding to cell line c and perturbagen type t.

Lastly, NCS scores are converted to percentile scores accounting for whether the connectivity between the queried (q) and reference signature (r) is significantly different from that observed between r and other queries. This is done by comparing each observed NCS value, $\mathrm{ncs}_{q,r}$, between q and r to a distribution of NCS values representing the similarities between a reference compendium of queries (Q_{ref}) and r. This procedure results in a standardized measure we refer to as Tau (τ), that ranges from −100 to +100 and represents the percentage of queries in Q_{ref} with a lower |NCS| than $|\mathrm{ncs}_{q,r}|$, adjusted to retain the sign of $\mathrm{ncs}_{q,r}$ and reliant on the following formula:

$$\tau_{q,r} = \mathrm{sgn}\left(\mathrm{ncs}_{q,r} \times \left(\frac{100}{N}\right) \times \sum_{i=1}^{N} \left[\left|\mathrm{ncs}_{q,i}\right| < \left|\mathrm{ncs}_{q,r}\right|\right]\right)$$

where $\mathrm{ncs}_{q,r}$ is the normalized connectivity score for signature r with respect to query q; $\mathrm{ncs}_{i,r}$ is the normalized connectivity score for signature r relative to the ith query in Q_{ref} (a set of query signatures obtained from exemplar signatures of perturbagens matching the cell line and perturbagen type of r; and N is the number of queries in Q_{ref}.

L1000 software packages used. L1000 data were analyzed using the 'tidyverse' suite[55] of R packages (v.1.2.1) and the 'cmapR' package[56] (v.1.0.1) in R v.3.5.0 (R Core Team 2018).

CRISPR sgRNA design. Single-guide RNAs for the ORFs in this study were designed using the Broad Institute GPP sgRNA designer for *Streptococcus pyogenes* Cas9 against genome coordinates for the GRCh38 assembly of the human genome (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design). Only exonic coding regions for ORFs were used. A maximum of eight unique sgRNAs were employed per gene: if fewer than eight were nominated due to small gene size and lack of available photospacer adjacent motif (PAM) sites, all nominated sgRNAs were used. If more than eight sgRNAs were nominated, the eight top-ranked were used according the Broad Institute GPP sgRNA designer pick analysis. For the secondary CRISPR screen, 147 ORFs were tested. These were chosen to include all ORFs with a viability phenotype in the primary screen in the appropriate cell lines (A375, MCF7 and HEPG2), as well as additional ORFs that did not have a viability phenotype.

For tiling sgRNA analyses, additional nominated sgRNAs for each ORF were selected. Also, we selected sgRNAs to putative 3′ UTR, 5′ UTR and promoter regions (defined as within 1,000 bp of the transcript start site). A maximum of 16 sgRNAs were designed for each region. If multiple UTR exons were present, a maximum of 16 sgRNAs were designed for each. Intronic sgRNAs were used where available and were limited to six sgRNAs per intron. sgRNAs for adjacent protein-coding genes were also employed as indicated, and designed in an identical manner. The number of sgRNAs for adjacent coding genes and various genome regions is detailed in Supplementary Tables 25 and 26.

Determination of infection conditions for CRISPR pooled screens. Optimal infection conditions were determined in each cell line to achieve 30–50% infection efficiency, corresponding to a multiplicity of infection (MOI) of ~0.5–1. Spin infections were performed in 12-well plate format with 3×10^6 cells per well. Optimal conditions were determined by infecting cells with different virus volumes at a final concentration of 4 μg ml^{-1} polybrene. Cells were spun for 2 h at 1,000g at 30 °C. Approximately 24 h after infection, cells were trypsinized and A375, HT-29 and PC-3 cells (2×10^5); A549 and HeLa cells (1.5×10^5); HepG2 cells (3×10^5); and MCF7 cells (5×10^5) from each infection were seeded in two wells of a six-well plate, each with complete medium; one was supplemented with the appropriate concentration of puromycin (1.5 μg ml^{-1} for A375; 2 μg ml^{-1} for A549, MCF7 and PC-3; and 1 μg ml^{-1} for HeLa, HA1E, HepG2 and HT-29). For the secondary screen, only HepG2, MCF7 and A375 were used. Cells were counted 4–5 d post selection to determine infection efficiency, comparing survival with and without puromycin selection. Volumes of virus that yielded ~30–50% infection efficiency were used for screening.

Primary and secondary CRISPR pooled proliferation screens. The lentiviral barcoded library used in the primary screen contains 5,235 sgRNAs, which includes an average of eight guides per gene and 500 nontargeting control guides. The validation library contains 6,996 sgRNAs targeting selected regions of the ORFs. Genome-scale infections were performed in three replicates with the predetermined volume of virus in the same 12-well format as the viral titration described above, and pooled 24 h post centrifugation. Infections were performed with sufficient numbers of cells per replicate, to achieve a representation of at least 1,000 cells per sgRNA following puromycin selection (~1.5×10^7 surviving cells). Approximately 24 h after infection, all wells within a replicate were pooled and were split into T225 flasks. Twenty-four hours after infection, cells were selected with puromycin for 7 d to remove uninfected cells. After selection was complete, $1.5–2.0\times10^7$ cells were harvested for assessment of the initial abundance of the library. Cells were passaged every 3–4 d and harvested ~21 d after infection. For all genome-wide screens, genomic DNA was isolated using Midi or Maxi kits according to the manufacturer's protocol (Qiagen). PCR and sequencing were performed as previously described[57,58]. Samples were sequenced on a HiSeq2000 (Illumina). For analysis, read counts were normalized to reads per million and then $\log_2$ transformed. The $\log_2$(fold change (FC)) of each sgRNA was determined relative to the initial time point for each biological replicate.

Analysis of CRISPR screening data. CRISPR data were analyzed as $\log_2$(FC) values computed between the day 21 time point and the input plasmid DNA: $\log_2(\mathrm{FC}) \leq -1$ was defined as a scoring sgRNA, which was depleted from the analysis. In the primary screen, a gene with at least two sgRNAs with $\log_2(\mathrm{FC}) \leq -1$ in at least one cell line was defined as a putative vulnerability hit. Because the vast majority of genes in the primary screen had eight sgRNAs per cell line, genes could be compared against each other with this metric. In the secondary screen, because of variation in the number of sgRNAs for each gene, a scoring candidate was defined as a gene in which at least 10% of sgRNAs had $\log_2(\mathrm{FC}) \leq -1$ and there were at least two sgRNAs with $\log_2(\mathrm{FC}) \leq -1$ in at least one cell line. sgRNAs were also analyzed using STARS v.1.3 and CERES scores as previously described[57,59].

Analysis of CRISPR tiling screen. $\log_2$(FC) values for each sgRNA at day 21 of the screen were considered as above. sgRNAs were then grouped into their respective genomic region (for example, UTR, ORF exon, adjacent gene exon or intron). The mean $\log_2$(FC) for each region was computed. A mean $\log_2(\mathrm{FC}) \leq -1$ was considered a scoring hit. Genes were then classified in the following manner according to the viability affect of sgRNAs: 'specific to ORF' if only the ORF region sgRNAs scored; 'specific to ORF and transcript subregion' if ORF sgRNAs and sgRNAs to only one other region of the RNA transcript scored; 'specific to transcript' if sgRNAs to any part of the ORF or RNA transcript scored, but not sgRNAs to introns or genomic regions; 'shared with adjacent gene' if the ORF and an annotated adjacent protein-coding gene both scored; and 'nonspecific to the genome' if sgRNAs to any part of the genomic region, intron, RNA transcript or ORF all demonstrated depletion.

Comparison of CRISPR screen data with Project Achilles. For each gene of ORF in each of the eight cell lines used in the primary ORF CRISPR screen, knockout was determined as having produced depletion if at least two guides produced at least 50% depletion from initial abundance after RPM normalization. The file 'Achilles_logfold_change' in DepMap_public_19Q4 was used for Achilles screens (available at https://depmap.org/portal/download). To determine the expected number of genes or ORFs that deplete in any cell line given n cell lines, all possible subsets of n lines were selected and the number of genes with at least one depleted line were counted. For a negative control, this process was repeated in Achilles screens using only genes proposed as nonessential by previously published RNA interference data[60], to generate a null distribution.

Off-target sgRNA effect prediction. For the 57 putative hits in the primary CRISPR screen, we analyzed scoring sgRNAs for off-target genomic homology sites using the Cas-OFFinder v.1.0 algorithm[61]. Homology sites were computed using default program settings with a mismatch tolerance of 0, DNA bulge of 0 and RNA bulge of 0. Predicted off-target sites are listed in Supplementary Tables 30 and 31. In addition, all ORF-targeting sgRNAs in the primary screen were analyzed for specificity or off-target sites using BLAT through the UCSC Genome Browser.

GREP1 annotation analysis and expression data. *GREP1* annotation status was evaluated using the indicated historical versions of the GENCODE database with graphic visualization of the locus. In cell lines, *GREP1* expression was evaluated through Cancer Cell Line Encyclopedia data for *LINC00514* (NR_033861.1), a RefSeq annotation that incorporates the first portion of *GREP1*. CCLE data were downloaded from https://portals.broadinstitute.org/ccle.

Pooled *GREP1* knockout. For the pooled *GREP1* CRISPR knockout assay, we used a pool of 486 barcoded, adherent human cancer cell lines developed at the Broad Institute[62]. The cell line pool was grown in RPMI 1640 medium supplemented with 10% FBS. sgRNAs used for this experiment were noncutting control sgLacZ (AACGGCGGATTGACCGTAAT), cutting control sgChr2 (GGTGTGCGTATGAAGCAGTGG), sg*GREP1* no. 1 (ACTCAAAATGGCTATAGACC) and sg*GREP1* no. 2 (AGGCTTTAGAGGGGACATGA). On Day 0, the cell line pool was plated in six-well plates at 400,000 cells per well in 3 ml of cell culture medium. Twenty-four hours later, using an all-in-one Cas9/sgRNA plasmid, the cell line pool was infected with each lentivirus at an MOI of 10; lentivirus was concentrated before use to obtain a concentration of $>1\times10^7$ particles ml^{-1}. Cells were also treated with 4 $\mu g\,ml^{-1}$ polybrene in 2 ml per well for lentiviral infection, and spun at 2,250 r.p.m. for 1 h at 37 °C. Twenty-four hours after transduction, cells were split from one well in a six-well plate into two T25 flasks; at this time point the baseline cell DNA lysate was harvested as a 'no infection' control. Seventy-two hours after infection, cell culture medium was changed and puromycin selection was started at a concentration of 1 $\mu g\,ml^{-1}$ puromycin. Thereafter, cell culture medium was changed every 72 h and cells were expanded as needed into T75 and T175 flasks. Pooled cell line DNA was collected from the input plasmid pool, on day 6 as an early time point and on day 15 as a late time point, to assess for dropout of cell line. At each sample time point, cells were counted and 2×10^6 cells were removed for lysis for DNA. For lysis, cells were pelleted, washed in PBS and genomic DNA was extracted with the DNA Blood and Tissue Kit according to the manufacturer's instructions (Qiagen). The remaining cells not required for lysis were reseeded into T75 and T175 flasks for continuation of cell growth.

For sequencing, time point DNA was subjected to PCR using universal barcode primers. PCR products were run on a 2% agarose gel to confirm amplicon size, then 10 µl from each PCR product was pooled and purified with AMPur beads (Beckman Coulter). DNA concentration was measured via Qubit fluorometric quantification (Thermo Fisher Scientific) and DNA was sequenced using NovaSeq (Illumina) at the Genomics Platform at the Broad Institute.

Analysis of pooled *GREP1* knockout sequencing data. Cell line abundance was calculated based on cell line barcode detection by next-generation sequencing as previously described[62]. To analyze the pooled *GREP1* CRISPR knockout data, we first calculated the theoretical number of cells in each well at each time point based on the experimental measurements of the total number of cells and the number of cells removed for sequencing. We accounted for these discarded cells by scaling the measured number of cells at a given time point by the ratio of the total number of cells at the previous time point to the number of reseeded, or retained, cells from the previous time point.

Next, for quality control, we computed the purity of each sample as the percentage of read counts mapping to cell lines not included in the pool. We removed samples with purity <95%, then filtered out cell lines with <12 reads in more than one replicate of either of the two negative control conditions, LacZ and Chr2. The conservative threshold of 12 was determined from the minimum number of counts at which we are able to distinguish between that number of counts and half that number, at a confidence level of 0.05, under a Poisson distribution.

Next, we added a pseudocount of 1 to each of the read counts and normalized updated read counts by library size and theoretical total cell count. We define the log(FC) of a cell line in a sample as the $\log_2$-transform of the ratio of the normalized read count of the cell line in the sample to the normalized read count of the cell line at day 0. Finally, we define viability as the difference between log(FC) in the cell line and treatment of interest and the average of log(FC) in the cell line and the two negative controls.

We then developed a series of data-processing steps to empirically improve the quality of the dataset (Supplementary Fig. 11). First, we excluded cell lines believed to be puromycin resistant based on the criterion of positive viability in the puromycin, no-virus condition. These filters resulted in a viability dataset of 400 out of 486 cell lines. We then removed cell lines exhibiting excessive lentiviral toxicity given the high MOI used for this experiment; this left 320 cell lines. Next, we eliminated cancer type cohorts with ≤5 cell lines, due to insufficient numbers for analysis, leaving 294 cell lines. Lastly, we calculated the number of cell lines per cancer cohort expressing *GREP1* above a minimal threshold, and excluded cohorts with insufficient expression because any change in those cohorts may be spurious due to population shifts in the cell line pool or off-target effects.

CRISPR-sequencing. ZR-75-1 cells infected with lentivirus for sgCh2-2 negative control, sgGREP1 nos. 1 or 2 and antibiotic-resistant cells were selected with 2 $\mu g\,ml^{-1}$ puromycin for 48 h as described previously. Ninety-six hours after infection, genomic DNA from cells was isolated using the Qiagen DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's instructions; 100 ng of DNA was amplified by PCR with the following thermocycler conditions: 94 °C for 2 min, followed by 30 cycles at 94 °C for 30 s, 52 °C for 30 s and 68 °C for 1 min; final elongation was at 68 °C for 7 min. PCR products were confirmed for specificity with a 1% agarose gel and then gel purified using a Qiagen Gel Extraction kit according to the manufacturer's instructions. DNA was diluted to a concentration of 25 $ng\,\mu l^{-1}$ and submitted to the Massachusetts General Hospital Center for Computational and Integrative Biology DNA Core for sequencing. FASTQ sequencing files were analyzed using CRISPResso[63] v.2 (http://crispresso.pinellolab.partners.org) according to default parameters.

Patient outcome analysis for *GREP1*. Expression data for *GREP1* in the TCGA samples was acquired from the publicly available MiPanda tool using the *LA16c-H380H5.3* gene annotation as a query[64]. Data for the GDC TCGA Breast Cancer and GDC TCGA Colon Cancer datasets were used. LINC00514 expression was extracted as a proxy for *GREP1* given that LINC00514 is a fragment of the longer gene. Overall survival was also extracted for these datasets. Kaplan–Meier curves and statistical significance by log-rank *P* value were generated using GraphPad Prism8 software, with $P<0.05$ being considered statistically significant.

***GREP1* copy number analysis.** CCLE copy number data from the 2013-12-03 segmentation were downloaded from https://depmap.org/portal/download. Data for LINC00514 (283875) were used as a proxy for *GREP1* given overlapping genomic regions. Copy number data were then aggregated by cell line lineage.

CRISPR validation experiments. Cells were plated in 96-well plates and allowed to grow for 4–8 h before infection with the indicated sgRNA or treatment condition; 1,000–5,000 cells per well were plated depending on the cell line. sgRNAs were obtained from either the Broad Institute Genomic Perturbation Platform or direct synthesis into the LentiGuide-Puro plasmid backbone via a commercial vendor (GenScript). sgRNA sequences are listed below.

Gene	sgRNA no.	sgRNA sequence
ASNSD1	1	GCTCACGTCCTACACTTGAG
ASNSD1	2	TTTGGGTGCCAACTGAAGAG
ASNSD1 uORF	1	GCTTAGATCCTCCTTGTGTG
ASNSD1 uORF	2	TAAAGAACAAAAAATTGTGG
chr2-2	NA	GGTGTGCGTATGAAGCAGTGG
COG7	2	TGTTGAAGCCCTAAAACAGG
COG7	1	CTACTACTACAAGTGTCACA
GREP1	1	ACTCAAAATGGCTATAGACC
GREP1	2	AGGCTTTAGAGGGGACATGA
GREP1	3	GCTCAAAATGGCTTTGGACC
HP08474	1	TGTGTTTGAGCCAGGCATGG
HP08474	2	AGTCCCAGCAGCTACTCCGG
RP11-277L2.3	1	CGCCTCCTGGGTTCCAGCAG
RP11-277L2.3	2	GGGACTAGATGGAGCCGAAG
RP11-54A9.1	1	TGGGTCTCCTCACAGAGTGA
RP11-54A9.1	2	TCCTCAGACCAACCAGCTCA
LacZ	NA	AACGGCGGATTGACCGTAAT
ZBTB11-AS1	1	GCGGGACTCTGTATTACCAG
ZBTB11-AS1	2	GCGACGCCGGGACCTCATCG
CTD-2270L9.4	1	CGTGAAGGAGTGGATCAATG
CTD-2270L9.4	2	GAACTTGGAGAAGTCCATGG
G083755	1	CCAACAGGTGACCTCAGCAA
G083755	2	GGACCTCTTACATCATGGAA
SF3B1	NA	AAGGGTATCCGCCAACACAG
ZBTB11	1	ACAGGTTGACACCAAAGGAG
ZBTB11	2	GCATATATTCGACTACACAA
OLMALINC	1	ACAGGGCACTGGTCTCCCAA
OLMALINC	2	CAAGGCTGTATATTTCACCT

All sgRNAs were sequenced and verified. Following sequence verification, constructs were transfected with packaging vectors into HEK293T with Fugene HD (Sigma-Aldrich). After plating, cells were then infected with sgRNA lentivirus to achieve maximal knockout but without viral toxicity. Sixteen hours after infection, cells were selected with 2 μg μl^{-1} puromycin (Invitrogen) for 48 h. Cell viability was measured using CellTiter-Glo reagent (Promega) at 16 h post transfection for baseline assessment, and at additional time points as needed. For stable knockout cell lines, cells were plated at equal densities and cell viability was measured by CellTiter-Glo every 24 h as indicated.

***GREP1* overexpression rescue experiments.** For CRISPR rescue experiments, Cas9-derived cell lines were infected with lentivirus GFP- or GREP1-coding plasmids cloned into the pLX_TRC313 vector, which has EF1a promoter and hygromycin resistance (https://portals.broadinstitute.org/gpp/public/vector). Cells were selected in 350 μg ml^{-1} hygromycin for 72 h before transitioning back to standard culture medium.

In 96-well plates, 5,000 ZR-75-1-derived cells were plated and infected with the indicated sgRNA lentivirus 4–6 h after plating; 16 h after infection, cells were selected with 2 μg ml^{-1} puromycin for 48 h and grown for 7 d before cell viability analysis using CellTiter-Glo reagent.

Conditioned media rescue experiments. On day –2, HEK293T cells were plated and transiently transfected with GFP and GREP1 using Fugene HD reagent. On the same day, either 5,000 ZR-75-1-derived cells or 2,500 AU565-derived cells were plated in wells of a 96-well plate. On day –1, ZR-75-1 and AU565 cells were switched to serum-free medium. On day 0, conditioned medium from GFP- or GREP1-expressing HEK293T cells was cleared of cellular debris by centrifugation and then 100 μl of conditioned medium was applied to each well. Conditioned medium was then refreshed daily and cell viability was determined with CellTiter-Glo reagent at the indicated time points.

Immunoprecipitation. HEK293T cells were transiently transfected with GFP-V5 or GREP1-V5 fusion proteins using OptiMem and Fugene HD (Sigma-Aldrich). Seventy-two hours after transfection, cell culture medium was harvested and cell debris sedimented by centrifugation twice at 1,500 r.p.m. for 5 min. The resulting cell culture medium was concentrated in a 10:1 ratio using a 3-kDa size-exclusion filter (Millipore), and concentrated culture medium treated with HALT protease inhibitor. Next, all immunoprecipitation steps were performed either on ice or in a cold room (4 °C). First, culture medium was cleared with Pierce magnetic A/G beads (Thermo Fisher Scientific) for 1 h while rotating at 18–20 r.p.m.. Beads were then discarded, and 10% of the medium was removed as an input sample and kept at 4 °C without freezing. The remaining culture medium was then treated with 50 μl of magnetic anti-V5 beads (MBL International) and rotated at 18–20 r.p.m. overnight at 4 °C. The following day, the supernatant was discarded and beads were washed four times in immunoprecipitation wash buffer (50 nM Tri-HCl pH 8.0, 150 nM NaCl, 2 mM EDTA pH 8.0, 0.2% NP-40 and 1 μg ml^{-1} PMSF protease inhibitor) with rotation for 10 min per wash. After the final wash, beads were gently centrifuged and residual wash buffer was removed. Then, proteins were eluted twice with 2 μg μl^{-1} V5 peptide in water (Sigma-Aldrich) at 37 °C for 15 min with shaking at 1,000 r.p.m. The two elution fractions were pooled and samples were prepared with 4× LDS sample buffer and 10× sample-reducing agent (Thermo Fisher Scientific), followed by boiling at 95 °C for 5 min. One-third of the eluate was then run on a 10–20% Tris-glycine SDS–PAGE gel and stained with SimplyBlue Coomassie stain (Thermo Fisher Scientific) for 2 h. Gels were destained with a minimum of three washes in water for at least 2 h per wash. Bands were visualized using Coomassie autofluorescence on LI-COR Odyssey in the 800-nM channel. Gel lanes were then cut into six equal-sized pieces using a sterile razor under sterile conditions, and stored in 1 ml of diethyl pyrocarbonate-treated water before MS analysis.

Methods for protein sequence analysis by liquid chromatography–MS/MS. Liquid chromatography–MS/MS was performed in the Taplin Biological Mass Spectrometry Facility at Harvard Medical School. Briefly, excised gel bands were cut into pieces of approximately 1 mm^3, which were then subjected to a modified in-gel trypsin digestion procedure[65]. Gel pieces were washed and dehydrated with acetonitrile for 10 min. followed by removal of acetonitrile and then completely dried in a speed-vac. Rehydration of gel pieces was performed with 50 mM ammonium bicarbonate solution containing 12.5 ng μl^{-1} modified sequencing-grade trypsin (Promega) at 4 °C. After 45 min, excess trypsin solution was removed and replaced with 50 mM ammonium bicarbonate solution to cover the gel pieces. Samples were then placed in a room overnight at 37 °C. Peptides were later extracted by removal of the ammonium bicarbonate solution, followed by one wash with a solution containing 50% acetonitrile and 1% formic acid. The extracts were then dried in a speed-vac (~1 h) and stored at 4 °C until analysis.

On the day of analysis, samples were reconstituted in 5–10 μl of high-performance liquid chromatography (HPLC) solvent A (2.5% acetonitrile, 0.1% formic acid). A nanoscale reverse-phase HPLC capillary column was created by packing 2.6 μm of C18 spherical silica beads into a fused silica capillary (100 μm inner diameter × ~30 cm length) with a flame-drawn tip[66]. After equilibration of the column, each sample was loaded using a Famos autosampler (LC Packings) onto the column. A gradient was formed, and peptides were eluted with increasing concentrations of solvent B (97.5% acetonitrile, 0.1% formic acid).

As peptides eluted they were subjected to electrospray ionization and then entered into an LTQ Orbitrap Velos Pro ion-trap mass spectrometer (Thermo Fisher Scientific). Peptides were detected, isolated and fragmented to produce a tandem mass spectrum of specific fragment ions for each peptide. Peptide sequences (and hence protein identity) were determined by matching protein databases with the acquired fragmentation pattern using the software program Sequest[67] (Thermo Fisher Scientific). All databases include a reversed version of all sequences, and data were filtered to FDR = 1–2%. Glycosylated peptides were defined using the A score method as previously described[68].

Immunoprecipitation–MS and gene ontology analysis. We analyzed Immunoprecipitation–MS (IP–MS) data from two independent experiments for V5 immunoprecipitation for GFP-V5- and GREP1-V5-conditioned media in HEK293T cells, and from one biological replicate for GFP-V5 and GREP1-V5 in CAMA-1 and ZR-75-1 cells. IP–MS data were merged for the two experiments and all proteins with fewer than two total peptides were removed to exclude technical artifacts. To the remaining proteins, a pseudocount of 1 was added to ensure a nonzero denominator. Next, the fold change of (GREP1 + 1)/(GFP + 1) peptide count was calculated and $\log_{10}$-transformed. Enriched peptides with a (GREP1 + 1)/(GFP + 1) ratio of ≥2 were further analyzed using the Gene Ontology database (http://geneontology.org) for cellular component analysis; corrected FDR values were plotted.

GREP1 disorder analysis. The GREP1 primary amino acid sequence was analyzed via the DISOPRED3 package[69] on the PsiPred server (http://bioinf.cs.ucl.ac.uk/psipred/) using default settings. Disorder scores were plotted as indicated.

GREP1 evolutionary analysis. The GREP1 amino acid sequence (ENST00000573315.2_prot) was aligned to nonredundant protein sequences using the NCBI BlastP suite, and manually aligned to the genomes of the common rat (RGSC 6.0/rn6, July 2014 assembly) and domestic dog (Broad CanFam3.1/canFam3 assembly). The resulting protein hits were then ranked by e-value and the most significant result was used for each organism. Predicted proteins and low-quality protein assemblies were included in this analysis. Resultant species-specific amino acid sequences were then aligned by the Clustal Omega sequence aligner (https://www.ebi.ac.uk/Tools/msa/clustalo/), and percentage similarity to human GREP1 was plotted.

GREP1 codon usage analysis. We calculated the triplet codon frequency for all triplet codons for the GREP1 amino acid sequence, the whole ORFeome in total and GENBANK genes by collating all mRNA sequences within these respective groups and calculating codon usage per group. Each codon usage was normalized to a standard rate of codon usage per 1,000 codons. Triplet codons were then collapsed into single amino acids by scaling codon usage rate to the relative frequency of usage for each codon per amino acid. Aggregate frequency of amino acid representation was then calculated and compared across groups.

Cytokine profiling array. Cytokine profiling was performed simultaneously using the Human XL Cytokine Array (R&D Systems, no. ARY022). Briefly, cell culture media were cleared of cellular debris and Halt protease inhibitor was added as above. Then, cytokine arrays were blocked in 2 ml of array buffer 6 (blocking buffer), each for 1 h on a shaker at room temperature. Samples were prepared with 300 μl of culture medium and diluted with 1,200 μl of array buffer 6. Cytokine arrays were then removed from the blocking buffer and incubated with samples overnight at 4 °C on a rocker. The following morning, array membranes were washed in 20 ml 1× wash buffer for a total of three washes. Then, arrays were placed in 1.5 ml of 1× array buffer 4/6 (a 1:2 mixture of array buffers 4 and 6), and 30 μl of reconstituted detection antibody cocktail was added. Samples were incubated for 1 h at room temperature on a shaker. Subsequently, membranes were washed in 20 ml of 1× wash buffer for a total of three washes and then transferred to 2.0 ml of 1× streptavidin-HRP for 30 min at room temperature on a shaker, followed by three more washes in 20 ml of 1× wash buffer. Subsequently the membranes were blotted on tissue paper to remove excess buffer, and signal was developed with chemiluminescent reagent mix. Images were developed with CareStream Kodak BioMax light film (Kodak).

Cytokine profiling analysis. Immunoblot images of cytokine arrays were scanned, and the signal intensity of all array antibody spots was determined using ImageJ v.2.0.0 (https://imagej.nih.gov/ij/index.html). Raw data values were then inverted using the formula $y = 255 - x$, where x is raw signal intensity. Inverted values were then normalized according to knockout or overexpression experiments. For the former, signal was analyzed as sgControl – sgGREP1; for the latter, signal was analyzed as GREP1 – GFP. The absolute value of signal change was then averaged across experiments and rank-listed according to the magnitude of average change.

GDF15 enzyme-linked immunosorbent assay. The GDF15 Quantikine ELISA kit (R&D Systems) was used. In brief, cell culture media samples were diluted 1:3 by volume in Diluent RD5-20. To prepare microplate wells, 100 μl of Assay Diluent

RD1-9 was added to each well, then 50 µl of standards, controls or diluted samples was added to a given well. The plates were incubated for 2 h at room temperature on a horizontal orbital microplate shaker at 500 r.p.m. Wells were then washed four times with 400 µl of 1× wash buffer for 5 min per wash; after the final wash, plates were inverted and blotted on tissue paper to remove excess. Then, 200 µl of Human GDF15 conjugate was added to each well and the plate was incubated for 1 h at room temperature on an orbital shaker. Following this, wells were then washed four times with 400 µl of 1× wash buffer for 5 min per wash; after the final wash, plates were inverted and blotted on tissue paper to remove excess. Then, 200 µl of substrate solution was added per well and plates were incubated for 30 min at room temperature without shaking and protected from light. Next, 50 µl of stop solution was added per well and samples were mixed with gentle tapping. The optical density of samples at 450 and 570 nM was determined on a microplate reader within 15 min of completion of the protocol. For analysis, background signal from 570 nM was subtracted per well from the 450-nM signal. Samples were then calculated based on a standard curve to obtain GDF15 concentration values. For pharmacologic treatments preceding GDF15 measurements, HEK293T cells were treated with 10 µM of either vorinostat, idarubicin, GSK265, bortezomib, GSK132 or luminespib for 24 h. Cells with transient transfection of GFP or GREP1 cDNA were treated with DMSO as controls. After 24 h, GDF15 abundance was measured in conditioned medium by enzyme-linked immunosorbent assay (ELISA).

Correlation of *GREP1* and *GDF15* expression. Expression of *GREP1*, *GDF15*, *FN1* and *EMIL2* was downloaded via the MiPanda portal[64] as TPM values. GTex and TCGA samples were used. Spearman rho correlation coefficients and Spearman *P* values were calculated using GraphPad Prism8 and plotted.

Recombinant GDF15 experiments. Recombinant human GDF15 (R&D Systems, catalog no. 957-GD-025) was resuspended in water at 10 µg µl^{-1}. Knockout with sgGREP1 no. 2 or controls in ZR-75-1 was performed as described above. Twenty-four hours after infection with lentiviral sgRNA, cell culture medium was refreshed with the addition of puromycin, as described above for antibiotic selection, and GDF15 or vehicle control was supplemented at the following concentrations: 0.01, 0.1, 1.0, 10 and 100 pg ml^{-1}. Thereafter, cell culture medium and recombinant GDF15 were refreshed every 24 h. Cell viability was measured 7 d after lentiviral infection using CellTiter-Glo reagent (Promega).

Generation of *GREP1* glycosylation mutants. V5-tagged *GREP1*, *T63V*, *T265V* and *T63V/T265V* double-mutant cDNA constructs were generated through a commercial service with GenScript in the plx307 vector. Briefly, for each respective construct, threonine at position 63 was mutated to valine with mutations A187G and C188T to change codon ACC to GTC; the threonine at position 265 was mutated to valine with mutations A748G and C749T to change codon ACC to GTC. The *GREP1 T63V/T265V* double-mutant construct harbored all four base pair changes. For GDF15 analyses, the indicated constructs were transiently transfected into HEK293T cells as described previously and GDF15 was measured in conditioned medium 48 h later as previously described.

ZBTB11-AS1 knockdown experiments. A549 cells with transduced lentivirus encoding GFP, ZBTB11-AS1 ORF or mutant ZBTB11-AS1 ORF with mutated ATG and antibiotic-resistant cells were isolated with 2 µg ml^{-1} puromycin for 72 h, and 500,000 cells of the given cell line were plated in six-well plates in serum-free medium. Four hours after plating, wells were individually transfected with 20 µM of the indicated siRNA oligonucleotide or nontargeting control mixed in 135 µl of OptiMem with 10 µl of Lipofectamine 2000 (Thermo Fisher Scientific). Twelve hours later, serum-containing medium was added and cells were grown for 48 h. Cells were then trypsinized and plated in 96-well plates at a density of 5,000 per well in six replicates. Cell viability was measured 72 h later using CellTiter-Glo reagent (Promega). siRNA sequences were as follows: Lincode ZBTB11-AS1 no. 1, 5′-GGACGAAUCUGCAGCGCUC-3′ (catalog no. N-188908-01-0002, Dharmacon, Horizon Discovery); Lincode ZBTB11-AS1 no. 3, 5′-GUUGAGAGUUCAGCCGAAA-3′ (catalog no. N-188908-03-0002, Dharmacon, Horizon Discovery); ON-TARGET plus nontargeting siRNA no. 1, 5′-UGGUUUACAUGUCGACUAA-3′ (catalog no. D-001810-01-20, Dharmacon, Horizon Discovery); and ON-TARGET plus nontargeting siRNA no. 3, 5′-UGGUUUACAUGUUUUCUGA-3′ (catalog no. D-001810-03-20, Dharmacon, Horizon Discovery). Knockout efficiency was monitored by qPCR.

Statistical analyses for experimental studies. All data are expressed as means ± s.d. All experimental assays were performed in duplicate or triplicate. Statistical analysis was performed by either two-tailed Student's *t*-test, one- or two-way ANOVA, Kolmogorov–Smirnov test, log-rank *P* value or other tests as indicated. $P < 0.05$ was considered statistically significant.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Processed data for CRISPR screens (Figs. 3 and 4d) are available in Supplementary Tables 22 and 27. Raw data are available in the Source data files accompanying this manuscript, as well as through the NCBI Sequence Read Archive at: SRR13126801, SRR13128583, SRR13132373, SRR13142215 and SRR13142421. Mass spectrometry data relating to Fig. 1 are available in Supplementary Table 14. Raw MS spectra are available through the original datasets at: https://cptac-data-portal.georgetown.edu/study-summary/S060 (CPTAC2_BRCA_prosp), https://cptac-data-portal.georgetown.edu/study-summary/S045 (CPTAC2_COAD_prosp), https://cptac-data-portal.georgetown.edu/study-summary/S050 (CPTAC3_ccRCC), https://cptac-data-portal.georgetown.edu/study-summary/S056 (CPTAC3_LUAD), https://cptac-data-portal.georgetown.edu/study-summary/S051 (CPTAC3_PTRC_DP1), https://cptac-data-portal.georgetown.edu/study-summary/S053 (CPTAC3_UCEC), ftp://massive.ucsd.edu/MSV000080527 (HLA_Abelin), ftp://massive.ucsd.edu/MSV000084787 (HLA_Ouspenskaia), ftp://massive.ucsd.edu/MSV000084172/; ftp://massive.ucsd.edu/MSV000080527; ftp://massive.ucsd.edu/MSV000084442/ (HLA_Sarkizova), ftp://massive.ucsd.edu/MSV000082644 (CPTAC Medulloblastoma) and http://www.peptideatlas.org (PeptideAtlas database). L1000 data relating to Fig. 2 and Supplementary Figs. 8 and 9 are available through the NIH LINCS program and at https://clue.io/data. The website lincsproject.org provides information about the LINCS consortium, including data standards. Source data are provided with this paper.

Code availability

L1000 data analysis code and preprocessed data are available via GitHub: https://github.com/cmap/cmapM. There is additional information about this database and tools at http://clue.io/connectopedia. L1000 data were analyzed via the following: the 'tidyverse' suite36 of R packages (v.1.2.1), the 'cmapR' package37 (v.1.0.1) in R v.3.5.0 (R Core Team 2018) and in-house code available through github (https://github.com/johnprensner/smORF_analyses). Mass spectrometry peptides were processed via Spectrum Mill MS Proteomics Workbench v.6.0. Additional code for computational tools used in this study is listed here: PhyloCSF (https://github.com/mlin/PhyloCSF/wiki) for 29-mammal alignment, Slncky (https://slncky.github.io), STARS v.1.3 (http://www.broadinstitute.org/rnai/public/software/index) and CERES v.1.0 (https://github.com/cancerdatasci/ceres).

References

45. Xie, W. et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
46. Chen, J. et al. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol.* **17**, 19 (2016).
47. Liu, S. J. et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, aah7111 (2017).
48. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
49. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
50. Domazet-Loso, T., Brajkovic, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
51. Domazet-Loso, T. et al. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol. Biol. Evol.* **34**, 843–856 (2017).
52. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
53. Yang, X. et al. A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* **8**, 659–661 (2011).
54. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
55. Ross, Z., Wickham, H., Robinson, D. Declutter your R workflow with tidy tools. Preprint at *PeerJ* https://peerj.com/preprints/3180.pdf (2017).
56. Enache, O. M. et al. The GCTx format and cmap{Py, R, M, J} packages: resources for optimized storage and integrated traversal of annotated dense matrices. *Bioinformatics* **35**, 1427–1429 (2019).
57. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
58. Piccioni, F., Younger, S. T. & Root, D. E. Pooled lentiviral-delivery genetic screens. *Curr. Protoc. Mol. Biol.* **121**, 32.1.1–32.1.21 (2018).
59. Meyers, R. M. et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
60. Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10**, 733 (2014).
61. Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).

62. Yu, C. et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat. Biotechnol.* **34**, 419–423 (2016).
63. Pinello, L. et al. Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.* **34**, 695–697 (2016).
64. Niknafs, Y. S. et al. MiPanda: a resource for analyzing and visualizing next-generation sequencing transcriptomics data. *Neoplasia* **20**, 1144–1149 (2018).
65. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858 (1996).
66. Peng, J. & Gygi, S. P. Proteomics: the move to mixtures. *J. Mass Spectrom.* **36**, 1083–1091 (2001).
67. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
68. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292 (2006).
69. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).

Acknowledgements

We thank D. Bondeson, P. Tsvetkov, S. Corsello, U. Ben-David and T. Ouspenskaia for helpful discussions and critical reading of the manuscript. We thank M. Zhong for technical assistance with cloning and Z. Demere for assistance with CRISPR-sequencing. We thank D. Nusinow and S. Gygi for insights into identification of small peptides in proteomics datasets. We thank R. Tomaino for assistance with mass spectrometry at the Talpin Biological Mass Spectrometry Facility at Harvard Medical School. We thank J. Chen for assistance with the Slncky algorithm. We thank J. Gould for assistance with gene datasets. We thank I. Cheeseman for provision of DOX-inducible HeLa Cas9 cells. J.R.P. was supported by the Harvard K-12 in Central Nervous System tumors (grant 5K12 CA 90354-18). V.L and M.W.K. were supported by the National Institutes of Health (grants R01 HD073104 and RO1 HD091846 to M.W.K.).

Author contributions

J.R.P. and T.R.G. conceived the project, designed experimental approaches, supervised the study and analyzed data. J.R.P. selected ORFs for screening and developed ORF prioritization methods. J.R.P. and X.Y. designed and generated the ORF cDNA library. J.R.P performed ORF library screening, in vitro CRISPR experiments, siRNA experiments, immunoblots, cell culture assays and all GREP1 functional experiments. B.F. executed the arrayed ORF screen for L1000. O.M.E. and N.J.L. performed gene expression profiling and analyzed L1000 gene expression data. Z.J. contributed ORF predictions and assisted in analysis of ORF candidates. V.L., A.K., M.K. and J.R.P. performed protein evolutionary analyses and analyzed phylostratigraphy data. K.K., K.R.C. and J.D.J. performed proteomic identification of ORFs from datasets. J.R.P., F.P. and D.E.R. designed and analyzed CRISPR screens. T.G., D.A. and A.B. assisted with sgRNA design. A.G. and Z.K. performed cell line CRISPR screens. L.W., K.S., G.B. and J.A.R. performed pooled CRISPR screening. V.M.W. and J.M.D. analyzed pooled CRISPR screen data. J.M.D. performed comparative analyses of ORF CRISPR data with publicly available CRISPR screens. J.R.P. and T.R.G. wrote the manuscript draft and all authors contributed to editing it.

Competing interests

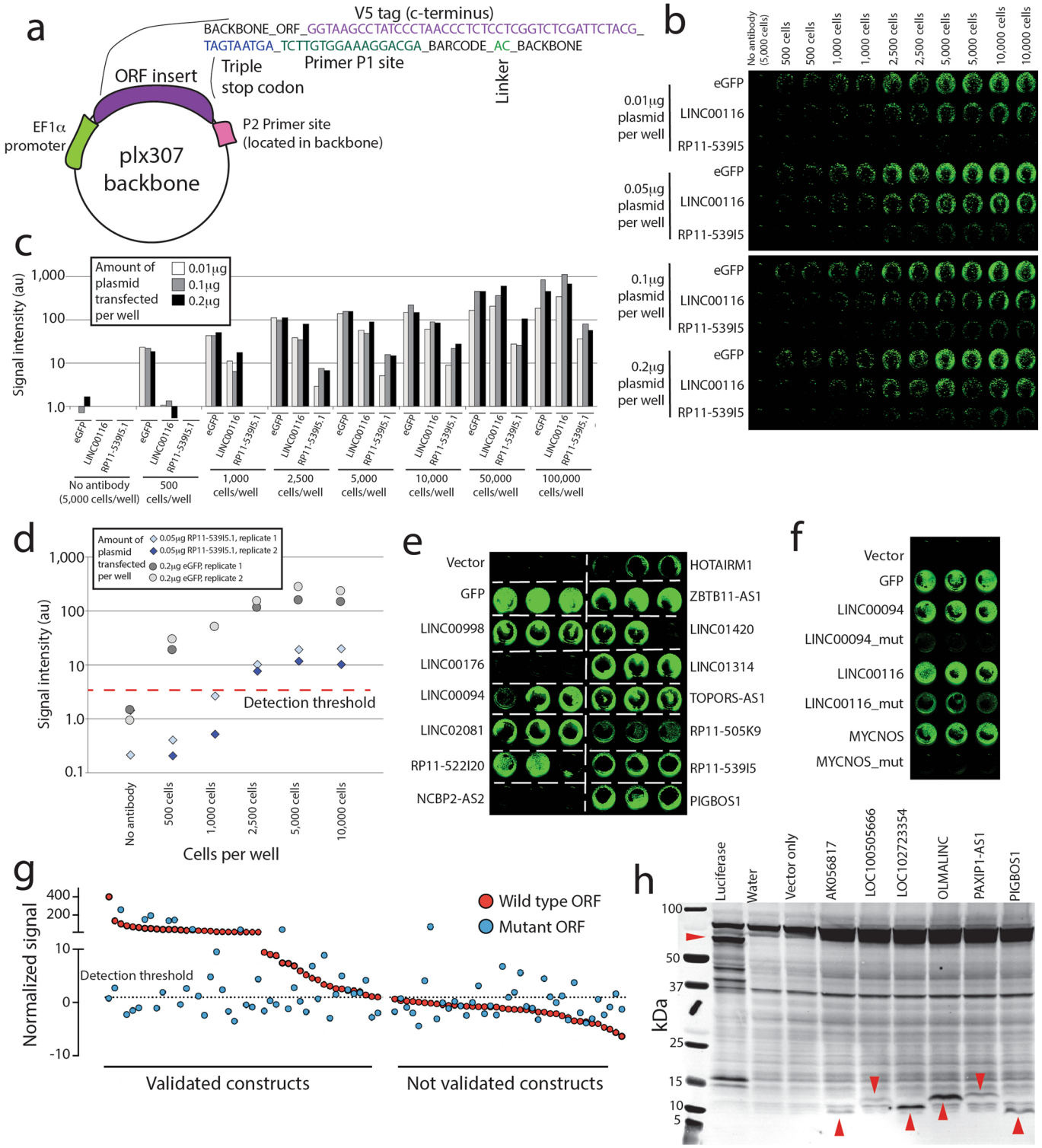
The authors declare no competing interests.

Additional information

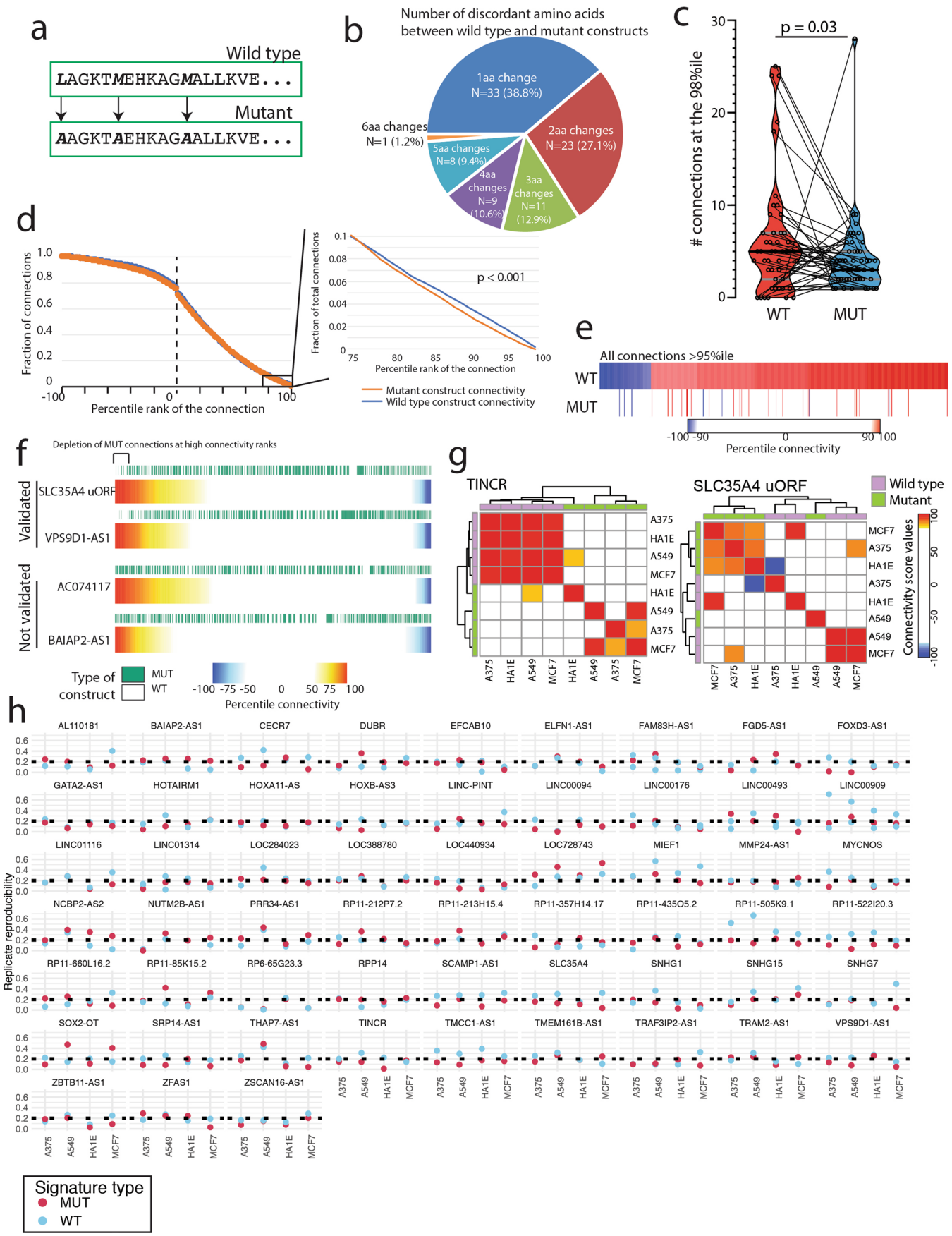
Extended data is available for this paper at https://doi.org/10.1038/s41587-020-00806-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-00806-2.

Correspondence and requests for materials should be addressed to T.R.G.

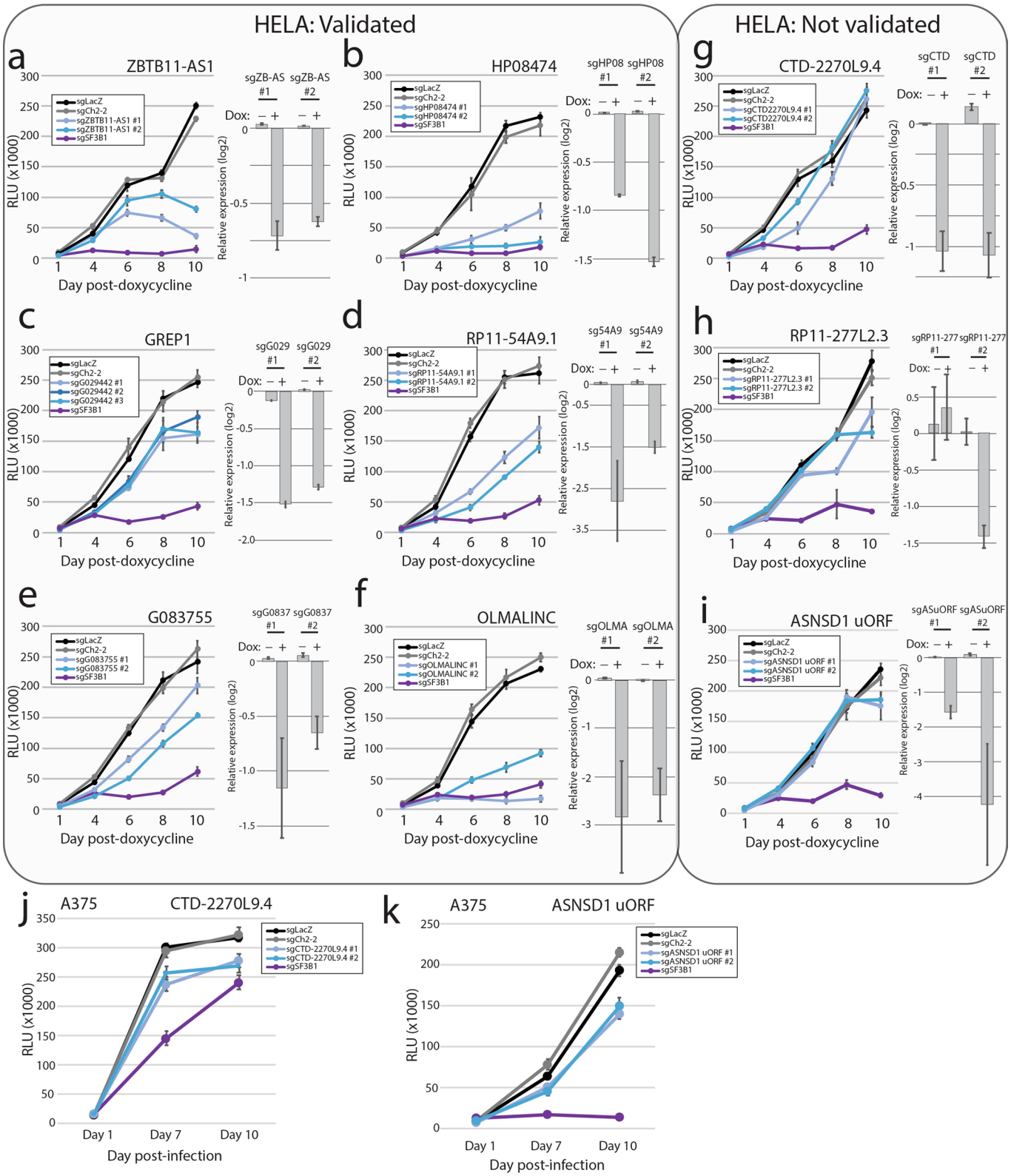


Extended Data Fig. 1 | Generation and validation of a non-canonical ORF cDNA library. a, Vector design and sequence details for the ORF library. The vector used is a modified version of the plx307 vector developed by the Genomic Perturbation Platform at the Broad Institute. **b**, Titration analyses of in cell western experiments. Three ORFs were chosen: eGFP (positive control), LINC00116 (high-expressing ORF), and RP11-539I5 (low expressing ORF). Increasing amounts of plasmid were transfected into increasing numbers of HEK293T cells as shown. **c**, Quantification the in cell western titration shown in **b**, demonstrating signal detection over noise and signal plateau. Signal was quantified using pixel density in the 800 nM green color channel. **d**, Replicate experiments assessing signal-to-noise thresholds for a low-expressing ORF transfected into HEK293T cells with a low DNA plasmid concentration, as well as a high-expressing ORF (eGFP) transfected into HEK293T cells at a high DNA plasmid concentration. **e**, Example in cell western data in triplicate experiments for selected ORFs. **f**, Abrogation of protein translation via mutation of the ORF for selected examples. **g**, A systematic evaluation of in cell western signal for wild type and mutant ORFs for all pairs. ORFs are separated into those with signal above the baseline threshold, and those without reproducible signal. **h**, An immunoblot showing *in vitro* transcription/translation of selected tag-free ORFs using a wheat germ lysate system. Red arrows indicate the translated ORFs. Results were repeated in two independent experiments.

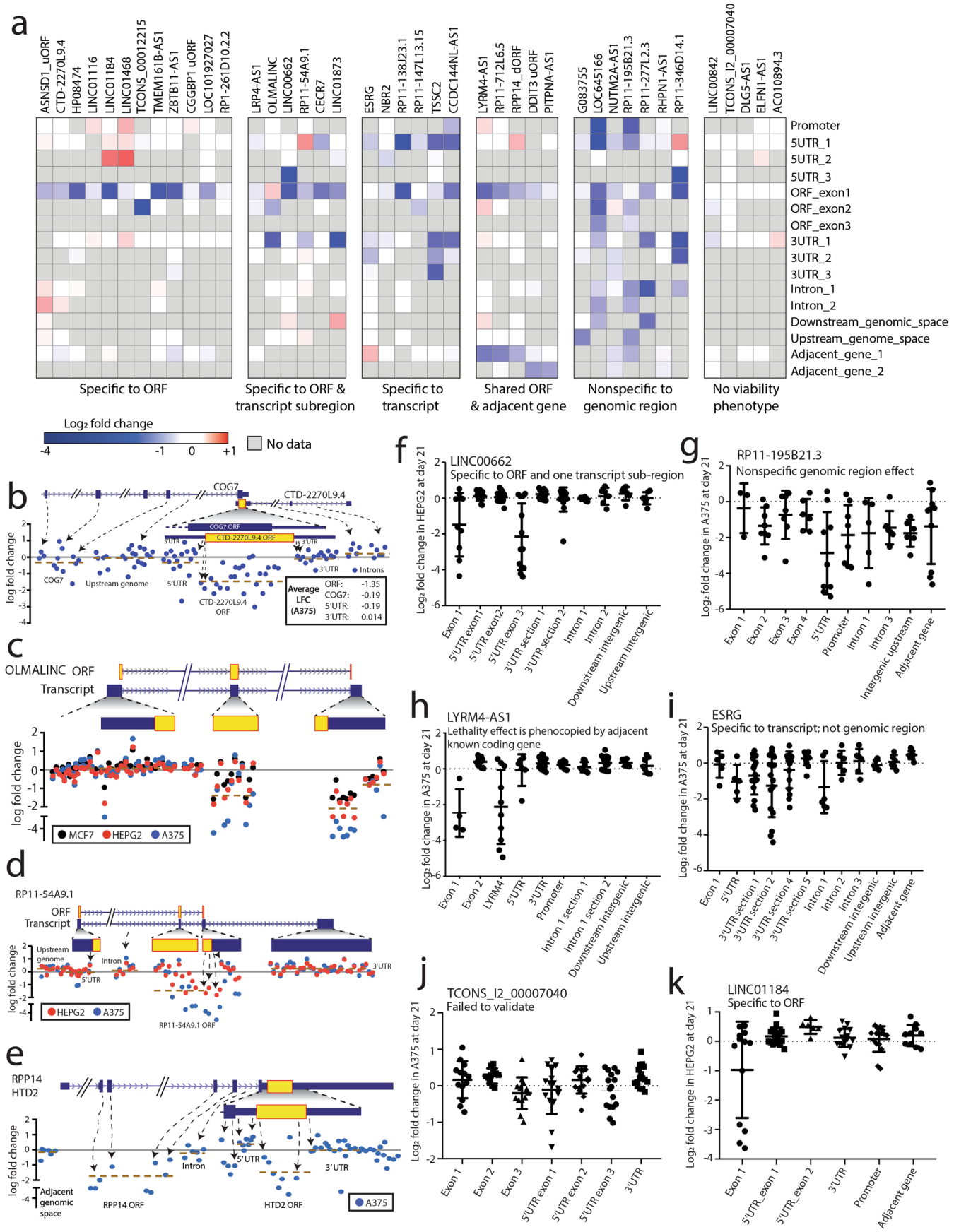


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Analysis of paired wild-type and mutant constructs in L1000 data. a, A strategy for ORF mutagenesis strategy in which the start codon and downstream methionines were mutated to alanine. The shown amino acid sequence is a fictional sequence. **b**, A pie chart showing the number and percentage of amino acids changed per ORF from the mutagenesis. **c**, A violin plot showing the number of Perturbational Class (PCL) connections made at the 98th percentile for matched mutant and wild type constructs (n = 47 for each, all data points are biologically independent experiments). P value by a two-tailed Wilcoxon matched pairs rank test. **d**, *Left*, the overall distribution of PCL connections across all ranks in wild type and mutant constructs (n = 19,012 independent comparisons for each). *Right*, an inset image of distribution of PCL connections at high connectivity, showing a bias in connections made with wild type compared to mutant constructs (n = 1,920 independent comparisons each). P value by a two-tailed Wilcoxon matched pairs rank test. **e**, All PCL connections in wild type constructs at either the >=95th percentile or <= -95th percentile, with the matched percentile connectivity in the mutant constructs. **f**, The distribution of percentile connectivity results in wild type or mutant constructs for the indicated genes. In brief, all ORF L1000 signatures were queried against all PCL classes and a percentile connectivity was generated for each individual cell line and for both wild type and mutant constructs. Cell line and construct data was then aggregated and ranked from highest to lowest connectivity. The rank positions of wild type and mutant ORFs were then plotted to reveal a depletion of mutant constructs at high connectivity scores. **g**, Two example heatmaps for the TINCR and SLC35A4 uORF plasmids showing clustering of PCL connectivity among wild type constructs that is not shared with mutant constructs. Purple bars denote wild type ORF experiments and green bars denote mutant ORF experiments. **h**, L1000 signature replicate reproducibility for all wild type and mutant pairs across all cell lines. All ORF signatures with at least one reproducible wild type signature are shown.

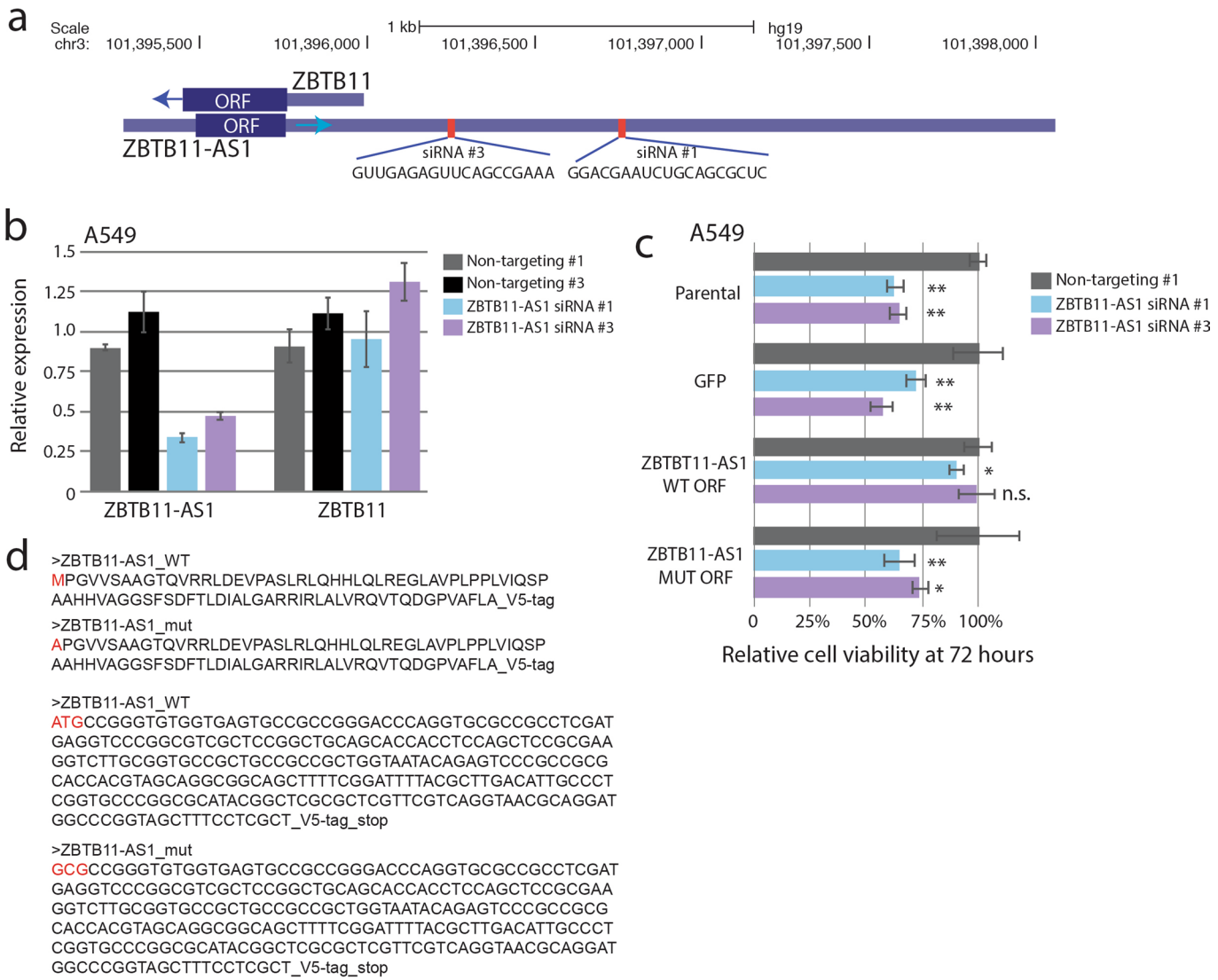


Extended Data Fig. 3 | Validation of CRISPR hits via manual assays. a–i, CRISPR assays using doxycycline-inducible Cas9 in HeLa cells. Targets are divided in ones that validated and ones that did not. For each experiment, the right-set panel is qPCR data of expression 96 hours after induction of Cas9 with doxycycline. **a**) *ZBTB11-AS1* **b**) *HP08474* **c**) *GREP1* **d**) *RP11-54A9.1* **e**) *G083755* **f**) *OLMALINC* **g**) *CTD-2270L9.4* **h**) *RP11-277L2.3* **i**) *ASNSD1 uORF*. **j-k**, CRISPR assays using stably-expressing A375 Cas9 cells. **j**) *CTD-2270L9.4* **k**) *ASNSD1 uORF*. For all data in this figure, n = 6 technical replicates for each data point. Error bars represent standard deviation. Data was also acquired a 3 independent biological replicates based on doxycycline dose level (0.2 ug/mL, 1.0 ug/mL and 2.0 ug/mL doxycycline, as well as 0 ug/mL doxycycline). The data shown are the 1.0 ug/mL dosing level, with similar results observed for the 0.2 ug/mL and 2.0 ug/mL doxycycline dosing levels.

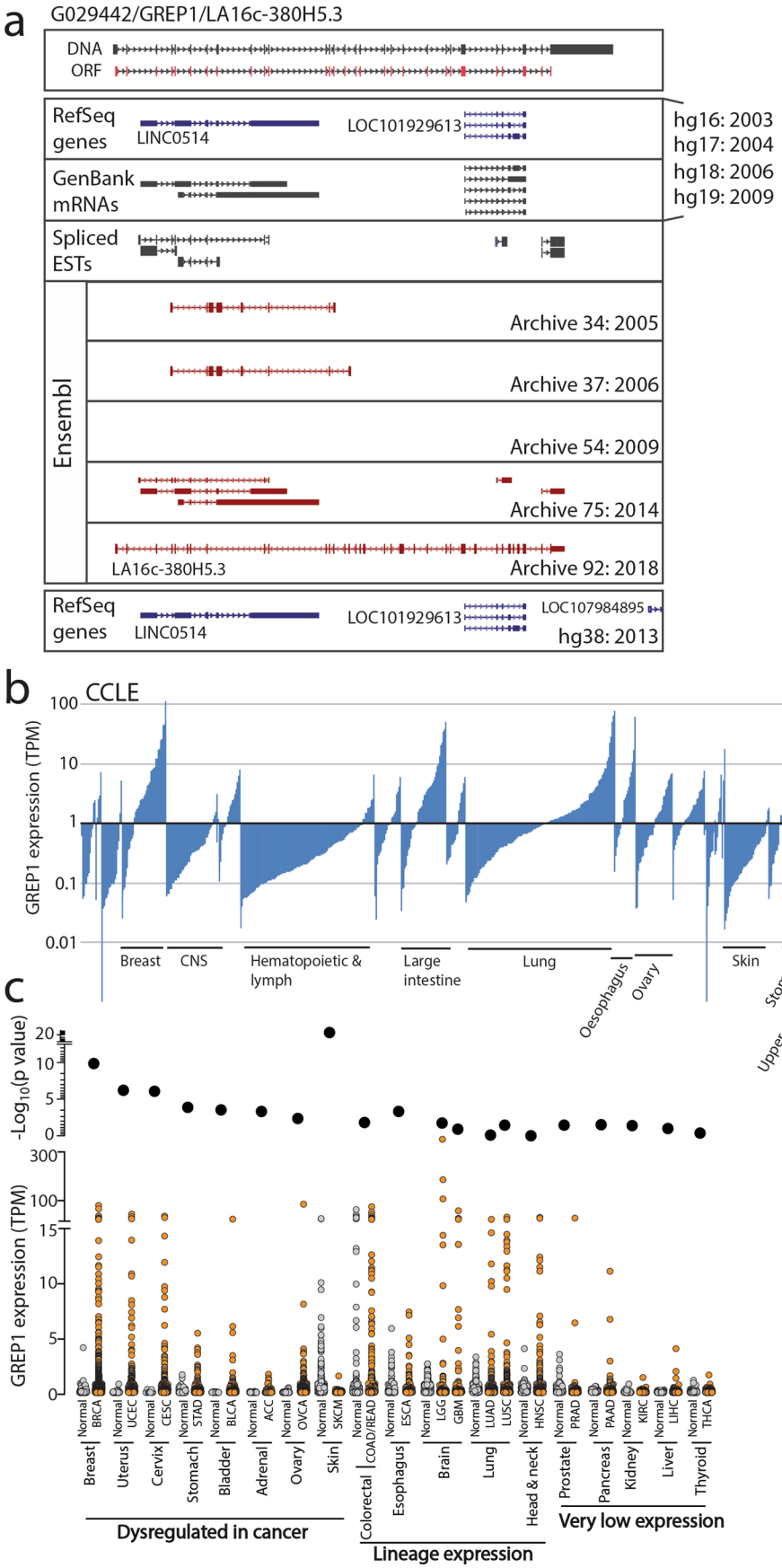


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Tiling CRISPR assays to elucidate functional non-canonical ORFs. a, A heatmap showing log fold change viability loss at Day +21 in the secondary CRISPR screen for the indicated non-canonical ORFs tested by multiple tiling sgRNA regions. **b-e**, Examples of non-canonical ORFs with a CRISPR tiling phenotype. **b-e**) Graphical representation of tiling CRISPR assays in which each dot represents an individual sgRNA. sgRNAs are mapped to their genomic loci and the genomic region of the tiling assay is shown. The location of the putative non-canonical ORF is shown in the gene annotation above. **b**) *CTD-2270L9.4* **c**) *OLMALINC* **d**) *RP11-54A9.1* **e**) *RPP14* dORF / *HTD2*. **f - k**, Representative sgRNA log fold change data for the indicated transcripts. Each tiling experiment is classified as indicated. **f**) *LINC00662* **g**) *RP11-195B21.3* **h**) *LYRM4-AS1* **i**) *ESRG* **j**) *TCONS_I2_00007040* **k**) *LINC01184*.

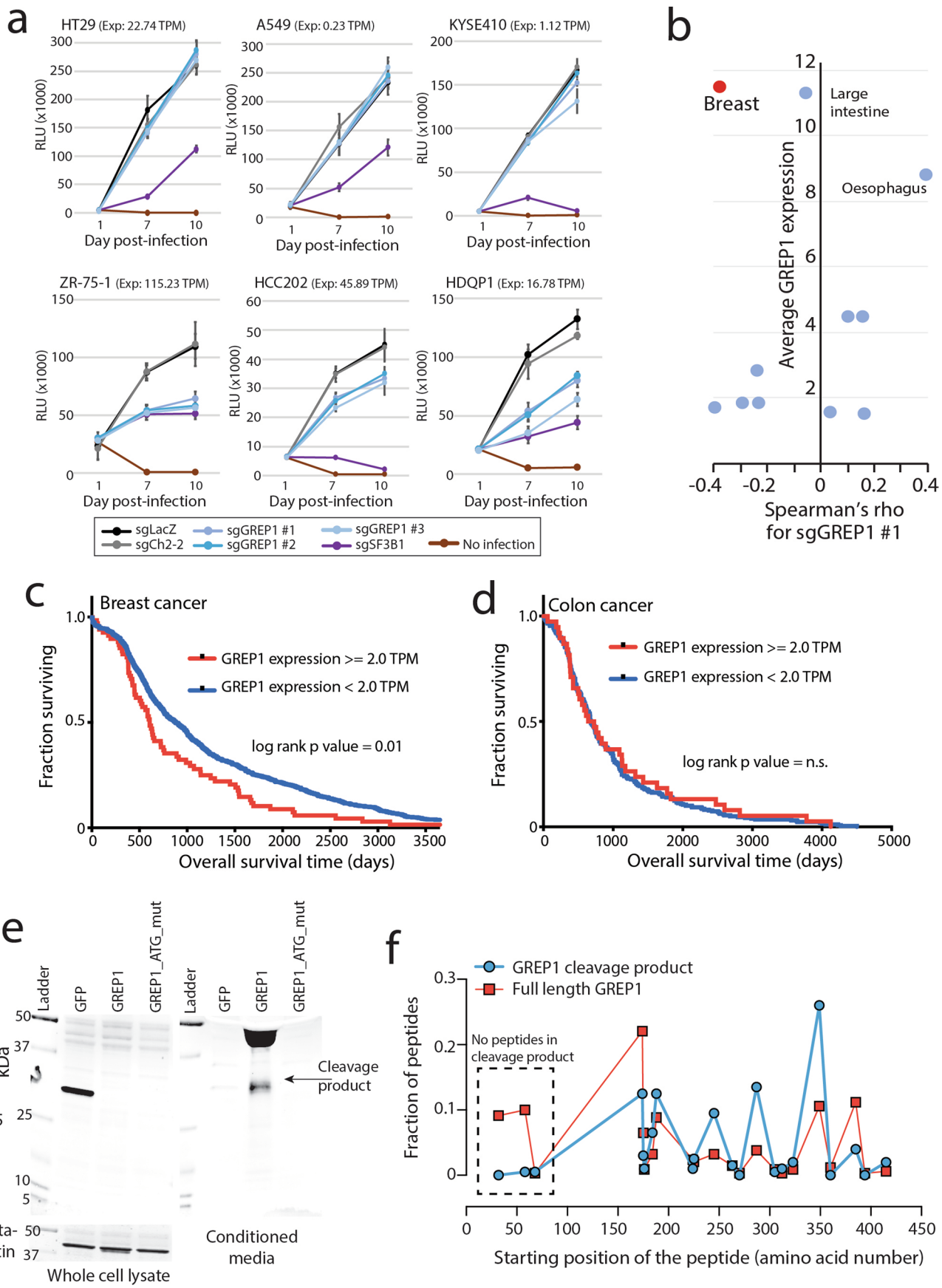


Extended Data Fig. 5 | Specific siRNA knockdown of ZBTB11-AS1 mRNA transcript causes a viability phenotype which is specifically rescued by the wild type ZBTB11-AS1 ORF. a, A schematic showing the genomic location and sequences for the two siRNAs used for *ZBTB11-AS1*. **b**, mRNA expression levels for *ZBTB11-AS1* or *ZBTB11* transcripts 48 hours after siRNA knockdown of *ZBTB11-AS1* in A549 cells. N = 3 independent replicates for all conditions. Barplots represent mean ± standard deviation. **c**, Relative cell viability of A549 cells treated with *ZBTB11-AS1* siRNAs at 72 hours. Parental A549 cells were used along with A549 cells expressing cDNAs for GFP, wild type *ZBTB11-AS1* ORF sequence, or mutant *ZBTB11-AS1* ORF lacking translational start sites. Only the wild-type *ZBTB11-AS1* ORF sequence rescues the viability phenotype. N = 6 independent replicates for all conditions. Barplots represent mean ± standard deviation. **d**, DNA and amino acid sequences of the wild type and mutant *ZBTB11-AS1* ORF cDNAs. *p < 0.05, **p < 0.01. n.s., non-significant. For P values: Parental, non-targeting vs siRNA #1 P < 0.0001, non-targeting vs siRNA #2 P < 0.0001; GFP, non-targeting vs siRNA #1 P = 0.0008, non-targeting vs siRNA #2, P < 0.0001; WT ORF, non-targeting vs siRNA #1 P = 0.04, non-targeting vs siRNA #2 P = 0.83; MUT ORF, non-targeting vs siRNA #1 P = 0.001, non-targeting vs siRNA #2 P = 0.02. P values by a two-tailed Student's T test.



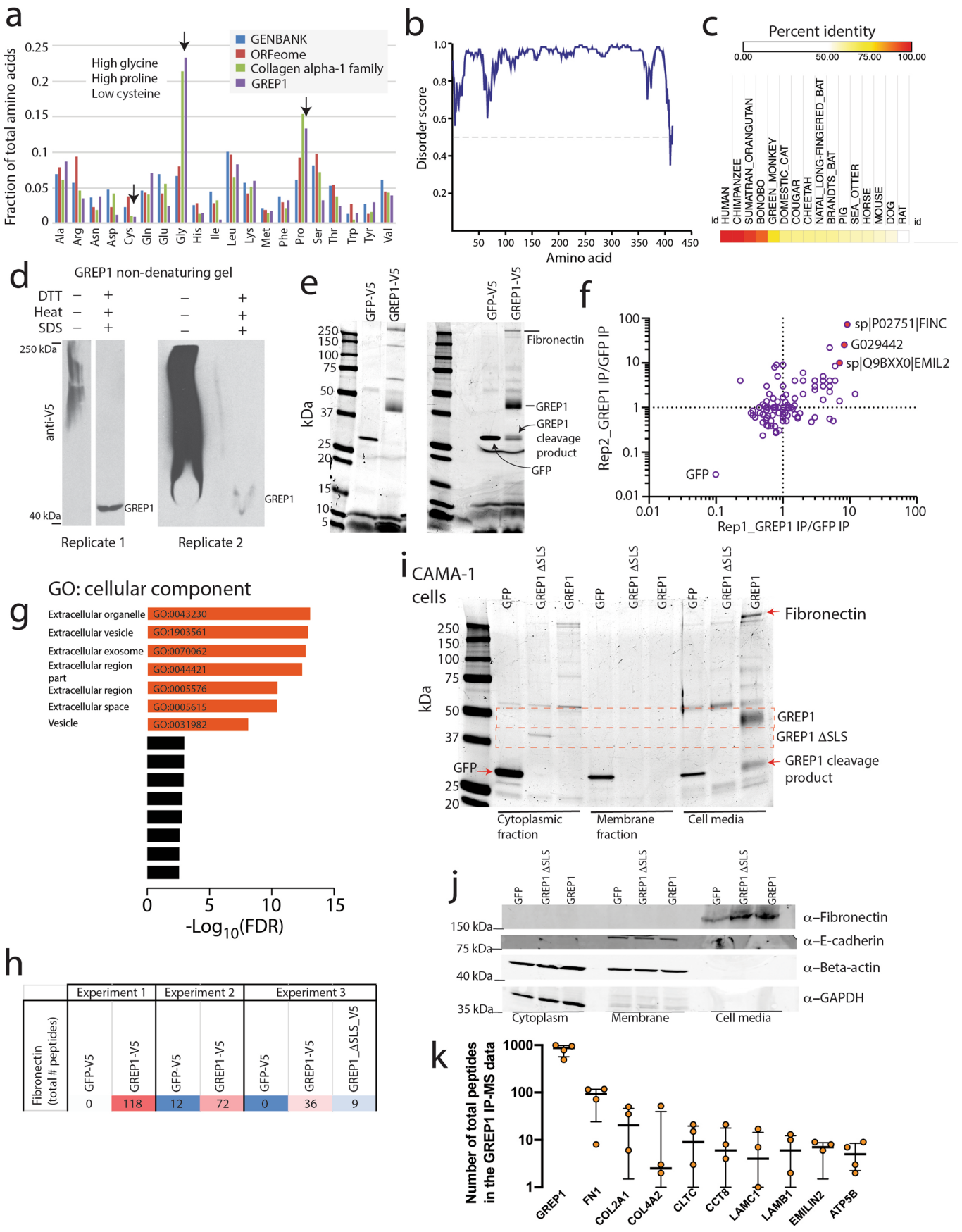
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | The GREP1 locus and expression. a, A schematic representation of the *GREP1* gene structure and the annotation of this locus in the indicated databases. The year of release for each database is indicated. **b**, mRNA expression level of *GREP1* across tumor lineages in the Cancer Cell Line Encyclopedia. The Y axis is in a log10 scale. **c**, mRNA expression of *GREP1* across tumor types using TCGA and GTex data. A two-tailed Student's t-test was used to calculate significance of change between normal and cancer tissues. Cell lineages are grouped according to whether *GREP1* expression is specifically modulated in cancer, universally expressed as a lineage gene, or not robustly expressed in the indicated lineage.



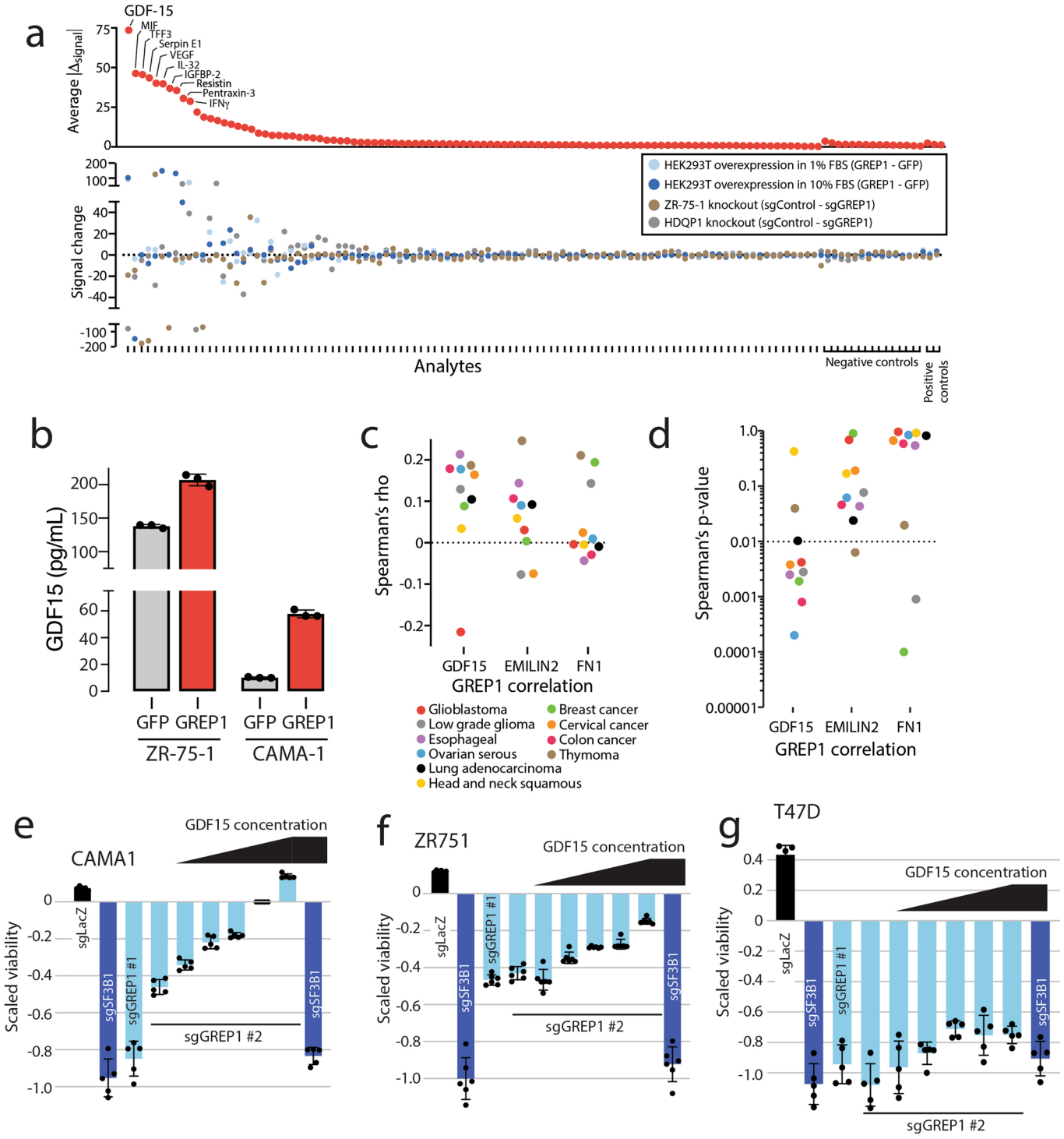
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | GREP1 is implicated in cell proliferation and breast cancer patient outcomes. a, Cell viability curves following *GREP1* knockout in three sensitive and three insensitive cell lines. *GREP1* expression in the Cancer Cell Line Encyclopedia is indicated in transcripts per million (TPM) **b**) A scatter plot showing lineage-specific correlation between cell viability and *GREP1* mRNA expression on the X axis with the average *GREP1* expression level on the Y axis. **c**, Overall survival for breast cancer patients in the TCGA database stratified by *GREP1* expression. N = 1,036 individual patients. N = 969 *GREP1*-low and N = 67 *GREP1*-high patients. Significance by a one-sided log-rank P value. **d**, Overall survival for colon cancer patients in the TCGA database stratified by *GREP1* expression. N = 296 individual patients. N = 38 *GREP1*-high and N = 258 *GREP1*-low patients. Significance by a one-sided log-rank P value. **e**, Immunoblot of V5-tagged GREP1 or GFP in HEK293T cells in both whole cell lysate and conditioned media. A mutant GREP1, in which translational start sites were mutated to alanine, lacks protein translation initiation ability. Results were repeated in three independent experiments. **i**, Abundance of mass spec peptides detected in the full length GREP1 or cleavage product GREP1 proteins. Peptide abundance is represented as a fraction of total peptides detected. All error bars represent standard deviation.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | GREP1 is associated with the extracellular matrix. a, Total fraction of amino acid usage in the ORFeome, GENBANK, GREP1, and the Collagen alpha-1 family. Sequence similarities between GREP1 and the collagen family are indicated. **b**, Predicted disorder score for the GREP1 amino acid sequence. **c**, Amino acid conservation for detected homologs of GREP1 in the indicated species. **d**, Non-denaturing native western blot of GREP1 in conditioned media from HEK293T cells expressing V5-tagged GREP1. **e**, Representative Commassie-stained gels for immunoprecipitation of GREP1 from the conditioned media of HEK293T cells. Two representative biological replicates are shown. **f**, Enrichment of extracellular matrix proteins in the IP-MS data for GREP1 compared to IP-MS data for GFP. **g**, Gene Ontology Cellular Component analysis of proteins >= 2 fold enriched in GREP1 immunoprecipitation compared to GFP immunoprecipitations. **h**, IP MS total peptide count for fibronectin shown for three separate experiments. **i**, Commassie stain of V5 immunoprecapitation of V5-tagged GFP, GREP1 del_SLS or GREP1 constructs expressed in CAMA-1 cells following fractionation of cell lysate into cytoplasmic, membrane and cell media components. Results were repeated in 2 independent experiments. **j**, Western blot of endogenous fibronectin, E-cadherin, beta-actin and GAPDH in cell lysate or cell culture media for CAMA-1 cells expressing GFP, GREP1 del_SLS or GREP1 constructs as in panel **i**. Results were repeated in two independent experiments. **k**, IP mass spectrometry data showing the total peptide count for GREP1 and other top-scoring proteins following IP of V5-tagged GREP1 in HEK293T, ZR-75-1, and CAMA-1 cells. N = 4 independent IP MS experiments. Lines represent median ± interquartile (25-75%) range.



Extended Data Fig. 9 | GREP1 regulates GDF15 in vitro and correlates with GDF15 expression in patient tumor tissues. a, Cytokine profiling in HEK293T cells with transient ectopic GREP1 or GFP overexpression, ZR-75-1 cells with stable *GREP1* knockout, or HDQP1 cells with stable *GREP1* knockout. The change in signal abundance was calculated for each control/GREP1 pair. To rank cytokines, the average of the absolute values for the individual signal changes was plotted. **b**, GDF15 abundance by ELISA in ZR-75-1 and CAMA-1 cells overexpressing a *GREP1* or *GFP* cDNA plasmid. N = 3 technical replicates. N = 2 independent experiments performed, with representative results shown. **c**, Spearman's rho for *GREP1* expression correlation with *GDF15*, *EMILIN2*, or *FN1* in the indicated TCGA datasets. **d**, Spearman's p value for the *GREP1* correlation coefficient for *GREP1* correlation with *GDF15*, *EMILIN2*, or *FN1* in the indicated TCGA datasets. **e-g**, Recombinant GDF15 partially rescues *GREP1* knockout. CAMA-1, ZR-75-1 or T47D Cas9 cells were infected with the indicated sgRNAs. 24 hours after infection, cells were treated with vehicle control or increasing concentration of recombinant human GDF15 as shown. Relative abundance was measured 7 days after infection. N = 5 for all conditions in panel **e**. N = 6 for all conditions in panel **f**. N = 5 for all conditions in panel **g**. All error bars represent standard deviation. Two independent experiments were performed for panels **e–g**.

nature research

Corresponding author(s): Todd R. Golub

Last updated by author(s): 11/23/2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All software used in data collection were described in the Methods section of the paper. No specialized software or custom code was used for data collection.
Data analysis	ORF candidates were analyzed with: PhastCons score (version hg19_20110909) for 100 placental mammals, PhyloCSF (https://github.com/mlin/PhyloCSF/wiki) for 29 mammal alignment, Pfam web server (http://pfam.xfam.org/search#tabview=tab1), the NCBI Conserved Domain finder (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), Slncky (https://slncky.github.io), Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index), SignalP v4.1 (http://www.cbs.dtu.dk/services/SignalP-4.1/). TimeTree (http://www.timetree.org), Cancer Cell Line Encyclopedia RNA expression data (https://portals.broadinstitute.org/ccle), DISOPRED3 available via PsiPred (http://bioinf.cs.ucl.ac.uk/psipred/), Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/), the MiPanda v1.0 gene expression portal (http://www.mipanda.org). Mass spectrometry peptides were processed via: Spectrum Mill MS Proteomics Workbench v6.0, an in-house SMDataCrawler code package,(9https://proteomics.broadinstitute.org/millhtml/SM_instruct/SMallman.htm). L1000 data analysis code and pre-processed data are available via GitHub https://github.com/cmap/cmapM. There is additional information about this database and tools at http://clue.io/connectopedia. L1000 data was further analyzed via: the 'tidyverse' suite36 of R packages (v1.2.1), the 'cmapR' package37 (v1.0.1) in R v3.5.0 (R Core Team 2018), in-house code available through github (https://github.com/johnprensner/smORF_analyses). sgRNA sequences were analyzed for off-target effects with Cas-OFFinder v1.0 (http://www.rgenome.net/cas-offinder/). CRISPR screen data was analyzed using: Bowtie v2 for read alignment (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) and analyzed with STARS v1.3 (python code, http://www.broadinstitute.org/rnai/public/software/index) and CERES v1.0 (R package, https://depmap.org/ceres, https://github.com/cancerdatasci/ceres). CRISPRseq data was analyzed with CRISPResso v2 (http://crispresso.pinellolab.partners.org). Cytokine array profiling images were analyzed with ImageJ v2.0.0 (https://imagej.nih.gov/ij/index.html) For all data types, PRISM Graphpad (version 8) was used for visualization.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Processed data for CRISPR screens (in Figure 3 and Figure 4d) are available in Supplementary Tables 22, 27, and 32. Raw data are available in the Source Data files accompanying this manuscript as well as through the NCBI Sequence Read Archive (SRA) at: SRR13126801, SRR13128583, SRR13132373, SRR13142215 and SRR13142421. Comparison of CRISPR screen data with prior data was completed using the DepMap_public_19Q4 data release for Achilles screens at https://depmap.org/portal/download.

Mass spectrometry data relating to Figure 1 are available in Supplementary Table 14. Raw MS spectra are available through the original datasets at:
- CPTAC2_BRCA_prosp, PMID 33212010, https://cptac-data-portal.georgetown.edu/study-summary/S060
- CPTAC2_COAD_prosp, PMID 31031003, https://cptac-data-portal.georgetown.edu/study-summary/S045
- CPTAC3_ccRCC, PMID 31675502, https://cptac-data-portal.georgetown.edu/study-summary/S050
- CPTAC3_LUAD, PMID 32649874, https://cptac-data-portal.georgetown.edu/study-summary/S056
- CPTAC3_PTRC_DP1, PMID 31988290, https://cptac-data-portal.georgetown.edu/study-summary/S051
- CPTAC3_UCEC, PMID 32059776, https://cptac-data-portal.georgetown.edu/study-summary/S053
- HLA_Abelin, PMID 28228285, ftp://massive.ucsd.edu/MSV000080527
- HLA_Ouspenskaia, PMID TBD, ftp://massive.ucsd.edu/MSV000084787
- HLA_Sarkizova, PMID 31844290, ftp://massive.ucsd.edu/MSV000084172/; ftp://massive.ucsd.edu/MSV000080527; ftp://massive.ucsd.edu/MSV000084442/
- CPTAC Medulloblastoma, PMID 30205044, ftp://massive.ucsd.edu/MSV000082644
- PeptideAtlas database, http://www.peptideatlas.org

L1000 data relating to Figure 2 and Supplementary Figures 8 & 9 is available through the NIH LINCS program and at https://clue.io/data and utilized the Touchstone reference database at https://clue.io/command. Raw ORF L1000 data is available at https://clue.io/data under the ORFeome Library Characterization (OFL) heading. The website lincsproject.org provides information about the LINCS consortium, including data standards.

Other databases employed to analyze ORF candidates are:
- Pfam web server (http://pfam.xfam.org/search#tabview=tab1)
- the NCBI Conserved Domain finder (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi),
- Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)
- SignalP v4.1 (http://www.cbs.dtu.dk/services/SignalP-4.1/)
- TimeTree (http://www.timetree.org)
- Cancer Cell Line Encyclopedia RNA expression data (https://portals.broadinstitute.org/ccle)
- DISOPRED3 available via PsiPred (http://bioinf.cs.ucl.ac.uk/psipred/), Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/)
- the MiPanda v1.0 gene expression portal (http://www.mipanda.org).
- CCLE copy number data from the 2013-12-03 segmentation was downloaded from https://depmap.org/portal/download.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No predetermined sample sizes were employed.
Data exclusions	No data was excluded from the analyses.
Replication	L1000 data was generated by biological triplicate; all data were used. CRISPR screens were analyzed via sequencing of three technical replicates; all data were used. ORF in cell western data was generated as biological triplicate; all data was used. Individual validation CRISPR assays were performed in biological duplicate with 4 technical replicates each. Non-denaturing western blots were performed in two biological replicates. Immunoprecipitation/mass spectrometry were performed in biological replicate. Conditioned media and GDF15 rescue experiments were performed in biological duplicate with 4 technical replicates each. ELISA assays were performed in biological duplicate with 3 or 4 technical replicates each, as indicated in the figure. Cytokine arrays were performed with technical replicates. qPCR data were performed with biological duplicate samples with 3 technical replicates each. All attempts for experimental replication were successful.
Randomization	For comparison of ORF CRISPR data to Dependency Map data, we selected 500 random genes in the Dependency Map in a random cell line as well as in the 8 cell lines used in our screen, in order to generate a reference distribution of values.
Blinding	Investigators were not blinded to experimental design. Blinding was not possible as this was an exploratory study and so blinding is

Blinding | impossible or unlikely to affect the results.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

V5, clone D3H8Q, species Rabbit, Monoclonal antibody, dilution 1:2000, catalogue number 13202S, manufacturer: Cell Signaling Technology, conditions: 4C overnight

ZBTB11, specieis Rabbit, Polyclonal antibody, dilution 1:1000, catalogue number A303-240A-M, manufacturer: Bethyl Laboratories, conditions: 4C overnight

Beta-Actin, species Mouse, Monoclonal antibody, dilution 1:4000, catalogue number: A5316, manufacturer: Sigma-Aldrich, conditions: 4C overnight

Goat anti-mouse secondary, species Goat, dilution 1:5000, catalogue number 926-32210, manufacturer: LI-COR, conditions: 20C for 1 hour

Goat anti-rabbit secondary, species Goat, dilution 1:5000, catalogue number 926-68021, manufacturer: LI-COR, conditions: 20C for 1 hour

Goat Anti-rabbit HRP secondary antibody, dilution 1:10000, catalogue number 7074S, manufacturer: Cell Signaling Technology, conditions 20C for 1 hour.

Validation

ZBTB11 antibody was validated in human HeLa cells by assessing CRISPR knockout cells and identifying a band of the correct protein size which showed decrease upon knockout. Manufacturer information at https://www.bethyl.com/product/A303-239A/ZBTB11+Antibody.

V5 antibody was validated in human HEK293T cells by assessing the antibody with and without transfection of a GFP-V5 plasmid, which showed a protein band of the correct size. Manufacturer information at: https://www.cellsignal.com/products/primary-antibodies/v5-tag-d3h8q-rabbit-mab/13202?site-search-type=Products&N=4294956287&Ntt=13202s%2C&fromPage=plp&_requestid=1764878

Beta-actin antibody was validated in human HEK293T cells by assessing a single protein band of the correct size in native lysates, with no off-target bands identified. Manufacturer information at: https://www.sigmaaldrich.com/catalog/product/sigma/a5316?lang=en&region=US

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

All Cas9-derived cell lines were obtained from the Broad Institute Genomics Perturbation Platform as listed here. All cell lines stably expressed the plx311 Cas9 construct (https://www.addgene.org/96924/) and were verified for Cas9 activity prior to usage via the Genomics Perturbation Platform. Cell lines include:
-- Primary/Secondary CRISPR screen: PC3 Cas9, HEPG2 Cas9, HeLa Cas9, HA1E Cas9, A549 Cas9, A375 Cas9, MCF7 Cas9, HT29 Cas9
-- GREP1 experiments: HT29 Cas9, A375 Cas9, A549 Cas9, HCC15 Ca9, KYSE410 Cas9, KYSE510 Cas9, MIAPACA2 Cas9, SNU503 Cas9, SW837 Cas9, AU565 Cas9, ZR751 Cas9, MCF7 Cas9, HCC202 Cas9, HDQP1 Cas9, JIMT1 Cas9, HCT116 Cas9, MDAMB468 Cas9, MDAMB231 Cas9, HCC1954 Cas9.

Parental (non-Cas9) cell lines used for GREP1 experiments include: MCF7, ZR-75-1, BT474, CAMA-1. These were obtained from ATCC.

HEK293T cells were obtained from ATCC.

	HeLa with doxycycline-inducible Cas9 were obtained from Iain Cheeseman at MIT, who purchased the original HeLa cells from ATCC.
Authentication	Cell line authentication for high-throughput screens was maintained by the Genomic Perturbation Platform.
Mycoplasma contamination	Cell lines were tested for mycoplasma upon receipt and were negative.
Commonly misidentified lines (See ICLAC register)	No commonly mis-identified cell lines were used in this study.