BMC Bioinformatics

RESEARCH

Open Access

Optimal construction of a functional interaction network from pooled library CRISPR fitness screens

Veronica Gheorghe[1,2] and Traver Hart[1,3*]

*Correspondence:
traver@hart-lab.org

[1] Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[2] Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center UTHealth, Houston, TX, USA
[3] Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Abstract

Background: Functional interaction networks, where edges connect genes likely to operate in the same biological process or pathway, can be inferred from CRISPR knockout screens in cancer cell lines. Genes with similar knockout fitness profiles across a sufficiently diverse set of cell line screens are likely to be co-functional, and these "coessentiality" networks are increasingly powerful predictors of gene function and biological modularity. While several such networks have been published, most use different algorithms for each step of the network construction process.

Results: In this study, we identify an optimal measure of functional interaction and test all combinations of options at each step—essentiality scoring, sample variance and covariance normalization, and similarity measurement—to identify best practices for generating a functional interaction network from CRISPR knockout data. We show that Bayes Factor and Ceres scores give the best results, that Ceres outperforms the newer Chronos scoring scheme, and that covariance normalization is a critical step in network construction. We further show that Pearson correlation, mathematically identical to ordinary least squares after covariance normalization, can be extended by using partial correlation to detect and amplify signals from "moonlighting" proteins which show context-dependent interaction with different partners.

Conclusions: We describe a systematic survey of methods for generating coessentiality networks from the Cancer Dependency Map data and provide a partial correlation-based approach for exploring context-dependent interactions.

Keywords: Essential genes, Coessentiality network, Functional network, CRISPR

Background

Functional interaction networks connect genes which operate in the same biological process or pathway. Systematic genetic interaction surveys in yeast showed that genes with similar genetic interaction profiles across dozens [1] to thousands [2, 3] of strains showed high likelihood of functional interaction. In human cells, genome-scale pooled library knockdown [4] or CRISPR-mediated gene knockout screens [5–9] enabled the comparison of gene loss of function fitness vectors across cell lines, which show the

same tendency toward co-functionality. As the number of cell line CRISPR screens has grown [10], these coessentiality networks have become a powerful method for inferring gene function and understanding the modular architecture of the cell [11].

Though informative, there has to date been no systematic evaluation of the quality of each of these networks—nor even consensus on how to measure quality. There are numerous algorithms for transforming raw gRNA read count data from pooled library CRISPR screens into quantitative measurements of gene fitness effect, including Bagel2 [12], Castle [13], Ceres [14], Chronos [15], JACKS [16], MAGECK [17], and a Z-score approach [18] optimized for finding genes with positive instead of negative knockout fitness effects.

After combining gRNA-level fold changes into gene-level scores, each published approach uses different sample-level variance normalization. Kim et al. [8] use quantile normalization with no reference distribution, which resets the value of each gene to the mean of all gene fitness scores at that rank. Boyle et al. [7] subtract the top principle components of the fitness matrix of a set of reference nonessential genes (olfactory receptors), to remove artifactual components. Wainberg et al. [9] go further and implement a covariance whitening method based on Cholesky decomposition, which normalizes both variance and covariance and minimizes the effect of uneven sample distribution—a potentially serious source of bias, since, for example, there are more than ten times as many lung cancer cell lines in DepMap as prostate cancer lines. The similarity of gene vectors is frequently measured by Pearson correlation coefficient (PCC), though Wainberg et al. argue that inflated P-values from PCC lead to numerous false positives and that ordinary least squares (OLS) after Cholesky whitening—collectively Generalized Least Squares (GLS) —is a better approach.

In this study, we systematically compare combinations of essentiality scoring algorithms, variance and covariance normalization methods, and similarity measures to determine an optimal strategy for building a functional interaction network from coessentiality data. We show that results are highly dependent on data processing steps, and that covariance whitening is a critical step in improving the predictive power of networks. We further show that, after covariance normalization, PCC and OLS are mathematically identical, and demonstrate how PCC and partial correlation can be employed to extract context-dependent interactions from global coessentiality networks.

Results

To assess an optimal functional interaction network from coessentiality data, we first acquired gene knockout fitness data from CRISPR knockout screens from the Cancer Dependency Map project [10]. A typical network construction pipeline involves converting raw read count data into a matrix of gene knockout fitness scores, performing variance and/or covariance normalization across samples, and measuring similarity of all pairs of gene fitness profiles across samples (Fig. 1). We constructed networks using all combinations of alternatives at each step and evaluated the quality of each network using a log likelihood framework that measures the enrichment for gene pairs that belong to the same annotated biological process or pathway.

Gene knockout fitness effects ("essentiality scores") were calculated using four recent algorithms designed explicitly for the analysis of CRISPR pooled library knockout

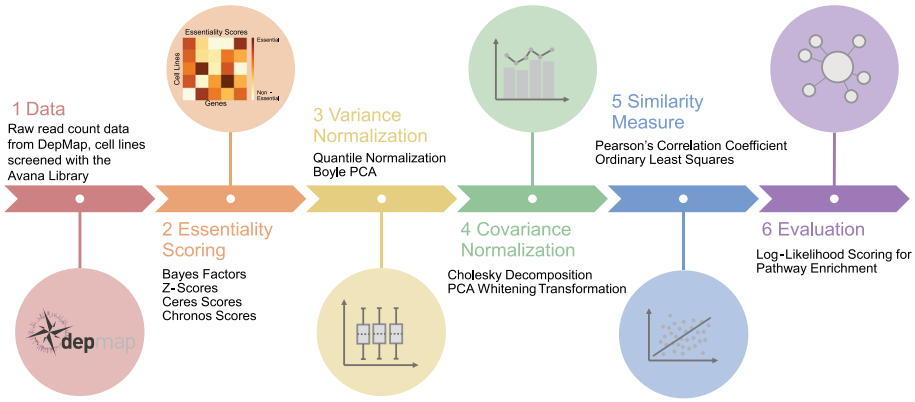

Fig. 1 The network construction pipeline. We generated coessentiality networks using all combinations of essentiality scoring, variance normalization, covariance normalization, and similarity measurement methods, and each network is evaluated for pathway enrichment. All networks are built from the same raw read count data

screens. Bayes Factors (BF) were calculated using the BAGEL2 algorithm [12]; Ceres [14] and Chronos [15] scores were downloaded from the Dep Map portal [10], and Z-scores were calculated from a two-component Gaussian mixture model of average fold changes as described in Lenoir et al. [18]. We selected the intersection of all genes and samples measured by all algorithms, resulting in a matrix of 17,834 genes by 730 cell lines, ensuring a comparable result from each algorithm.

We previously used quantile normalization as a method of sample-level variance normalization, where each gene's fitness score is normalized to the rank mean (see Methods). An alternative approach described in Boyle et al. [7] that uses singular value decomposition of a reference set of nonessential genes to identify and remove "artifactual" components is functionally equivalent to sample-level variance normalization (Additional file 1: Fig. S1) and is therefore included here.

Covariance normalization or "whitening" can be an important step to prevent correlated samples from biasing results; e.g. when specific tissue types or oncogenic mutations are overrepresented in the cell lines. PCA-based whitening is one such approach. Wainberg et al. [9] introduce Cholesky decomposition as a covariance normalization in their coessentiality network which, coupled with ordinary least squares (OLS) regression, they apply as a generalized least squares (GLS) approach. We apply Cholesky whitening (covariance normalization) and OLS (similarity measure) separately to evaluate the relative contributions of each and include the commonly used Pearson correlation coefficient (PCC) as an alternative similarity measure. See Fig. 1 for an overview of the processing steps applied, and Methods for implementation details.

To evaluate the ability of a coessentiality network to identify co-functional gene pairs, we compared the most similar (top 50 k) gene pairs from each network to annotated pathway databases. We measured enrichment by applying the log likelihood framework developed for functional interaction networks [19], where gene pairs belonging to the same annotated pathway are considered true positives and pairs belonging to different pathways are false positives (Fig. 2A). The log likelihood scheme evaluates each network's ability to recreate known gene pathways, by measuring the frequency of true

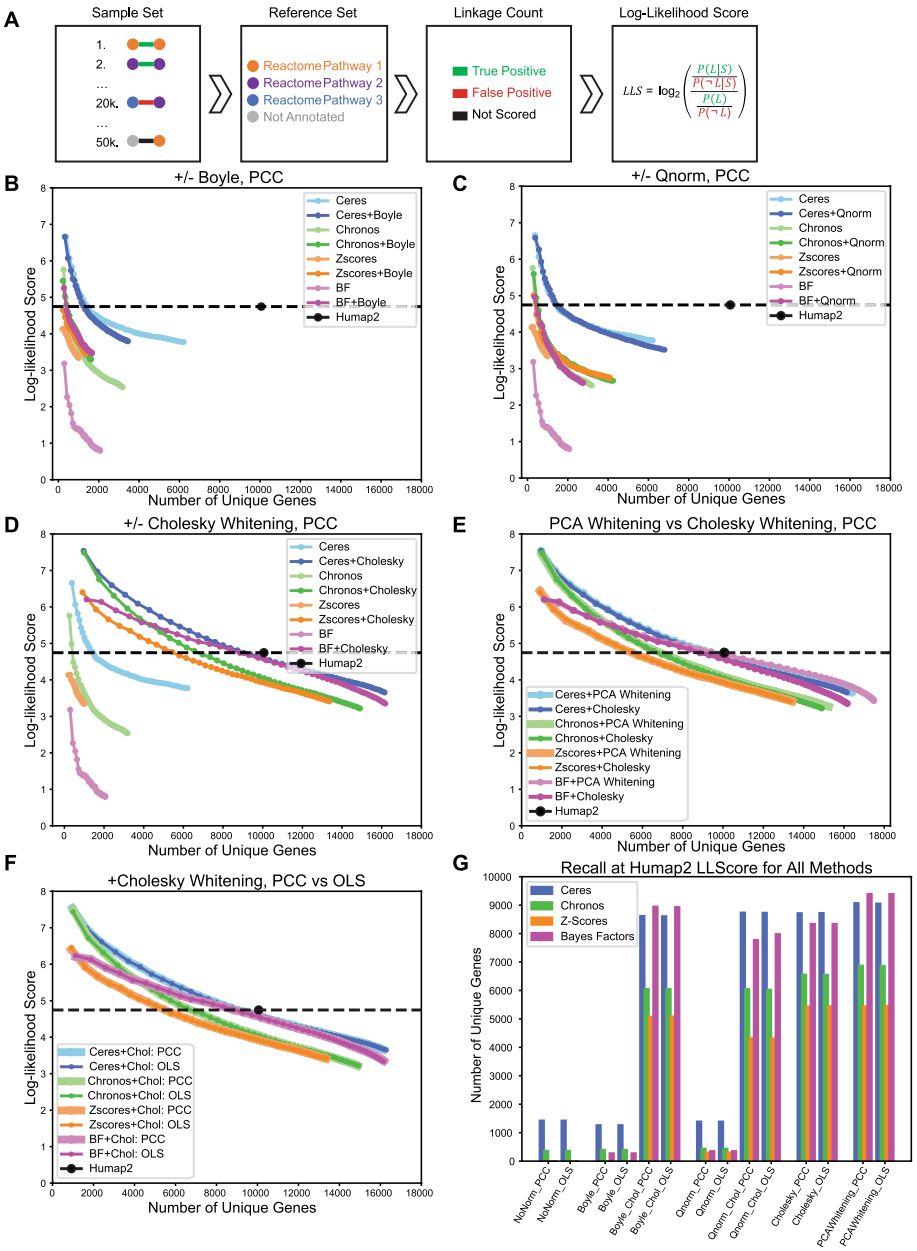

Fig. 2 Evaluation of network construction methods. **A** The Log-likelihood framework for evaluating a network for pathway enrichment. Gene pairs in the sample are compared against reference set CleanReactome. Pairs annotated in the same Reactome pathway are counted are true positives, while pairs annotated in different pathways are false positives. **B** Top ranking 50 k gene pairs from networks created with and without applying Boyle PCA normalization are binned in bins of 1000 pairs, and evaluated with LLS scheme. For each network, LL scores are plotted versus the number of unique genes in each bin, calculated cumulatively. **C** LLS and recall for networks created with and without Quantile-normalization applied. **D** LLS and recall for networks created with and without Cholesky Whitening. **E** LLS and recall for networks where PCA-whitening was used as covariance normalizations, compared to networks where Cholesky-whitening was used. **F** LLS and recall, for networks with Cholesky-whitening applied, using PCC as similarity measure versus using OLS. **G** Recall for all 56 networks at the bin where the networks LLScore is approximately equal to the Humap2 LLS

positives versus false positives observed in the given network and comparing it to the background expectation. The background expectation is given by the frequency of linkages between all annotated genes operating in the same pathway versus operating in different pathways. We explored several pathway annotation databases, including Reactome [20], KEGG [21], and the GO Biological Process tree [22], and determined that Reactome offered the most complete coverage (Additional file 1: Fig. S2). Interestingly, we found that a very large number of genes and gene pairs in the annotated set belonged to just six annotated pathways involving mitochondrial translation and oxidative phosphorylation. Since the number of edges in a fully connected network grows with the square of the number of nodes, the large size of the mitochondrial ribosome and ETC Complex I result in a very large number of "true positive" hits that can overestimate the quality of the remaining network (Additional file 1: Fig. S2). Since oxphos genes are known to be a source of bias in coessentiality networks and differential essentiality [23], we removed these six pathways from the Reactome reference set, hereafter referred to as CleanReactome.

We evaluated the cumulative (Fig. 2B–F) and local (Additional file 1: Fig. S3) functional enrichment LLS for the top 50,000 gene pairs in each network, in bins of 1000 gene pairs, and plotted coverage (total number of genes in the network) versus functional enrichment (LLS score). Both Boyle (Fig. 2B) and quantile normalization (Fig. 2C) significantly improved both the coverage and accuracy of the Bayes Factor and Z-score networks, but had lesser impact on the Ceres and Chronos data, likely because both of these methods contain a functional equivalent to sample variance normalization.

In contrast to sample variance normalization, global covariance normalization dramatically improved both the coverage and accuracy of networks derived from all four essentiality scores. Cholesky whitening (Fig. 2D) increased recall and functional enrichment for both Ceres and BF networks to nearly match that of the hu.MAP 2.0 compendium of human protein complexes, derived from integration of numerous large-scale affinity purification/mass spectrometry and other protein–protein interaction data [24]. PCA whitening (Fig. 2E) closely matched the improvement of the Cholesky approach, with trivial incremental improvement in some cases.

A key conclusion of the Wainberg et al. study is that GLS, implemented there as Cholesky whitening plus OLS, is superior to PCC-derived functional interaction networks. However, after covariance normalization, OLS is mathematically identical to Pearson correlation (see "Methods" section). This is reflected in the identical LLS curves generated by PCC or OLS derived similarity scores after Cholesky whitening (Fig. 2F).

We summarized each network's performance by calculating its recall at the functional enrichment level offered by hu.MAP2 (Fig. 2G). Raw or sample-normalized networks offered weak performance compared to pipelines that include covariance normalization. Unsurprisingly, sample variance normalization steps (Boyle, quantile normalization) provide no incremental improvement when covariance normalization is applied, OLS and PCC are identical after covariance normalization, and different covariance normalization methods give very similar results. Perhaps more surprising is the significantly better coverage of networks generated with Ceres or BF scores relative to Z-scores or the newer Chronos scores, the current method of choice for reporting DepMap hits.

Though these networks may yield equivalent scores of global accuracy, it is not necessarily the case that they contain identical information. To assess systematic differences between high-scoring networks, we examined differences between the covariance

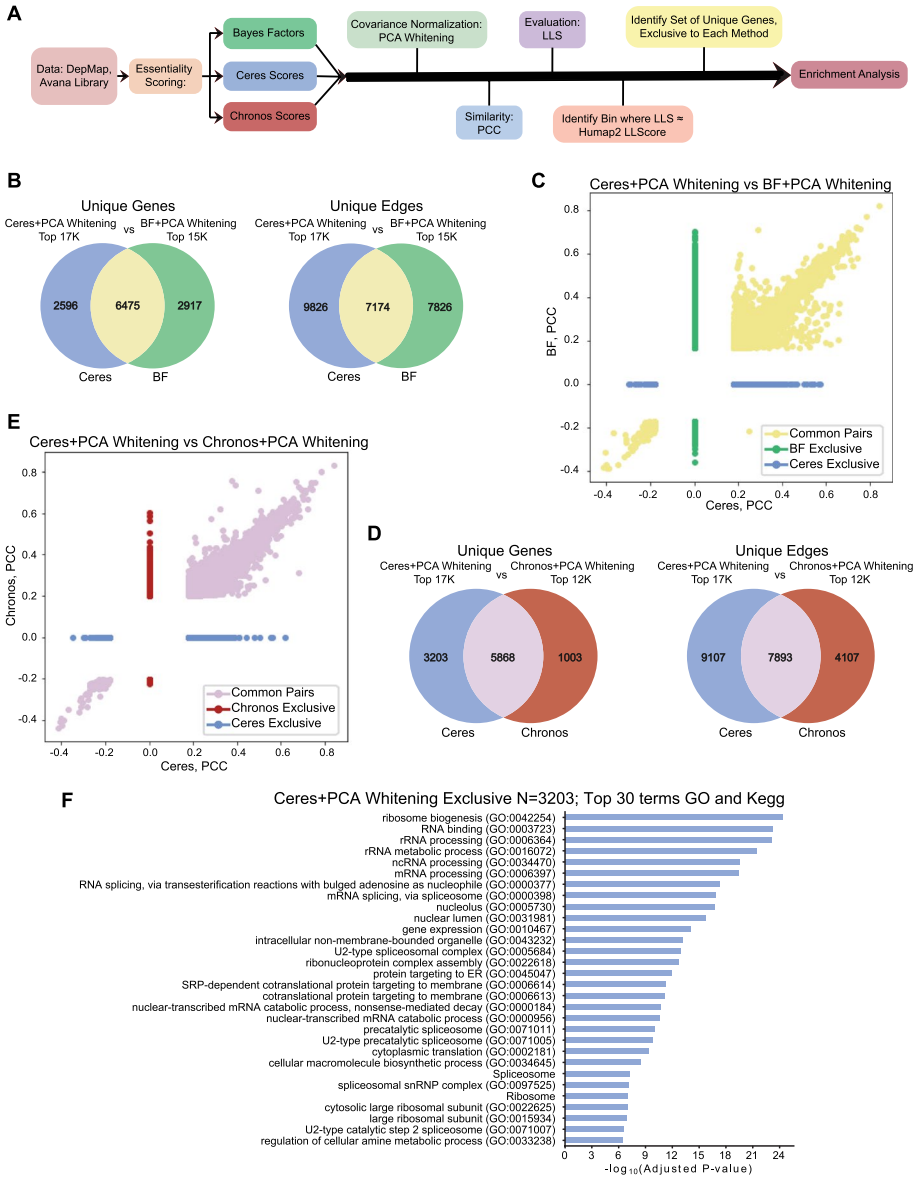

Fig. 3 Comparison of top networks. **A** Schema for enrichment comparison of top scoring networks. Ceres, BF, and Chronos scores were subjected to identical processing pipelines and the resulting networks compared. **B** Venn diagrams depicting the numbers of genes and edges exclusive to the top 17 k edges in the network created with Ceres + PCAwhitening + PCC versus the top 15 k edges in the network created with BF + PCAwhitening + PCC. **C** Pearson's correlation coefficients of common edges, and of exclusive edges to each of the networks (Ceres vs BF). **D** Venn diagrams depicting the numbers of genes and edges exclusive to the top 17 k edges in the network created with Ceres + PCAwhitening + PCC versus the top 12 k edges in the network created with Chronos + PCAwhitening + PCC. **E** Pearson's correlation coefficients of common edges, and of exclusive edges to each of the networks (Ceres vs Chronos). **F** Enrichment of the gene set exclusive to Ceres + PCAwhitening + PCC (when compared to Chronos). Only the top 30 enriched GO and Kegg terms are listed in the graph

normalized Ceres network (top 17,000 edges) and the normalized BF (top 15,000 edges) and Chronos (top 12,000 edges) networks (Fig. 3A). The functional enrichment level of hu.MAP2 was used as a threshold for the three top scoring networks. The Ceres and BF networks share only ~ 7000 edges and ~ 6500 genes, while each network contains more than 2500 unique genes (Fig. 3B). Common edges are generally correlated but both networks contain high-scoring edges that are unique to that processing pipeline (Fig. 3C). However, the genes unique to each pipeline show no functional enrichment for GO or KEGG pathways, suggesting no functional bias. Likewise, the Ceres and Chronos networks share nearly 8000 edges and 6000 genes (Fig. 3D) and common edges are highly correlated (Fig. 3E). The Ceres network, larger because of its greater recall at the hu.MAP2 LLS threshold, has 3202 unique genes that are highly enriched for core cellular processes such as transcription and translation, suggesting that these genes are depleted in the Chronos network (Fig. 3F). This is not solely a result of the smaller Chronos network; a similar analysis of identical-sized networks yields the same results (Additional file 1: Fig. S4), suggesting a systematic limitation of the Chronos scoring scheme relative to Ceres.

Wainberg et al. argue that the GLS approach, combining covariance normalization with OLS, is superior to PCC because the P-values of PCC networks can be inflated by unequally weighted samples [9], leading to false positives. However, PCC and OLS are mathematically equivalent after covariance normalization, and this is reflected in the identical performance curves of the PCC and OLS similarity measures as shown in Fig. 2F. One advantage of Pearson correlation is that it is far less computationally intensive; an all-by-all gene correlation matrix takes only a few seconds (and one line of code) to calculate while OLS requires setting up the regression for each gene pair individually, which can take orders of magnitude longer. Notably, global PCC values after whitening are significantly smaller in magnitude than PCC before whitening, leading to some counterintuitive results; at the same rank, post-whitening PCC is often considerably weaker than pre-whitening PCC (Additional file 1: Fig. S5) despite showing a significantly higher enrichment for functional interaction (Fig. 2D).

A second advantage of the Pearson correlation approach is that it facilitates the use of partial correlation to detect conditional interactions between sets of genes. Partial correlation can measure the similarity of two gene knockout profiles after removing the effect of a third gene, set of genes, or another vector. If the functional relationship between two genes varies across the assayed cell lines, the observed correlation can be diluted by the presence of samples in which the relationship is severed [25]. For example, consider common oncogenes *KRAS* and *NRAS*, two members of the MAP kinase signaling pathway whose mutually exclusive mutations mean they are essential in different cell lines (Fig. 4A). Downstream signaling partner RAF1 is also essential in the presence of either *KRAS* or *NRAS* (Fig. 4A). The PCC of *RAF1-NRAS* $\rho_{RAF1,NRAS}$ is significant, ranking in the top 0.003% of global correlation pairs, but the partial correlation of RAF1-NRAS with respect to KRAS $\rho_{RAF1,NRAS\_KRAS}$ is even higher. Likewise, $\rho_{RAF1,KRAS\_NRAS} > \rho_{RAF1,KRAS}$, indicating higher correlation after removal of the NRAS signal, while the mutual exclusivity of KRAS and NRAS mutations results in $\rho_{KRAS,NRAS} < 0$ (Fig. 4B). This can be considered a case of a protein "moonlighting" where *RAF1* interacts with either *KRAS* or *NRAS* depending on the context.

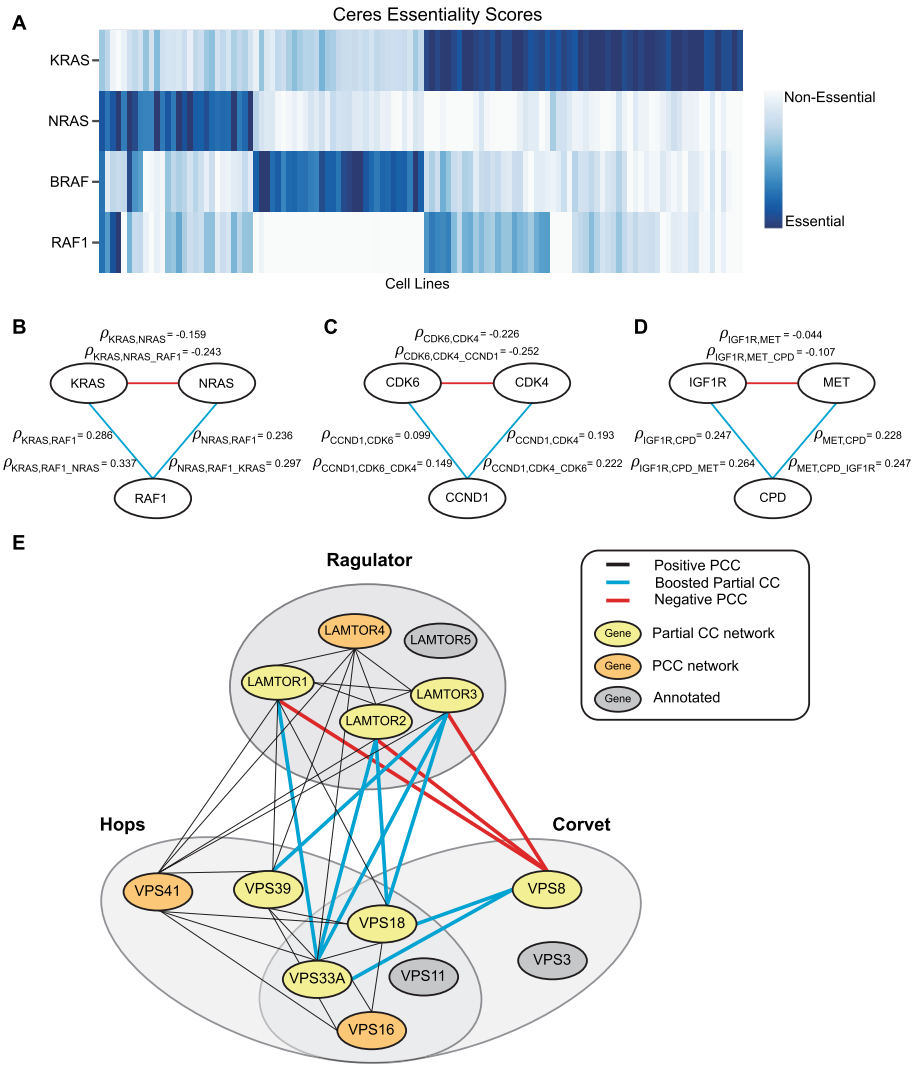

Fig. 4 Detecting moonlighting interactions using partial correlation. **A** Heatmap showing KRAS, NRAS, BRAF essentiality in selected cell lines. KRAS, NRAS, and BRAF are mutually exclusive, while RAF1 is essential in KRAS and NRAS backgrounds. **B** KRAS-NRAS-RAF1 moonlighting trio, with PCC and Partial correlation coefficients listed by the respective edges. Red edges indicate negative correlation coefficient, while blue edges indicate positive correlation coefficient. **C** CDK6-CDK4-CCND1 moonlighting trio, with correlation coefficients of edges listed. **D** IGF1R-MET-CPD moonlighting trio, with correlation coefficients of edges listed. **E** Network illustrating Ragulator complex, HOPS and CORVET complexes with shared subunits. Red colored edges indicate negative PCC, blue colored edges indicate boosted partial correlation coefficient, black colored edges are positive PCC. The yellow nodes depict genes recovered using partial correlation, orange nodes are genes recovered with PCC, and gray nodes depict genes annotated to be in the complex, that were not present in our top-ranking pairs

We conducted a systematic search for these moonlighting trios, where the candidate moonlighting gene has a positive PCC with each candidate "parental" gene, the partial correlation with each parent gene is boosted after controlling for the other parent, and the PCC of the two parents is negative. We discovered several cases that reflect known examples of context-dependent interaction. D cyclin *CCND1* interacts with cyclin-dependent kinases *CDK4* and *CDK6*, which are mutually exclusive in DepMap data (Fig. 4C). Similarly, protease *CPD* interacts independently with insulin-like growth

factor receptor *IGF1R* and hepatocyte growth factor receptor *MET* (Fig. 4D). CPD protein is known to process pro-IGF1R into its mature form [26] and we recently showed that *CPD* is also involved in maturation of *MET* receptor protein specifically in glioma cells [25].

More complicated conditional interaction subnetworks can also be informative, including at least one instance where partial correlation can disambiguate protein complexes with overlapping membership (Fig. 4E). The CORVET and HOPS complexes are functionally related molecular machines that play important roles in endocytosis [27], with the CORVET complex being primarily involved in early endocytosis and being replaced by HOPS at the late endosome/lysosome stage [28]. The Ragulator complex is a multisubunit complex that sits at the lysosome and acts as an activator of mTORC1 complex in the presence of amino acids, and is itself regulated by the HOPS complex [29]. Partial correlation analysis differentiates shared CORVET/HOPS subunits VPS33A and VPS18 from CORVET-specific subunit VPS8, while their partial correlation with Ragulator subunits LAMTOR1-4 are boosted after removing VPS8 effects (Fig. 4E). HOPS-specific subunit VPS39 is correlated with the Ragulator independent of CORVET-specific VPS8.

Discussion

Coessentiality networks offer a powerful method for inferring gene function from panels of CRISPR knockout screens in cell lines, but "coessentiality" is a catchall term describing an informatic pipeline with a variety of choices. We systematically explored the most common of these options to determine which combination of fitness scores, variance normalizations, and similarity measures gave maximal coverage of known co-functional relationships. We found that covariance normalization or "whitening" gives the largest boost to performance, regardless of fitness scoring, but also that coessentiality networks derived from covariance-normalized Bayes Factor and Ceres fitness scores markedly outperformed both the Chronos and Z-score approaches. Interestingly, after covariance normalization, the top two similarity measures, Pearson correlation and Ordinary Least Squares (OLS), are mathematically identical and give the same ranking for gene pairs.

Though genes which operate in the same biological process or pathway have similar knockout fitness profiles, genes whose functional interaction is context-dependent can have their profile similarities weakened by inclusion of cell lines or contexts where the interaction is not present. We [25] and others [30] have recently shown how analysis of CRISPR genetic screen data can reveal proteins that have multiple roles in a cell, leading to functional interactions with mutually exclusive partners. We extend the Pearson correlation network by using partial correlation to identify these "moonlighting" genes, whose interactions with one gene are boosted when the effect of a second gene is removed (and vice versa).

Conclusions

This work reinforces studies that show that coessentiality networks are among the most powerful predictors of mammalian gene function. Properly constructed, these networks contain more than 10,000 unique genes—more than half of the protein-coding genome—despite individual screens rarely recording more than 2000 genes with fitness

defects and entire categories of genes, e.g. those involved in secretory pathways or cell–cell communication, being systematically absent from pooled library screens. Moreover, the accuracy of these networks is comparable to that of the hu.MAP integrated map of protein complexes. Nevertheless, the systematic discovery and elucidation of context-dependent and pleiotropic gene functions across the DepMap cell lines has only just begun and promises to increase our insight into the organization and function of mammalian cells.

Methods

Data and essentiality scoring

The data used in this study comes from publicly available CRISPR knockout screens datasets, downloaded from the Cancer Dependency Map database (Avana dataset) [10]. Four different pipelines for measuring gene knockout fitness effects (gene essentiality) were used: BAGEL2 [12], Z-score model [18], Ceres [14], and Chronos [15]. The BAGEL2 algorithm generates log Bayes Factors (BF) to report gene essentiality for each cell line, with positive scores indicating essentiality. Z-score values are generated using a Gaussian mixture model, with negative scores indicating essentiality. The Ceres algorithm removes principle components highly related to copy-number-specific effects and scales the data so that median essential score is -1, and median non-essential score is 0. Chronos models the read-count data assuming a negative binomial distribution and removes copy-number related bias, with the scores scaled similar to Ceres.

Bayes Factors (BF) and Z-scores were calculated using raw read count data from the DepMap 20Q4 release. Ceres scores were downloaded from the 20Q4v2 release, and Chronos gene effect scores were downloaded from the 22Q2 release. We considered only the common genes and cell lines of BF, Z-scores and Ceres scores, which resulted in genes-by-cell lines data matrices of size $17{,}834 \times 730$. The Chronos data was processed to include only genes and cell lines present in the intersection with Ceres, BF and Z-score, resulting in a data matrix of size $17{,}662 \times 727$.

Normalization techniques

We compared four normalization techniques, two of which perform as variance normalization methods and the other two are covariance normalization methods. The quantile normalization technique executes variance normalization, applied to mitigate screen quality bias and to allow comparison between different samples. Quantile normalization first ranks the genes by magnitude, calculating the mean for genes in the same rank, and substituting the values of all genes in that rank with the mean value, and then reorders the genes in the original order. The Boyle principal component analysis approach aims to remove the technical confounding introduced by olfactory receptor genes that have highly correlated profiles across different genetic backgrounds [7]. This method applies principal component analysis of the gene-by-cell-line essentiality matrix across olfactory receptors, and then subtracts the first four principal components from the original score matrix, implemented here as in Wainberg et al. methods [9].

The covariance normalization methods, also known as whitening or sphering transformations, aim to remove dependencies between features in the data matrix. These methods involve linear transformation of the data matrix X using a "whitening" matrix W,

such that the resulting normalized data matrix $\tilde{X}$ has a covariance matrix equal to the identity matrix (Eq. 1). To perform this, the data matrix is first centered, subtracting the mean across all samples, and the covariance matrix is computed, denoted by Σ (Eq. 2). The covariance matrix is positive semi-definite, meaning it is symmetric with non-negative eigenvalues, thus it can be inverted, and it can be decomposed as a product of two simpler matrices. There are many "whitening" matrices that can do the linear transformation described above, and we used two different options that satisfy the condition. One of the techniques of calculating W, described in Wainberg et al. [9], uses a Cholesky decomposition of the inverse covariance matrix (Eq. 3), and the linear transformation is done as described in (Eq. 4). Another technique termed PCA whitening utilizes the Eigen-decomposition of the covariance matrix (Eq. 5), and the PCA whitening transformation is done as described in (Eq. 6).

$$\tilde{X} = WX,\ s.t.\ Cov\left(\tilde{X}\right) = I \tag{1}$$

$$Let\ Cov(X) = \Sigma \tag{2}$$

Cholesky whitening:

$$\Sigma^{-1} = LL^T \tag{3}$$

$$Let\ W = L^T \rightarrow \tilde{X} = L^T X. \tag{4}$$

PCA whitening:

$$\Sigma = EDE^{-1} \tag{5}$$

$$Let\ W = D^{-1/2}E^T \rightarrow \tilde{X} = D^{-1/2}E^T X \tag{6}$$

where $D = diag(\boldsymbol{\lambda})$, a diagonal matrix containing the eigenvalues λ_i on the diagonal; and E is the orthogonal matrix of eigenvectors.

Statistical measures of similarity

We utilized two statistical measures to quantify co-essentiality, Pearson's correlation and Ordinary Least Squares. Using Pearson's correlation, we calculate the correlation coefficient for all possible gene pairs (Eq. 7), resulting in a gene-by-gene correlation matrix. We rank the gene pairs by the correlation coefficient values.

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y},\ \text{where } \sigma \text{ denotes standard deviation} \tag{7}$$

Using ordinary least squares (implemented in Python3 using the numpy.linalg.lstsq function), we estimate the parameter vector b in the linear regression model $y = bx + \varepsilon$, where y is set to be the essentiality scores for one of the genes and x is a two-column matrix, with the first column being the other gene's essentiality scores and the second column is the intercepts, set to a vector of all ones. We ran OLS for each gene pair,

and calculated log P-values, resulting in a gene-by-gene matrix of P-values. When the Cholesky whitening transformation is applied, both x and y are transformed by the triangular Cholesky matrix, and this constitutes the Generalized least squares method described in Wainberg et al. [9]. We rank the gene pairs by the log P values.

It is important to note that in least squares linear regression the slope b is given by (Eq. 8). Rewriting it as in (Eq. 9) and substituting (Eq. 7), shows how the correlation coefficient ρ factors in. When covariance normalization is applied, whether with Cholesky or PCA whitening, the transformed data has identity covariance, meaning all features are independent and the variance along each of the features is one. With the variance (σ^2) equal to one in the covariance normalized data, PCC and OLS yield equivalent results.

$$b = \frac{cov(x,y)}{\sigma_x^2} \tag{8}$$

$$\frac{cov(x,y)}{\sigma_x^2} = \frac{cov(x,y)}{\sigma_x \sigma_y} \times \frac{\sigma_y}{\sigma_x} = \rho_{x,y} \frac{\sigma_y}{\sigma_x} \tag{9}$$

Evaluation of pathway enrichment with log-likelihood scoring

For each method, we took the top 50,000 ranking gene pairs, and bin them into bins of 1000 gene pairs for further analysis. To evaluate the performance of these methods, we conducted pathway enrichment tests with a log-likelihood scheme described by Lee et al. [19], using multiple annotated pathway databases. The log likelihood scoring quantifies the accuracy of each network and their ability to reconstruct known biological pathways and processes. The pathway databases we considered are: Kegg [21], Gene Ontology (GO) Biological Processes [22], and Reactome [20]. Additionally, we considered pre-processed versions of GO and Reactome, in which pathways that contained gene sets involved in mitochondrial translation and oxidative phosphorylation were removed. We refer to these versions as CleanGO and CleanReactome. Only six pathways were removed from Reactome in this process, and the bulk of our analysis was done with the resulting CleanReactome. In the log-likelihood scoring pipeline, we filter through the pathways within a reference file, to only include those pathways that have a minimum of 5 genes and maximum of 400 genes.

Using the gene pairs from each bin and the annotated reference set, we measure if both genes in the sample pair are annotated in the same pathway (true positive), or if they are annotated in different pathways (false positive). If a gene from our sample is not annotated in any pathways in the reference set, we do not count the pair in our calculation. We then compare the ratio of the frequencies of true positives and false positives in our sample with the background expectation, meaning the ratio of frequencies of all annotated genes operating in the same pathway and genes operating in different pathways in the reference set. The log-likelihood score is given by (10).

$$LLS = \log_2 \left(\frac{\Pr\left(true\ pos\right) / \Pr\left(false\ pos\right)}{\Pr\left(all\ reference\ pos\right) / \Pr\left(all\ reference\ neg\right)} \right) \tag{10}$$

Scores above zero indicate that the method tends to link genes in the same pathway, and high scores indicate more confident linkages. The log-likelihood scoring was performed cumulatively, as well as locally for each method.

The hu.MAP 2.0 Protein Complexes List was downloaded from the humap2 website (http://humap2.proteincomplexes.org/static/downloads/humap2/humap2_complexes_20200809.txt) [24]. We created a list of unique gene pairs from each protein complex, a total of 57,911 gene pairs, and calculated overall LLS for this list. The LL score for hu.MAP 2.0 was 4.75, with coverage of 10,060 unique genes. We used this score as a reference against which to compare the performance of the networks formed with the various methods.

Gene set enrichment analysis

To analyze the differences between the highest scoring networks, we looked at the enrichment from genes exclusive those networks. We used the hu.MAP2 LL Score as a cut-off for the PCA-whitening covariance normalized Ceres, Bayes Factors, and Chronos networks. We identified edges and genes exclusive to each network, and conducted enrichment analysis on the exclusive gene sets using the GSEAPY "enrichr" python module [31] with the reference sets 'KEGG_2021_Human', 'GO_Biological_Process_2021', 'GO_Cellular_Component_2021', 'GO_Molecular_Function_2021'.

Partial correlation

We utilized partial correlation to reveal the conditional relationship between two genes, after controlling the effect of a third gene. We calculate partial correlation of two genes x and y, while controlling the effect of a third gene z, using the recursive formula (Eq. 11). For each pair of genes x and y with $|\rho_{x,y}| > 0.15$, we calculate partial correlation with respect to every other gene in the network, and look for the gene that yields the highest partial correlation coefficient. To quantify the change in correlation coefficient after accounting for the effect of a third gene, we calculate a ratio between the two coefficients with the formula (Eq. 12).

$$\rho_{x,y\_z} = \frac{\rho_{x,y} - \rho_{x,z}\rho_{y,z}}{\sqrt{\left(1-\rho_{x,z}^2\right)\left(1-\rho_{y,z}^2\right)}} \tag{11}$$

$$\frac{\rho_{x,y\_z}^2}{\rho_{x,y}^2} \tag{12}$$

Our partial correlation analysis revealed many gene trios (x–y–z), where positive correlation coefficients between genes x and y and between genes y and z are boosted after controlling for the effect of the other, while the correlation coefficient between genes x and z is negative. A master data list of these trios is available in Additional file 2: Table S1. We created a network with these moonlighting trios, consisting of gene pairs with the edge being the positive partial correlation coefficient, as well as gene pairs with negative PCC edge. The network can be viewed in Cytoscape [32], and the data table used to create the network is available in Additional file 2: Table S2.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-05078-y.

Additional file 1. Figure S1. Boyle PCA as variance normalization. **A** Bar plot showing the percentage of variance explained by each principal component of the Boyle PCA approach across Olfactory receptor genes, applied to the Z-score data matrix. **B** Scatter plot of the standard deviation of the screen, using Z-scores data matrix, versus the screen-wise projection onto the first Principal component from the Boyle approach. **Figure S2.** LLS with other reference sets. **A** Ceres+PCC network evaluated using 3 different reference sets in the Log-likelihood analysis. Cumulative LL Scores and number of co-functional interactions in bins are plotted for the network, evaluated with Kegg, Reactome and GO reference sets. **B** Pathways containing genes associated with mitochondrial translation and oxidative phosphorylation were removed from Reactome and GO, to create CleanReactome and CleanGO. Cumulative LL Scores and number of co-functional interactions in bins are plotted for the Ceres+PCC network, evaluated with Kegg, Reactome and CleanReactome, GO and CleanGO reference sets. **C** Comparison of LLS scores and co-functional interactions of Ceres based networks using the full Reactome reference set and **D** the CleanReactome reference set in the LLS evaluation. **Figure S3.** Local LLS. **A** Local log-likelihood scores calculated per bin, using CleanReactome, for Ceres-based networks; **B** Chronos-based networks; **C** Bayes Factors based networks and **D** Z-scores based networks. **Figure S4.** Enrichment analysis for Ceres and Chronos based networks of equal size. **A** Venn diagrams depicting the numbers of genes and edges exclusive to the top 17k edges in the network created with Ceres+PCAwhitening+PCC versus the top 17k edges the network created with Chronos+PCAwhitening+PCC. **B** Enrichment of the gene set exclusive to Ceres+PCAwhitening+PCC. Only the top 30 enriched GO and Kegg terms are listed in the graph. **C** Enrichment of the gene set exclusive to Chronos+PCAwhitening+PCC. **Figure S5.** Pearson's correlation coefficients of the edges of the same rank in the Ceres+PCA Whitening+PCC and Ceres+PCC networks.

Additional file 2. Table S1. Table of the moonlighting gene trios with PCC and partial correlation details.

Additional file 3. Table S2. Table of the network constructed using the partial correlation approach.

Acknowledgements
Not applicable.

Author contributions
VG developed the approach and performed all bioinformatic analysis. VG and TH drafted and edited the manuscript. All authors read and approved the final manuscript.

Funding
VG and TH were supported by NIGMS Grant R35GM130119. TH is a CPRIT Scholar in Cancer Research and an Andrew Sabin Foundation Family Fellow. This work was supported by the NCI Cancer Center Support Grant P30CA16672.

Availability of data and materials
The data used in this study is available at Figshare. (https://doi.org/10.6084/m9.figshare.21379761). Python code for the data analysis pipeline is available at Github. (https://github.com/hart-lab/Optimal-Construction-of-Functional-Interaction-Network).

Declarations

Ethics approval and consent to participate
Ethics approval was not required for this study.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 22 September 2022 Accepted: 23 November 2022
Published online: 28 November 2022

References

1. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science. 2001;294:2364–8.
2. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The genetic landscape of a cell. Science. 2010;327:425–31.
3. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. A global genetic interaction network maps a wiring diagram of cellular function. Science. 2016;353:aaf1420.
4. Hart T, Koh C, Moffat J. Coessentiality and cofunctionality: a network approach to learning genetic vulnerabilities from cancer cell line fitness screens. 2017;134346.
5. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350:1096–101.

6. Rauscher B. Toward an integrated map of genetic interactions in cancer cells. Mol Syst Biol. 2018;14:e7656.
7. Boyle EA, Pritchard JK, Greenleaf WJ. High-resolution mapping of cancer cell networks using co-functional interactions. Mol Syst Biol. 2018;14:e8594.
8. Kim E, Dede M, Lenoir WF, Wang G, Srinivasan S, Colic M, et al. A network of human functional gene interactions from knockout fitness screens in cancer cells. Life Sci Alliance. 2019;2:e201800278.
9. Wainberg M, Kamber RA, Balsubramani A, Meyers RM, Sinnott-Armstrong N, Hornburg D, et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat Genet. 2021;53:638–49.
10. Broad Institute. DepMap: The Cancer Dependency Map. 2019.
11. Kim CY, Baek S, Cha J, Yang S, Kim E, Marcotte EM, et al. HumanNet v3: an improved database of human gene networks for disease research. Nucleic Acids Res. 2022;50:D632–9.
12. Kim E, Hart T. Improved analysis of CRISPR fitness screens and reduced off-target effects with the BAGEL2 gene essentiality classifier. Genome Med. 2021;13:2.
13. Morgens DW, Wainberg M, Boyle EA, Ursu O, Araya CL, Tsui CK, et al. Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. Nat Commun. 2017;8:15178.
14. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49:1779–84.
15. Dempster JM, Boyle I, Vazquez F, Root DE, Boehm JS, Hahn WC, et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. Genome Biol. 2021;22:343.
16. Allen F, Behan F, Khodak A, Iorio F, Yusa K, Garnett M, et al. JACKS: joint analysis of CRISPR/Cas9 knockout screens. Genome Res. 2019;29:464–71.
17. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. 2014;12.
18. Lenoir WF, Morgado M, DeWeirdt PC, McLaughlin M, Griffith AL, Sangree AK, et al. Discovery of putative tumor suppressors from CRISPR screens reveals rewired lipid metabolism in acute myeloid leukemia cells. Nat Commun. 2021;12:6506.
19. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. Science. 2004;306:1555–8.
20. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39(2):D691–7.
21. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
23. Rahman M, Billmann M, Costanzo M, Aregger M, Tong AHY, Chan K, et al. A method for benchmarking genetic screens reveals a predominant mitochondrial bias. Mol Syst Biol. 2021;17:e10013.
24. Drew K, Wallingford JB, Marcotte EM. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Mol Syst Biol. 2021;17.
25. Kim E, Novak LC, Lin C, Colic M, Bertolet LL, Gheorghe V, et al. Dynamic rewiring of biological activity across genotype and lineage revealed by context-dependent functional interactions. Genome Biol. 2022;23:140.
26. Han K, Pierce SE, Li A, Spees K, Anderson GR, Seoane JA, et al. CRISPR screens in cancer spheroids identify 3D growth-specific vulnerabilities. Nature. 2020;580:136–41.
27. Solinger JA, Spang A. Tethering complexes in the endocytic pathway: CORVET and HOPS. FEBS J. 2013;280:2743–57.
28. Balderhaar HJK, Ungermann C. CORVET and HOPS tethering complexes—coordinators of endosome and lysosome fusion. J Cell Sci. 2013;126:1307–16.
29. Sancak Y, Bar-Peled L, Zoncu R, Markhard AL, Nada S, Sabatini DM. Ragulator-Rag complex targets mTORC1 to the lysosomal surface and is necessary for its activation by amino acids. Cell. 2010;141:290–303.
30. Pan J, Kwon JJ, Talamas JA, Borah AA, Vazquez F, Boehm JS, et al. Sparse dictionary learning recovers pleiotropy from human cell fitness screens. Cell Syst. 2022;13:286-303.e10.
31. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44:W90–7.
32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

Publisher's Note