nature portfolio

https://doi.org/10.1038/s41588-021-00840-z

Supplementary information

A genome-wide atlas of co-essential modules assigns function to uncharacterized genes

In the format provided by the authors and unedited

Supplementary Information

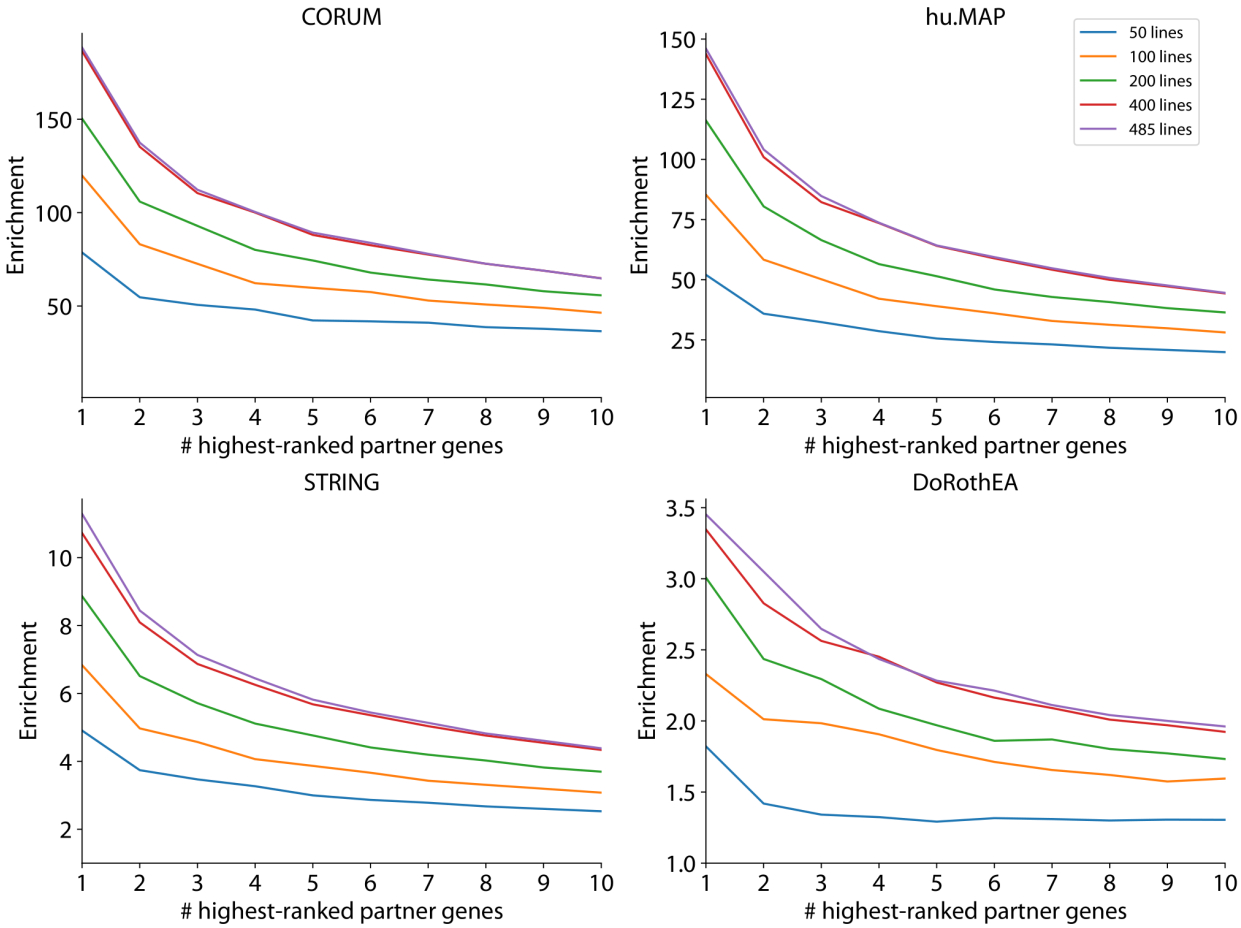
A genome-wide atlas of co-essential modules assigns function to uncharacterized genes

Michael Wainberg[1,2,5], Roarke A. Kamber[1,5], Akshay Balsubramani[1,5], Robin M. Meyers[1], Nasa Sinnott-Armstrong[1], Daniel Hornburg[1], Lihua Jiang[1], Joanne Chan[1], Ruiqi Jian[1], Mingxin Gu[1], Anna Shcherbina[1], Michael M. Dubreuil[1], Kaitlyn Spees[1], Wouter Meuleman[3], Michael P. Snyder[1], Michael C. Bassik[1,4]*, Anshul Kundaje[1,2]*

1. Department of Genetics, Stanford University, Stanford, California, USA
2. Department of Computer Science, Stanford University, Stanford, California, USA
3. Altius Institute for Biomedical Sciences, Seattle, Washington, USA
4. Chemistry, Engineering, and Medicine for Human Health (ChEM-H), Stanford University, Stanford, California, USA.
5. These authors contributed equally

*Correspondence to akundaje@stanford.edu or bassik@stanford.edu

Supplementary Fig. 1: Benchmarking after subsampling cell lines



Performance of GLS with PCA-based bias correction on the full 485 lines from DepMap as well as subset to only the first 50, 100, 200 and 400 lines. Using the full 485 lines barely improves performance over just using 400 lines, suggesting a saturation effect.

Supplementary Table 1: Co-essential and co-expressed partners of *TP53*, *KRAS* and *BRCA1.*

Gene	Significant co-essential partners (*p*-value, direction)			Top co-expressed partners (correlation)	
TP53	***USP28***	**3×10^{-28}**	**+**	***CPEB4***	**-0.32**
	CDKN1A	**1×10^{-24}**	**+**	*SYNJ1*	-0.32
	TP53BP1	**5×10^{-23}**	**+**	*PFN1*	0.32
	MDM2	**4×10^{-18}**	**-**	*RNPEP*	0.31
	CHEK2	**2×10^{-16}**	**+**	***RCC2***	**0.31**
	ATM	**1×10^{-14}**	**+**	*CNN2*	0.30
	PPM1D	**2×10^{-9}**	**-**	*SERINC1*	-0.30
	XPO7	6×10^{-8}	+	*FAM126B*	-0.30
	UBE2K	**9×10^{-6}**	**+**	***FOXM1***	**0.30**
	CNOT2	1×10^{-5}	-	*CHST14*	0.30
KRAS	***RAF1***	**1×10^{-12}**	**+**	*ZDHHC20*	0.55
	DOCK5	**4×10^{-7}**	**+**	*PTAR1*	0.49
	SHOC2	**2×10^{-6}**	**+**	*MATR3*	0.49
	ERGIC2	4×10^{-6}	+	*SPCS3*	0.47
	TM7SF3	8×10^{-6}	+	*SUZ12*	0.47
BRCA1	***BARD1***	**1×10^{-25}**	**+**	*KIF14*	0.67
	PALB2	**2×10^{-8}**	**+**	*MCM10*	0.65
	RPL21	1×10^{-6}	-	*KIF11*	0.63
	BRCA2	**2×10^{-5}**	**+**	***FANCD2***	**0.63**
	HIST2H2AA3	4×10^{-5}	-	*NCAPH*	0.62
	HEY1	6×10^{-5}	+	*ARHGAP11A*	0.62

Significant GLS co-essential versus top co-expressed partners of *TP53*, *KRAS* and *BRCA1*. Genes in bold have strong evidence of being part of the same pathway.

Supplementary Notes

Comparison of co-essentiality and co-expression in detecting interactions for cancer drivers

Co-essentiality outperformed co-expression in detecting interactions for a number of key cancer drivers, perhaps because our co-essentiality network was generated from cancer cell lines, while the co-expression network was not cancer-specific. For example, 8 of *TP53*'s 10 significant co-essential partners are known interactors (*USP28*, *CDKN1A*, *TP53BP1*, *MDM2*, *CHEK2*, *ATM*, *PPM1D*, *UBE2K*) compared with only 3 of the top 10 co-expressed partners in COXPRESdb (**Supplementary Table 1**). For *KRAS*, 3 of 5 significant co-essential partners are known interactors versus none of the top 5 co-expressed partners; for *BRCA1*, 3 of 6 co-essential partners are known interactors versus 1 of 6 for co-expression.

We performed a systemic analysis of driver gene co-essentiality using 82 curated oncogenes and 63 tumor suppressors[1]. Co-essential pairs were enriched for both types of driver genes, but especially for tumor suppressors. Only 0.11% of all gene pairs are co-essential, compared to 0.90% of tumor suppressor pairs (8.4× enrichment, Fisher $p = 7 \times 10^{-11}$), and 0.45% of oncogene pairs (4.2× enrichment, Fisher $p = 1 \times 10^{-5}$).

Generation of co-essential modules using ClusterOne

One major parameter affecting ClusterONE module detection quality is the module density *d*, which determines (on a 0-to-1 scale) how strong the internal connections within a cluster must be relative to the connections on the edge of the cluster between members and non-members. Inferring networks at multiple scales has been observed to provide the most complete picture of biological systems[2,3]; accordingly, we found that lower values of *d* (e.g. 0.2) led to larger modules and better performance on STRING, while larger values (e.g. 0.9) led to smaller

modules and better performance at recapitulating complexes and physical interactions from CORUM and hu.MAP, with intermediate values (e.g. 0.5) striking a balance between the two (**Extended Data Fig. 5**). Because modules generated with different values of *d* capture different types of biological pathways, we generated a combined list of modules using *d* = 0.2 (N = 168), 0.5 (N = 1892) and 0.9 (N = 3169) (**Supplementary Data 2**).

Benchmarking of co-essential modules

We performed thorough benchmarking to demonstrate how our modules substantially improved upon predictions based on Pearson correlations between individual gene pairs. Of the 9,616 genes with GO term predictions from our modules and at least one GO term themselves, 2480 (25.8%) had one of the predicted GO terms (a notably high percentage given the substantial incompleteness of GO terms and rigorous 100-fold enrichment threshold), 55% more genes than the 1600 (16.6%) successfully predicted by modules of the same size constructed from the top Pearson correlation partners (Fisher $p = 3 \times 10^{-36}$). At an even more stringent 1000-fold GO enrichment threshold, this difference widened from 55% to 269% (our modules: 395 of 1427 genes, 27.7%; same-size Pearson modules: 107 of 1427 genes, 7.5%; Fisher $p = 1 \times 10^{-34}$). Of the 1600 genes correctly predicted by Pearson, only 193 were incorrectly predicted by our modules, whereas of the 2480 genes correctly predicted by our modules, a full 1073 were incorrectly predicted by Pearson, 5.5× as many; at a 1000-fold GO enrichment threshold, this difference increased to 20× as many (**Fig. 2c**). Even among the 1407 genes correctly predicted by both approaches, our modules gave more confident predictions than the same-size Pearson modules (median enrichment 534 [interquartile range 247 to 979] versus 314 [interquartile range 169 to 587], Mann-Whitney $p = 2 \times 10^{-42}$; **Extended Data Fig. 8**).

As a second benchmarking strategy, we investigated how our modules compare to predictions based on the single top Pearson correlation partner. It is difficult to make an apples-to-apples

comparison between modules and single pairs, but perhaps the most analogous comparison is to ask whether any GO terms are shared between a gene and its top Pearson partner. Unfortunately this runs into the problem that most pairs of genes will share at least one generic GO term (e.g. biological_process:cellular_process). To get around this issue, we benchmarked our modules versus top Pearson on only "leaf" GO terms at the bottom of the GO term hierarchy, with no more-specific "child" GO terms underneath them. Here, of the 6745 genes with functional predictions from our modules according to the definition above, 1587 (23.5%) have one of the predicted GO terms themselves, compared to only 981 (14.5%) from the top Pearson partner alone (and 990 (14.7%) from same-sized Pearson modules). Thus, our modules give correct predictions for 60% more genes than the top Pearson partner (Fisher $p = 3 \times 10^{-27}$).

Qualitatively, these improvements greatly improve co-essentiality-based functional annotation. For example, *LIF* (leukemia inhibitory factor), an interleukin 6 class cytokine and activator of the JAK/STAT and MAPK pathways, resides in a 12-gene module with its receptor *LIFR*, the interleukin 6 signal transducer *IL6ST*, the JAK/STAT core pathway members *JAK1* and *STAT3*, the JAK/STAT inhibitor *SOCS3*, the MAPK regulators *PTPN1* and *PTPN2*, the lymphopoiesis regulators *BATF3* and *IKZF1*, and the LIF-upregulated strawberry notch homolog *SBNO2*, whereas the top Pearson partners on the first-quarter 2020 edition of the DepMap portal (depmap.org/portal/gene/LIF) are *SLC39A1*, *MICALL1*, *C22orf23*, *TCN2* and *CCDC157*. Similarly, the BCL2 family member *BAK1* resides in a 14-gene module with *BAX*, *BCL2*, *BCL2L2*, *MARCH5*, *MCL1* and *VDAC2*, whereas its top Pearson partners on the DepMap portal are *ZNF479*, *WBP1L*, *HMGN2*, *C7orf50* and *HSFX1*. This pattern also holds for multiple members of well-characterized multisubunit complexes, including *MCM7* (in a 12-gene module with *MCM2*, *MCM3*, *MCM4*, *MCM5*, *MCM6* and *MCMBP*) and *LSM2* (in a 12-gene module with *LSM3*, *LSM4*, *LSM5*, *LSM6*, *LSM7* and *LSM8*), which are missed by Pearson. As yet another

example, *CEBPA* forms a 4-member module with three other transcription factors that are key regulators of hematopoiesis, *MYB*, *SPI1* and *GFI1*, whereas its top Pearson partners on the DepMap portal are *NBEAL1*, *ANAPC11*, *TSPYL5*, *TIMM8A* and *ZNF559*.

Co-essential modules capture multiple levels of biological organization

An additional key feature of the co-essential modules we identify, by virtue of their overlapping nature, is their ability to recapitulate multiple levels of biological organization as well as relationships between distinct pathways and complexes, as exemplified by the set of modules corresponding to distinct complexes of the mitochondrial respiratory chain (**Fig. 3k**). The set of modules enriched for the 10-subunit endoplasmic reticulum (ER) membrane complex (EMC) provides an additional striking example. The EMC complex was recently shown to function as a transmembrane insertase required for the biogenesis of a subset of transmembrane proteins, particularly tail-anchored and polytopic proteins[4,5] and is additionally noted for its role in cholesterol homeostasis[6]. We identify several modules enriched for EMC subunits, including a 7-gene module (#5037) containing 6 EMC subunits and *TMEM147,* a gene recently shown to function in multipass transmembrane protein biogenesis[7], and a 16-gene module (#740) containing 8 EMC subunits and 2 genes (*MBTPS1*[8] and *SCAP*[9]) required for cholesterol homeostasis. In addition, we identify a 23-gene module (#534) that contains 5 EMC subunits and 11 subunits of the lysosomal V-ATPase complex (8 of which are separately contained in 9-gene module #2450). Several of these V-ATPase subunits were recently identified in an unbiased proteomic study as among the proteins most dependent on EMC function for their stability[10]. Thus, the set of modules we identify that are enriched for EMC subunits correspond to known inter-pathway interactions, and demonstrate the power of co-essential modules to not only identify individual pathways but to point to possible inter-pathway relationships.

Comparison between co-essentiality mapping and recent genetic interaction mapping studies

We also compared our co-essential modules to recent discoveries from genetic interaction mapping studies that employed pairwise gene perturbation. In the most comprehensive yeast genetic interaction map to date[11], the uncharacterized gene YJR141W/IPA1 was linked to the cleavage/polyadenlylation factor (CPF) complex and other mRNA procsesing genes. The human homolog of IPA1 is the E3 ligase UBE3D, which is contained in module #881 in our dataset, whose top GO term is “mRNA cleavage factor complex” and which contains CPF genes and other mRNA processing factors (**Supplementary Data 2**). UBE3D is not itself well functionally characterized in humans, though it was reported in BioPlex to physically interact with a CPF complex subunit, CPSF3[12]. Thus, our approach predicts a function for UBE3D in the mRNA processing pathway that is consistent with the demonstrated role of its yeast homolog IPA1. Similarly, in the most comprehensive mammalian genetic interaction map to date[13], the authors identified *TMEM261* as a member of Complex I of the electron transport chain. *TMEM261* resides in co-essential module #2104 with most-enriched GO term “biological process:mitochondrial respiratory chain complex I assembly”. These examples showcase our work’s concordance with yeast and human genetic interaction mapping studies.

Possible role of TMEM189 in regulating sphingolipids

We noted that *TMEM189* is also present in a co-essential module, module #808, which is highly enriched for genes involved in the biosynthesis of sphingolipids, a distinct class of lipids that predominantly localizes to the plasma membrane and, similarly to ether lipids, contributes to both signaling and membrane structure and fluidity. In lipidomic analyses, cell extracts derived from cells expressing sgRNAs targeting *SPTLC2*, a subunit of serine palmitoyltransferase, the rate limiting enzyme in sphingolipid biosynthesis, were highly depleted of several sphingolipid species, whereas abundances of most sphingolipid species were largely unaltered in cell

extracts from HeLa cells expressing *TMEM189*-targeting sgRNAs, ruling out a central role for *TMEM189* in sphingolipid biosynthesis (**Supplementary Data 4**). However, we observed that the relative abundances of several very long chain sphingomyelin species were altered in cells lacking *TMEM189*, with sphingomyelins with C26 fatty-acids decreased in abundance and C22 and C24 sphingomyelins increased in abundance (**Extended Data Fig. 9a, Supplementary Data 4**). We additionally found that affinity-purified TMEM189-GFP complexes, isolated from HeLa cells, were highly enriched for SPTLC2 **(Extended Data Fig. 9b, Supplementary Data 8**). Further work is required to determine whether this pair of observations – that TMEM189 and SPTLC2 appear to physically interact, and that the abundances of very long chain sphingomyelin species are subtly altered in *TMEM189*-knockout cells – reflects a direct role for *TMEM189* in the regulation of fatty acid incorporation into ceramides. Alternatively, because sphingolipid composition is tightly regulated to maintain membrane fluidity[14], the altered sphingolipid profile observed in *TMEM189*-knockout cells may reflect a compensatory response to loss of plasmalogens and resulting disrupted membrane composition in *TMEM189*-knockout cells. Regardless of these possibilities, our results provide conclusive evidence for a primary role for *TMEM189* as the orphan desaturase required for the final step of plasmalogen biosynthesis and provide a striking example of the power of co-essential modules to predict gene function.

Using co-essential modules to suggest novel roles for previously characterized genes

Beyond nominating functions for uncharacterized genes, the modules identified in this study have the potential to suggest novel roles for partially characterized genes. Within the modules validated in this study, we note that *SEC14L1,* an inhibitor of the antiviral RIG-1 pathway[15], is now the only gene in 8-gene module #2213 not yet linked to ether lipid synthesis. Notably, *SEC14L1* contains a conserved lipid-binding domain, and its yeast homolog *SEC14* is involved in non-vesicular lipid transport[16]. Plasmalogens are known to traffic to the cell surface after

being synthesized in the endoplasmic reticulum, but the factors that regulate plasmalogen trafficking, and the route plasmalogens take to the plasma membrane, remain undefined; the possibility that *SEC14L1* regulates either plasmalogen synthesis or transport is a prime example of an experimentally testable hypothesis motivated by our findings. Indeed, preliminary support for this hypothesis is provided by *in vitro* studies of yeast Sec14, which revealed that purified Sec14 is sufficient to catalyze ether lipid transport between lipid membranes[17]. Further study is required to confirm that human *SEC14L1* can similarly drive ether lipid transport between membranes and to discern whether non-vesciular lipid transport by *SEC14L1* could mediate ether lipid trafficking to the cell surface in living cells.

Numerous additional modules (**Fig. 3b, d, e, f, j, k, Supplementary Data 2**) that are highly enriched for one GO term contain components of seemingly unrelated pathways. For example, two essential components of the mitotic spindle checkpoint, *BUB1B* and *MAD2L1*, are present in a module (#1360) of kinetochore/spindle checkpoint genes but also, unexpectedly, in a module (#739) highly enriched for interferon (IFN) response genes (**Fig. 3f**). *BUB1B* and *MAD2L1* have well-defined roles in the prevention of chromosome instability (CIN)[18] but have not previously been linked to IFN gene function. One possible connection is that tumor cells with high levels of CIN generate cytosolic DNA that activates an IFN response, driving cancer progression[19]. The roles of *BUB1B* and *MAD2L1* in this process, which prevent a particular form of CIN - aneuploidy resulting from premature sister chromatid separation - were not addressed in that study and may warrant further investigation.

Comparison of GLS and diffusion maps with previous methods for gene network reconstruction

Generalized least squares accounts for the correlations between cell lines in computing gene-gene relationships, allowing us to effectively compute the neighborhoods of genes in the

network with a panel of non-independent cell lines. Our use of GLS and diffusion maps is distinct from methods previously used for gene network reconstruction, typified by thresholded Pearson correlation[11,20] and neighborhood regression[21] approaches. Incorporating more global approaches for network inference, like graphical Lasso or Bayesian network models[22], may well uncover further structure at various scales of the network.

An interactive resource for biological discovery

We created a web tool (coessentiality.net) (**Extended Data Fig. 10**; **Supplementary Video 1**) to facilitate exploration of our co-essentiality map shown in Fig. 1C, for which we anticipate several distinct uses. First, this tool can be used as a starting point for gaining insight into the function of any gene: users can search for a gene of interest, for example, KRAS (**Extended Data Fig. 10**, **Supplementary Video 1**), which is then highlighted on the interactive 2D layout in the context of its neighborhood. To understand relationships between sets of genes in a given neighborhood, users can then directly select gene neighborhoods with the cursor (**Supplementary Video 1**) to generate two data panels: a biclustered heatmap of the genes' essentiality profiles across the 485 cell lines, and a plot of the gene set's enrichment for annotated pathways, complexes and gene ontology terms[23,24] (**Extended Data Fig. 10**, **Supplementary Video 1**). Users can compare the selected gene set's essentiality profile with the mutational status and expression level of other genes. For example, with the *KRAS*-containing neighborhood selected, users can plot the lines in which *KRAS* is mutated, revealing that *KRAS* and several neighboring genes are selectively essential in *KRAS*-mutant lines (**Extended Data Fig. 10**, **Supplementary Video 1)**. Users can also gain insight into pathways required for the growth of of individual cancer lines or for cancers derived from a certain tissue by selecting cell lines (e.g. U937, **Supplementary Video 1**) or tissue types (e.g. kidney cancers, **Extended Data Fig. 10**, **Supplementary Video 1**) from drop-down menus, causing the two-dimensional network to be colored according to each gene's essentiality in the selected cell line or tissue. Finally, users can upload gene sets to

understand relationships between multiple genes of interest (**Supplementary Video 1**). We anticipate this tool will be a powerful hypothesis-generating platform for functional characterization of genes and gene sets and identification of cancer-type specific pathway dependencies.

Supplementary Methods

Pearson correlation simulations

For the simulations in Extended Data Figs. 1a and b, the x coordinates of the 8 red data points were sampled to be uniformly distributed between -4 and 0. The y coordinates were then sampled from $0.9\ x + \text{Normal}(0, (1 - 0.9^2)^{1/2})$ to have a Pearson correlation of approximately 0.9. To be visually pleasing, points were repeatedly re-simulated until two constraints were satisfied: the most extreme x and y coordinates had to be between 0.15 and 0.4 from the edge of the interval [-4, 0], and the minimum x and y differences between each pair of points had to be at least 0.2.

A second set of blue points were added alongside the red points. The blue points and red points share the same x coordinates, but the blue points' y coordinates were sampled to be uniformly distributed between -1 and 0 to avoid having any significant correlation with the x coordinates. To enforce this lack of correlation, the y coordinates were repeatedly sampled until both the Pearson and Spearman correlation p-values were greater than 0.5.

In the right half of Extended Data Fig. 1b, the same red and blue points were plotted, in addition to 20 duplicates of each of these points, shifted by a small amount of noise sampled from Normal(0, 0.1).

Manual Annotation of Modules and Neighborhoods

Co-essential modules in Fig. 3 were arranged into pathway schematics based on figures in textbooks or in peer-reviewed review papers, the latter of which were identified using Google Image searches for "[search term] pathway", using the pathway-descriptive search terms (simple 1-2 word phrases chosen to encapsulate the 3 most-enriched GO terms in that

co-essential module) listed in **Supplementary Data 2**. The textbook or literature pathway diagram(s) used for each module are listed in **Supplementary Data 2**. All genes in each module were shown in each panel, either as part of the main schematic diagram (if determined to be known members of the pathway; see below), or in one of two insets: “Uncharacterized” if the Uniprot annotation score for that gene was <3 (as of 11/2019), or, otherwise, “Other”. Genes were determined to be known members of a pathway, if they met one of three criteria: 1) the gene was depicted in one of the reference schematics for that pathway; 2) a Google Scholar search for the gene name (or alternate name; see **Supplementary Data 2)** and the indicated search term for that module revealed a paper (among the first 10 search results) indicating experimental evidence for involvement of that gene in that pathway; 3) the gene was defined as a member of a complex in the HUGO Gene Nomenclature Committee Database. References for each manual annotation are included in **Supplementary Data 2.**

In Fig. 1c, neighborhoods were manually annotated as concise 1-3 word encapsulating terms based on the most-enriched GO and KEGG terms (using gProfiler) for each neighborhood. In Fig. 1d, genes were shown as colored points if literature evidence linking that gene to the pathway was identified (see below), otherwise the gene was shown as a gray point. To determine whether literature evidence exists that supported the involvement of that gene in that pathway, a Google Scholar search was first performed to identify an overarching review of genes in that pathway. For genes in the neighborhood not linked to the pathway in that review (or if a suitable review was not identified), individual literature searches were performed in Google Scholar using the following query: “[Gene name] [Neighborhood name]”. References for each gene-pathway manual annotation in Fig. 1d are included in **Supplementary Data 1**.

Browser heatmap

The 481-cell-line bias-corrected CERES essentiality scores are plotted alongside the global co-essentiality network in the browser (**Extended Data Fig. 10**), and update interactively when a subset of genes is selected. The heatmap's rows and columns are ordered by co-clustering them to find latent components, using the *sklearn.cluster.bicluster.SpectralCoclustering* implementation of the SVD-based algorithm of Dhillon et al., 2001[25].

GO enrichment

Below the essentiality score heatmap in the browser is the enrichment analysis panel, which displays hypergeometric p-values for the selected gene set against various database annotation terms, as computed by gProfiler[26].

Module heatmaps

To create heatmaps for each module (**Supplementary Data 2**), the bias-corrected CERES scores for genes in the module were hierarchically biclustered with Ward linkage using the scipy.cluster.hierarchy.linkage function from the *scipy* Python package, with the *method* argument set to 'ward' and the *optimal_ordering* argument set to True. This biclustering was then visualized with the seaborn.clustermap function from the *seaborn* Python package.

Module GO term enrichments

In Supplementary Data 2, GO term enrichment *p*-values were calculated via a hypergeometric test implemented using the scipy.stats.hypergeom.sf function from the *scipy* Python package. When calculating enrichments and *p*-values, genes not found in any module were excluded. GO term enrichments and *p*-values were calculated for all GO terms from the GO consortium[23,24], except for GO terms with fewer than 20 total genes across all modules and three overly broad GO terms (biological process:biological process, cellular component:cellular

component and molecular function:molecular function), which were excluded. The top 3 most-enriched GO terms for each module were listed with their hypergeometric *p*-values, provided the *p*-value was significant at a per-module Bonferroni threshold of 0.05, corrected across all GO terms.

Generation of HeLa cell lines expressing individual sgRNAs and tagged genes

HeLa cells were maintained on tissue culture plastic and cultured in DMEM supplemented with 100 units/mL penicillin, 100 µg/mL streptomycin, and 10% fetal calf serum. Cells were passaged by incubation with Accutase, pelleting with centrifugation at 300 g for 5 min at room temperature, and replating in fresh growth medium. To transduce cells with individual constructs, HeLa cells stably expressing Cas9-BFP were lentivirally infected with constructs expressing either individual sgRNAs or GFP- or mCherry-tagged genes and a puromycin resistance (PuroR) or blasticidin resistance (BlastR) gene. At 3 d after infection, cells were selected with 2 µg/mL puromycin or 10µg/mL blasticidin for 3 d, and cultured for at least 3 d without selection agent before use in experiments.

Immunoblotting

Cleared cell extracts prepared in lysis buffer (50 mM Tris-Hcl pH 7.5, 150 mM NaCl, 1mM EDTA, 1% Triton X-100, 1 x cOmplete protease inhibitor cocktail (Roche) were heated in SDS loading buffer and subjected to SDS-PAGE, transferred to nitrocellulose, blotted and imaged using an Odyssey CLx (LI-COR Biosciences) or Supersignal West Femto Maximum Sensitivity Substrate with a Chemidoc System (Bio-Rad). The following antibodies were used: Rabbit polyclonal anti-TMEM189 (HPA059549, Sigma, 1:250), mouse monoclonal anti-GAPDH (AM4300, Fisher, 1:5000), rabbit polyclonal anti-mCherry (ab167453, Abcam, 1:1000), mouse monoclonal anti-GFP (A-11120, Thermo Fisher, 1:1000) and rabbit polyclonal anti-beta actin (ab8227, Abcam, 1:1000). In Fig. 4f, the species shown is that which migrated at the expected

molecular weight of TMEM189 (see Source Data for uncropped immunoblot). Cell extracts from HeLa-Cas9 cells expressing sgRNAs targeting either control loci or the *TMEM189* locus (with >70% knockout efficiency verified using ICE analysis (Synthego)) were used to validate the specificity of the anti-TMEM189 antibody (**Fig. 4f**).

Immunoprecipitation and mass spectrometry

For immunoprecipitations, HeLa cells that had been transduced with tagged constructs as described above were cultured in either T-150 flasks or 15 cm plates and harvested near confluency. Cell lysates for each cell line were prepared by detaching cells with trypsin, washing in PBS, resuspending in 1 mL IP buffer (50 mM HEPES, pH 6.8, 150 mM NaCl, 2mM EDTA, 1% Triton X-100, 1x cOmplete protease inhibitor cocktail (Roche)) and incubating for 30 min on ice. Cell lysates were cleared by centrifugation at 5,000 g for 5 min before incubation with 50 µl pre-washed GFP-TRAP MA beads (Chromotek) for 1 h at 4 degrees Centigrade, with end-over-end rotation. Beads were washed 4 times for 5 min with 1mL IP buffer prior to elution with 30 µl SDS sample buffer at 70 degrees Centigrade. In Fig. 5e, a similar procedure was followed, except RFP-TRAP MA beads (Chromotek) were used.

For analysis by mass spectrometry, elutions were loaded on 4-12% Bis-Tris NuPage SDS-PAGE gels (Thermo Fisher) and run at 100V for 30 minutes. Gels were stained with SimplyBlue SafeStain (Thermo Fisher) and equivalent gel fragments for each lane were extracted, sliced into small fragments, and stored in 1% acetic acid. Samples were processed as described previously[27], with the following modifications. Briefly, gel slices were first resuspended in 100 µL 50 mM ammonium bicarbonate supplemented with 10 µl 50 mM dithiothreitol and incubated for 30 min at 55˚C, and subsequently alkylated with 10 µl 100 mM acrylamide for 30 min at room temperature. Solution phase was discarded, and gel pieces were washed 3 times with 100 µl 50 mM ammonium bicarbonate/50% acetonitrile for 10 min at room temperature. 100 µL of 50 mM

ammonium bicarbonate and 1 µg trypsin was added to digest bound proteins during an overnight incubation at 37˚C. The overnight digests were spun down and the solution was collected. Peptides were extracted more two additional times with 50 µl of 70% acetonitrile/29% water/1% formic acid and incubated for 10 min at 37˚C and centrifuged at 10,000 x *g* for 2 minutes, and all three extractions were combined. The combined extracts were dried using a Speedvac and reconstituted in 100 mM triethylammonium bicarbonate for TMT10plex labelling (Thermo Fisher) following the manufacturer's instructions, and samples were mixed to generate the final peptide mixture.

Protein digests were loaded on a Waters Liquid Chromatography column coupled to an Orbitrap Fusion mass spectrometer (Thermo Fisher). Peptides were separated using a 25 cm long and 100 µm inner diameter capillary column packed with Sepax 1.8 µm C18 resin. Peptides were eluted off in a 60 min gradient at a flow rate of 600 nl/min from 5% to 35% acetonitrile in 0.1% formic acid. Mass spectrometry data was acquired by one full MS scan at 120k resolution followed with MS2 using HCD at 30k resolution. The instrument was set to run in top speed mode with 3 s cycle.

Raw data was processed using Thermo Proteome Discoverer software version 2.2. MS data were searched against a human proteome database with 1% FDR at peptide level. Protein quantification was based on the precursor ion peak intensity using the label free quantitation workflow. For Figs. 5d and Extended Data Fig. 9b, keratins and proteins identified with only one peptide were excluded from analysis. *P*-values were generated from Student's t-tests between duplicate samples of indicated tagged genes and all 6 other samples analyzed in the same run (including duplicate samples derived from cells expressing GFP-tagged JTB (an unrelated gene), and from cells expressing GFP alone).

Time-lapse microscopy for transferrin endocytosis

HeLa cells that had been transduced with sgRNAs, co-expressed with GFP and PuroR, and selected with puromycin as described above were lifted, centrifuged, and re-plated in 24-well tissue culture plates in triplicate at a density of 50,000 cells per well. After 1 d, cells were washed once in dPBS, incubated in dPBS for 30 minutes, and incubated in dPBS containing 25μg/mL transferrin-pHrodo (Thermo Fisher). Plates were transferred to an incubator and imaged every 20 minutes using an Incucyte (Essen). Total red intensity for each well, averaged over 9 images per well, was calculated after applying a threshold of 1 RCU using top-hat background subtraction. Reported values represent the mean total red fluorescence intensity after 24 h, normalized to the total green fluorescence signal to account for small variations in plating density, of triplicate wells. Similar results were obtained in three independent experiments using two sets of independently-generated cell lines.

Confocal microscopy

HeLa cells were transduced with a lentiviral construct coexpressing *C15orf57*-GFP and PuroR, and then transduced with a lentiviral construct expressing A*P2S1*-mCherry and BlastR. Cells cultured in glass-bottom 24-well plates and imaged in a single plane near the glass surface using an inverted Nikon Eclipse Ti-E spinning disk confocal microscope and an Andor Ixon3 EMCCD camera using an oil-immersion 100x objective (NA=1.45) using NIS Elements software (v4.4, Nikon). Images were assembled and adjusted for brightness and contrast in Photoshop (Adobe).

Supplementary References

1. Walker, E. J. *et al.* Monoallelic expression determines oncogenic progression and outcome in benign and malignant brain tumors. *Cancer Res.* 72, 636–644 (2012).
2. Dutkowski, J. *et al.* A gene ontology inferred from molecular networks. *Nat. Biotechnol.* 31, 38–45 (2013).
3. Kramer, M., Dutkowski, J., Yu, M., Bafna, V. & Ideker, T. Inferring gene ontologies from pairwise similarity data. *Bioinformatics* 30, i34–42 (2014).
4. Shurtleff, M. J. *et al.* The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. *Elife* 7, (2018).
5. Guna, A., Volkmar, N., Christianson, J. C. & Hegde, R. S. The ER membrane protein complex is a transmembrane domain insertase. *Science* 359, 470–473 (2018).
6. Volkmar, N. *et al.* The ER membrane protein complex promotes biogenesis of sterol-related enzymes maintaining cholesterol homeostasis. *J. Cell Sci.* 132, (2019).
7. McGilvray, P. T. *et al.* An ER translocon for multi-pass membrane protein biogenesis. *Elife* 9, (2020).
8. Duncan, E. A., Brown, M. S., Goldstein, J. L. & Sakai, J. Cleavage site for sterol-regulated protease localized to a leu-Ser bond in the lumenal loop of sterol regulatory element-binding protein-2. *J. Biol. Chem.* 272, 12778–12785 (1997).
9. Hua, X., Nohturfft, A., Goldstein, J. L. & Brown, M. S. Sterol Resistance in CHO Cells Traced to Point Mutation in SREBP Cleavage–Activating Protein. *Cell* 87, 415–426 (1996).
10. Tian, S. *et al.* Proteomic Analysis Identifies Membrane Proteins Dependent on the ER Membrane Protein Complex. *Cell Reports* vol. 28 2517–2526.e5 (2019).
11. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* 353, (2016).

12. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* 545, 505–509 (2017).
13. Horlbeck, M. A. *et al.* Mapping the Genetic Landscape of Human Cells. *Cell* 174, 953–967.e22 (2018).
14. Breslow, D. K. & Weissman, J. S. Membranes in balance: mechanisms of sphingolipid homeostasis. *Mol. Cell* 40, 267–279 (2010).
15. Li, M.-T. *et al.* Negative regulation of RIG-I-mediated innate antiviral signaling by SEC14L1. *J. Virol.* 87, 10037–10046 (2013).
16. Saito, K., Tautz, L. & Mustelin, T. The lipid-binding SEC14 domain. *Biochim. Biophys. Acta* 1771, 719–726 (2007).
17. Szolderits, G., Daum, G., Paltauf, F. & Hermetter, A. Protein-catalyzed transport of ether phospholipids. *Biochim. Biophys. Acta* 1063, 197–202 (1991).
18. Ricke, R. M., van Ree, J. H. & van Deursen, J. M. Whole chromosome instability and cancer: a complex relationship. *Trends Genet.* 24, 457–466 (2008).
19. Bakhoum, S. F. *et al.* Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* 553, 467–472 (2018).
20. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* 327, 425–431 (2010).
21. Meinshausen, N. & Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* vol. 34 1436–1462 (2006).
22. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804 (2012).
23. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29 (2000).
24. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338 (2017).
25. Dhillon, I. S. Co-clustering Documents and Words Using Bipartite Spectral Graph

Partitioning. in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 269–274 (ACM, 2001).

26. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* vol. 47 W191–W198 (2019).
27. Haney, M. S. *et al.* Identification of phagocytosis regulators using magnetic genome-wide CRISPR screens. *Nat. Genet.* 50, 1716–1727 (2018).