

# Predicting functional gene interactions with the hierarchical interaction score

Berend Snijder<sup>1,3</sup>, Prisca Liberali<sup>1</sup>, Mathieu Frechin<sup>1</sup>, Thomas Stoeger<sup>1,2</sup> & Lucas Pelkmans<sup>1</sup>

**Systems biology aims to unravel the vast network of functional interactions that govern biological systems. To date, the inference of gene interactions from large-scale 'omics data is typically achieved using correlations. We present the hierarchical interaction score (HIS) and show that the HIS outperforms commonly used methods in the inference of functional interactions between genes measured in large-scale experiments, making it a valuable statistic for systems biology.**

With the ever-increasing quantification of biological systems comes the need for powerful methods to analyze relationships between their measured variables. Statistical relations between genes measured from large-scale experiments can be used to predict functional or physical interactions, thereby suggesting gene function and regulatory pathways in healthy tissue<sup>1–6</sup> and subsequently leading to new insights in disease<sup>7–9</sup>. Here we present the HIS, which falls within a category of methods that use subsets or nested effects in the analysis of large-scale data sets<sup>10,11</sup>; among these methods are nested effects models (NEMs)<sup>10</sup>. Unlike previous methods, the HIS identifies directed hierarchical relationships between pairwise variables on the basis of the combined evidence for the hierarchical relationship and the phenotype strengths, and it applies to large data sets of diverse origins. The HIS can be calculated and visualized online at <http://www.his2graph.net/>, where we further offer source code and all published resources for reproduction of the results presented here. Source code is also available as **Supplementary Software**.

We developed the HIS to infer pairwise interactions between genes studied in parallel RNAi screens<sup>12,13</sup>. For RNAi screens, the subset principle of the HIS can be explained by the following example (**Fig. 1a**). Gene 'A', which is a hit in a given set of screens, is placed upstream of gene 'B' in the phenotypic hierarchy if B is a hit with the same sign for an exact subset of those screens and if there is no gene 'C' with an intermediate subset of hits; this method thereby avoids indirect interactions (Online Methods).

In addition to the HIS, we evaluated cosine and Pearson correlations, which essentially find linear relations between variables. Correlations have been applied most successfully for identifying gene interactions from yeast and bacterial gene knockout studies<sup>2–4</sup>, and because of this success they have been further applied in the analysis of parallel<sup>14</sup> and combinatorial<sup>6,15,16</sup> RNAi screens in human cells. We also evaluated the maximal information coefficient<sup>17</sup> (MIC), which captures a wider range of relations between variables. The performance of NEMs could not be systematically compared owing to technical limitations (**Supplementary Results**).

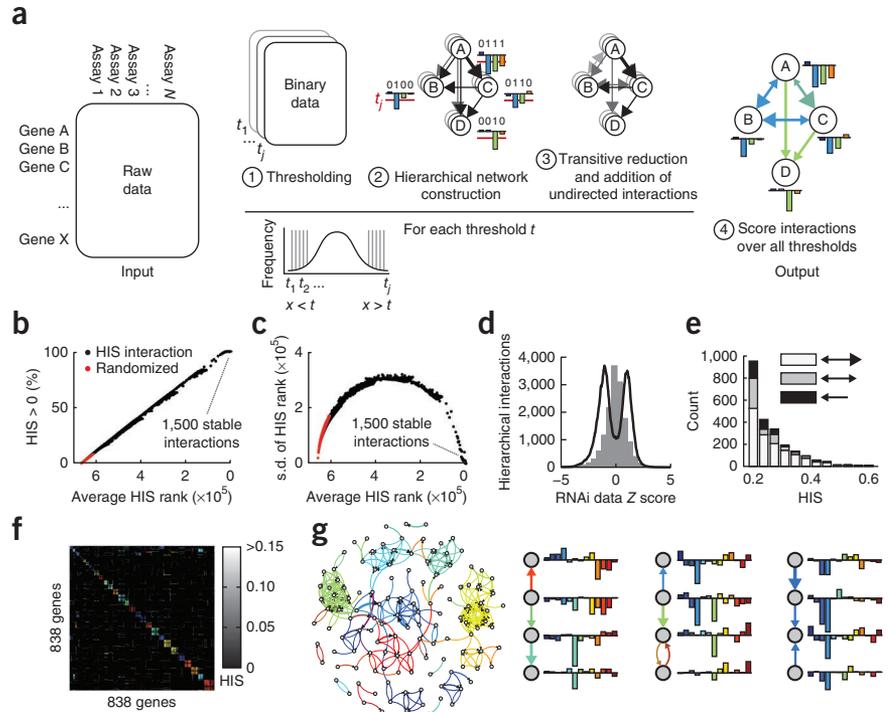
The HIS is robust with respect to parameter changes (**Fig. 1b,c**), missing data, large heterogeneous data sets and different data distributions (**Supplementary Fig. 1**; see also **Supplementary Results** and <http://www.his2graph.net/>). The resulting HIS network and the direction of the interactions reflect the phenotypic hierarchies present in the data (**Fig. 1d,e**). This leads to hierarchically organized modular networks (**Fig. 1f**), features commonly observed in biological networks<sup>18,19</sup>. The top-scoring hierarchical interactions are inferred between genes with the highest number of strong and nested phenotypes (**Supplementary Fig. 2**) and, as a consequence, typically connect multiple subnetworks (**Fig. 1g**).

We benchmarked the HIS and other methods in the retrieval of known interactions from four sets of parallel RNAi screens: our endocytome data set, which contains 13 image-based RNAi screens on 1,132 genes in endocytic activities and organelles in human cells (unpublished data, P.L., B.S. and L.P.); our largely unpublished infectome data set, which contains seven RNAi screens on the human druggable genome in virus infection<sup>12,13</sup>; and the publicly available human and fly RNAi screening results collected in the GenomeRNAi database<sup>20</sup>. In the endocytome and infectome data sets, we applied a normalization of population context-determined effects<sup>21</sup>, thereby consistently improving both the data quality and the cross-comparability of the different screens<sup>13</sup>. In contrast with benchmarks on synthetic data<sup>22</sup>, large-scale experimental data sets such as these do not have ground-truth interactions that should be inferred. We used three independent and publicly available descriptions of known gene interactions as the reference data sets: (i) genes sharing coannotations retrieved from DAVID<sup>23</sup> (v.6.2), (ii) genes with functional interactions retrieved from the search tool for the retrieval of interacting genes/proteins (STRING)<sup>24</sup> (v.9) and (iii) proteins with physical interactions retrieved from Pathway Commons<sup>25</sup> (Online Methods). Surprisingly, we found relatively little overlap in the reported gene interactions from the three sources (**Fig. 2a** and **Supplementary Fig. 3**), a result emphasizing the importance of using all three reference data sets in our benchmarks.

<sup>1</sup>Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. <sup>2</sup>Systems Biology PhD program, Life Science Zurich Graduate School, ETH Zurich and University of Zurich, Zurich, Switzerland. <sup>3</sup>Present address: CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. Correspondence should be addressed to B.S. ([bsnijder@cemm.oeaw.ac.at](mailto:bsnijder@cemm.oeaw.ac.at)) or L.P. ([lucas.pelkmans@imls.uzh.ch](mailto:lucas.pelkmans@imls.uzh.ch)).

**Figure 1** | Design and properties of the HIS.

(a) Global outline of the HIS algorithm. (b) The fraction of times an interaction is retrieved (at  $HIS > 0$ ) is plotted against the average interaction rank (where rank 1 is the highest-scoring interaction) calculated over different parameter settings on the endocytome data. Red dots indicate statistics of randomized interactions. (c) Same as in b but showing the s.d. of the HIS ranks on the y axis. (d) Number of hierarchical interactions inferred at each threshold of the endocytome RNAi data set (black line). Gray bars are a histogram of the input data (corresponding axis not shown). (e) Number of different interactions per HIS category. Black, fully hierarchical; gray, equal; white, unequal hierarchy. (f) HIS represented as heat map for the 838 genes of the endocytome data set with interactions. Intensity represents the HIS, and colors are determined according to the contributing assays (as in Fig. 1a). (g) Graph visualization of the HIS for values above 0.3 as inferred on the endocytome data set. Visualization is as on <http://www.his2graph.net/>. Selected interactions are shown with their corresponding 13 RNAi phenotypes displayed as bar graphs.



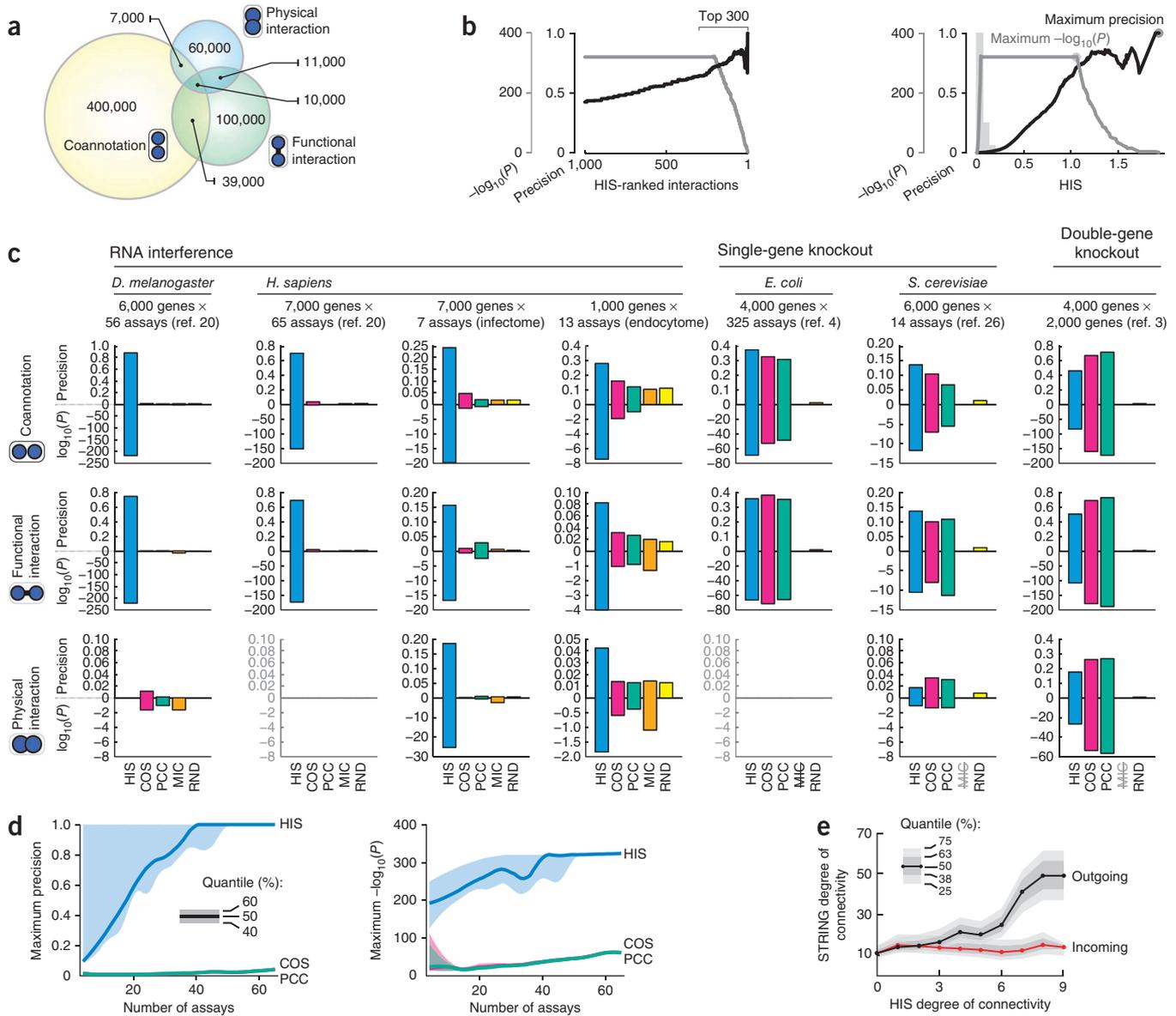
To compare the performance of the different methods, we calculated the precision (i.e., the true positive rate) and corresponding  $P$  values (based on the hypergeometric distribution) for the top-predicted interactions at each position of the ranked list, starting from the highest scoring interaction (Fig. 2b and Online Methods). When we calculated the precision and significance of the top 100 hierarchical interactions from the fly GenomeRNAi data set and compared them against the functional interactions reported in STRING, we found a precision of 79%, meaning that 79 of the top 100 predicted interactions were also reported in STRING. This is a 244-fold enrichment over background ( $P < 10^{-175}$ ). To globally compare the performance of the HIS with that of the other methods, we calculated the average precision and significance values up to the top 300 predicted interactions from HIS and four other methods (cosine and Pearson correlations, MIC and random predictions) (Fig. 2c), an analysis strategy that rewards methods that assign higher scores to confirmed interactions (Online Methods). For all four independent sets of parallel RNAi screens of human (*Homo sapiens*) and fly cell (*Drosophila melanogaster*) origin, the HIS strongly outperformed the other tested methods in the inference of known functional and physical interactions (Fig. 2c).

We next evaluated the performance of the HIS in retrieving known gene interactions from single- and double-gene-knockout screens performed in bacteria (*Escherichia coli*)<sup>4</sup> and yeast (*Saccharomyces cerevisiae*)<sup>3,26</sup> (Fig. 2c). The MIC could not be tested on these data sets, as the calculations repeatedly failed to finish (crashed). We found that, consistent with previous reports<sup>2-4</sup>, correlation-based methods retrieved known functional gene interactions and coannotated genes from parallel gene-knockout screens in bacteria and yeast in a significant manner (Fig. 2c), with roughly equal precision to the HIS for the single-gene-knockout data sets. No method inferred known physical interactions for the single-gene-knockout data sets on a significant level. On measurements of yeast growth for double-gene

knockouts on a single medium<sup>3</sup>, the HIS retrieved known gene and protein interactions at very high significance ( $P < 10^{-100}$  for functional interactions). However, only on this data set did we find that the correlation-based methods outperformed the HIS for retrieval of known interactions (Fig. 2c). Interestingly, the top-ranking interactions predicted by HIS were largely different from those predicted by the correlation-based methods, a result suggesting that the HIS may be complementary to correlation-based methods in the inference of functional interactions from single- and double-gene-knockout experiments in yeast and bacteria. We examined the robustness of the different methods to incremental dilution of the true signal with Gaussian noise (Supplementary Fig. 4). The HIS was significantly more robust to noise than the correlation-based methods, even on those data for which the correlation-based methods initially outperformed the HIS.

As both NEMs<sup>10</sup> and HIS infer interactions on the basis of subsets of phenotypes, we compared performance of both methods on the endocytome data set. Analysis of the other data sets with the NEM was not possible owing to its dependency on  $P$  value-transformed data, and the NEM-based transitive reduction could not be applied to any of the data sets because the network sizes were too big. The HIS significantly outperformed the NEM (Supplementary Fig. 5). A systematic analysis of the contribution of individual HIS features confirmed that subset effects alone do not prioritize known functional interactions between genes from parallel RNAi screens. This prioritization of known interactions requires the integration of phenotype width and strength and transitive reduction at multiple thresholds (Supplementary Fig. 6), which are unique features of the HIS.

We determined whether the high precision of the HIS on the *D. melanogaster* GenomeRNAi<sup>20</sup> data set was dependent on any single RNAi screen by bootstrapping over randomly selected sets of 3–65 RNAi screens. When we calculated the bootstrapped maximum precision and maximum  $-\log_{10}$ -transformed  $P$  value as performance measures (Fig. 2d), we found that both measures



**Figure 2** | HIS performs best in the inference of functional interactions across species. **(a)** Venn diagram of the overlap of reference interactions of the druggable genome screened in the infectome data set<sup>12,13</sup>. **(b)** Performance of the HIS inferred interactions from the *D. melanogaster* GenomeRNAi<sup>20</sup> data, when benchmarked to functional interactions (STRING<sup>24</sup> v.9 with scores above 400). Significance and precision are shown per rank (left) and per HIS (right). **(c)** Average statistics for the top 300 interactions are given for the seven data sets<sup>3,4,12,13,20,26</sup> (columns) when validated against the three reference sets<sup>23–25</sup> (rows). MIC<sup>17</sup> with gray strikethrough indicates that this calculation could not be completed. Gray graphs indicate a lack of reference interactions. COS, cosine correlation; PCC, Pearson correlation; RND, random. **(d)** Maximum precision (left) and maximum  $-\log_{10}(P)$  for the bootstrapped inference of interactions from the human GenomeRNAi data set. Median bootstrap values are indicated by solid lines; 40th and 60th percentiles are indicated by the lighter shading. **(e)** Degree of connectivity in STRING v.9 for genes with increasing outgoing edges (black) and incoming edges (red) as calculated by the HIS from the endocytome data set. Shaded regions (legend) indicate bootstrap statistics resulting from shifting the STRING interaction score threshold from 1 to 1,000.

increased steeply for the HIS as the number of analyzed RNAi screens increased, a result independent of any individual RNAi screen. In comparison, both correlation methods showed a much lower maximum precision and  $-\log_{10} P$  value, with hardly any increase in precision as more screens were analyzed. The same analysis on all possible sets of the endocytome screens reproduced this result (**Supplementary Fig. 7**). We consistently found that precision of the HIS results was higher for top-scoring hierarchical interactions, and this was also independent of any specific set of RNAi screens used (**Supplementary Fig. 8**). Consequently, we conclude

that individual interactions subjected to follow-up studies should be inferred from as many independent experimental data sets as possible and selected from the top-scoring HIS interactions to maximize the likelihood of the interactions being true. Furthermore, visualization of a top-scoring HIS network can be based on a lower HIS threshold that maximizes the significance of the predicted set, as was done in **Figure 1g** and **Supplementary Figure 9**.

Upon analyzing the properties of incoming and outgoing hierarchical interactions, we found that the degree of outgoing

hierarchical interactions, but not the degree of incoming hierarchical interactions, positively correlated with the STRING degree of connectivity (Fig. 2e and Supplementary Results). As more essential genes have been shown to have a higher degree of connectivity in interaction networks<sup>3,5,27,28</sup>, this may suggest that genes placed upstream in the phenotypic hierarchy might have a higher degree of connectivity owing to their broader involvement in cellular activities. We additionally compared the hierarchical interactions with computationally predicted directionality of signal flow from membrane receptors to transcription factors from protein-protein interactions<sup>29</sup>, and we found that agreement of inferred directionality depended on the type and sign of the interaction (Supplementary Fig. 10 and Supplementary Results).

We also applied the HIS to the analysis of transcriptomics in *S. cerevisiae*<sup>22</sup> (Supplementary Fig. 11) and immunohistochemical characterization of protein levels from human cancer samples<sup>30</sup> (Supplementary Fig. 12). As on the other data sets, the HIS greatly outperformed the correlation-based methods in the inference of gene interactions. The resulting interactions learned from variations in protein expression levels of 20 different human cancer types may be relevant for understanding the mechanism of human cancer (Supplementary Fig. 13).

Finally, we compared predicted interactors of protein tyrosine kinase 2 (encoded by *PTK2*, or *FAK*) inferred from the endocytome data set with comparative phosphoproteomics (Supplementary Table 1) and transcriptomics analysis of *PTK2* knockout, wild-type and rescue cell lines (Supplementary Fig. 14 and Online Methods). Both transcriptomics and proteomics identified similar processes typically associated with *PTK2* function (Supplementary Fig. 14 and Supplementary Table 2). A comparative network analysis of the predicted interactors showed the strongest association of the HIS predicted interactors to the experimental results, and strengthened a HIS-predicted role for *PTK2* in lysosome homeostasis (Supplementary Fig. 14 and Supplementary Results).

We present here a new statistic that accurately identifies relevant biological interactions from various 'omics data. The power of the HIS across all tested data sets<sup>3,4,12,13,20,22,26,30</sup>, and the low overlap with previously inferred interactions from these data sets, suggests that the set of hierarchical interactions is a largely unexplored resource for unraveling biological complexity. The improved performance of the HIS on parallel RNAi screens may be explained by the more complex organization of functional gene interactions in cells from multicellular organisms, by the reduced complexity of phenotypes measured solely from colony fitness data, and by the elimination of indirect hierarchical interactions, which recently was shown to also improve the predictive power of correlative analyses<sup>31</sup>. Whereas correlation-based methods appear useful for identifying interactions between components within one protein complex, cellular structure or cellular activity, the HIS explicitly identifies functional interactions between genes that function in multiple cellular activities, which likely represent the majority of regulatory interactions in higher organisms. Driven by the advent of systematic characterizations of genome-wide gene-knockout libraries in mammalian cells<sup>32,33</sup>, the HIS will greatly aid systematic inference of functional interactions from large-scale gene-knockout experiments in higher organisms.

## METHODS

Methods and any associated references are available in the [online version of the paper](#).

**Accession codes.** NCBI Gene Expression Omnibus: the microarray data set is available at accession [GSE43873](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

We would like to acknowledge E.-M. Damm and A. Schmidt for the phosphoproteomics data, A. Patrignani for analysis of the microarray data, Y. Yakimovich for help on the accompanying website, D. Schlaepfer (University of California, San Diego) for the PTK2-rescue cell line, F. Markowetz and X. Wang for help with NEM analysis, and all members of the Pelkmans lab for useful comments on the manuscript. L.P. acknowledges financial support from the SystemsX.ch RTD projects PhosphoNetX and LipidX and the University of Zurich, and B.S. acknowledges financial support from the Swiss National Science Foundation.

## AUTHOR CONTRIBUTIONS

B.S. and L.P. conceived of the study. B.S. developed the method and performed computational analyses. P.L., M.F. and T.S. performed experiments. B.S. and L.P. wrote the manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Moffat, J. & Sabatini, D.M. *Nat. Rev. Mol. Cell Biol.* **7**, 177–187 (2006).
- Tong, A.H. *et al. Science* **303**, 808–813 (2004).
- Costanzo, M. *et al. Science* **327**, 425–431 (2010).
- Nichols, R.J. *et al. Cell* **144**, 143–156 (2011).
- Gavin, A.C. *et al. Nature* **415**, 141–147 (2002).
- Horn, T. *et al. Nat. Methods* **8**, 341–346 (2011).
- Bernards, R. *Cell* **151**, 465–468 (2012).
- Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. *Mol. Syst. Biol.* **3**, 140 (2007).
- Lehner, B. *J. Exp. Biol.* **210**, 1559–1566 (2007).
- Markowetz, F., Kostka, D., Troyanskaya, O.G. & Spang, R. *Bioinformatics* **23**, i305–i312 (2007).
- Boutros, M., Agaisse, H. & Perrimon, N. *Dev. Cell* **3**, 711–722 (2002).
- Mercer, J. *et al. Cell Rep.* **2**, 1036–1047 (2012).
- Snijder, B. *et al. Mol. Syst. Biol.* **8**, 579 (2012).
- Wang, X., Castro, M.A., Mulder, K.W. & Markowetz, F. *PLoS Comput. Biol.* **8**, e1002566 (2012).
- Lauffer, C., Fischer, B., Billmann, M., Huber, W. & Boutros, M. *Nat. Methods* **10**, 427–431 (2013).
- Roguev, A. *et al. Nat. Methods* **10**, 432–437 (2013).
- Reshef, D.N. *et al. Science* **334**, 1518–1524 (2011).
- Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabási, A.-L. *Science* **299**, 1551–1555 (2002).
- Ryan, C.J. *et al. Mol. Cell* **46**, 691–704 (2012).
- Schmidt, E.E. *et al. Nucleic Acids Res.* **41**, D1021–D1026 (2013).
- Snijder, B. *et al. Nature* **461**, 520–523 (2009).
- Marbach, D. *et al. Nat. Methods* **9**, 796–804 (2012).
- Huang, D.W. *Nat. Protoc.* **4**, 44–57 (2009).
- Szklarczyk, D. *et al. Nucleic Acids Res.* **39**, D561–D568 (2011).
- Cerami, E.G. *et al. Nucleic Acids Res.* **39**, D685–D690 (2011).
- Steinmetz, L.M. *et al. Nat. Genet.* **31**, 400–404 (2002).
- Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. *Nature* **411**, 41–42 (2001).
- Butland, G. *et al. Nature* **433**, 531–537 (2005).
- Vinayagam, A. *et al. Sci. Signal.* **4**, rs8 (2011).
- Uhlen, M. *et al. Nat. Biotechnol.* **28**, 1248–1250 (2010).
- Barzel, B. & Barabasi, A.L. *Nat. Biotechnol.* **31**, 720–725 (2013).
- Bürkstümmer, T. *et al. Nat. Methods* doi:10.1038/nmeth.2609 (25 August 2013).
- Mali, P. *et al. Science* **339**, 823–826 (2013).

## ONLINE METHODS

**Data and code availability.** HIS-predicted interactions from the experimental data sets and source code can be found online at the accompanying website: <http://www.his2graph.net/>. This website additionally gives access to all published experimental data sets and their corresponding reference sets, as well as to the source code of the HIS method. Source code is also available as **Supplementary Software**. The microarray data set has been uploaded to the NCBI Gene Expression Omnibus as record [GSE43873](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43873). The phosphoproteomics results are available in **Supplementary Table 1**.

**Omics and reference data sets.** All published experimental data sets analyzed in this study were downloaded from their respective online resources and normalized such that the most abundant phenotypes have values around 0. Reference data sets from DAVID<sup>23</sup> version 6.2, STRING<sup>24</sup> version 9 and Pathway Commons<sup>25</sup> were downloaded from their respective online sources. DAVID annotation tables were downloaded for the full set of proteins or genes analyzed at default DAVID settings, and genes were scored as being coannotated when they share two or more annotations that contain maximally 90 genes or proteins within the analyzed data set. This limit was set to avoid ubiquitous annotations (for example, “splice variant”) from leading to spurious interactions. STRING interactions were considered with a STRING interaction score above 400, unless stated otherwise. All ‘omics and corresponding reference data sets are available for download from <http://www.his2graph.net/>. Homology mapping between mouse and human genes for the *PTK2* analysis was performed using BioMart.

**The hierarchical interaction score.** The HIS works as outlined in **Figure 1a** and as described above: a data set (of size  $m \times n$ ) is discretized at a set of thresholds into 0s (nonphenotypes) and 1s (hits or phenotypes). At each threshold, all hierarchical relations are inferred between each pair of variables (rows) of the data set, where a hierarchical relation between two variables is defined by a variable having an exact subset of ones of the other variable. Such an interaction is directed from the variable with the most 1s toward the variable with the subset of 1s. Transitive reduction is applied to remove ‘shortcuts’ or indirect links in the graph at each threshold, reducing the graph to its minimal representation. For each interaction and at each threshold, an intermediate score is kept that is the number of phenotypes that the two variables have in common. At each threshold, variables with identical phenotypes get assigned a bidirectional interaction, scored as well for the number of phenotypes that the variables have in common. Finally, the HIS is calculated as the mean number of phenotypes in common for each interaction in each direction over all thresholds. See <http://www.his2graph.net/> for the source code of the calculation. The calculation has been optimized in its use of memory and for speed.

**Benchmarking statistics.** We calculated the precision and its corresponding *P* value at each rank of the sorted list of predicted interactions, where rank 1 represents the single highest-scoring interaction. We use the term ‘precision’ for a set of predicted set of interactions being the fraction of interactions that are also present in the reference data set. In the absence of a true ground truth,

the more commonly used ‘true positive rate’ and accompanying statistics (ROC curves, and their area-under-curve statistic) are not applicable (see also **Supplementary Fig. 15**). Corresponding *P* values are calculated as the probability of getting  $x$  positive interactions out of a total predicted set of interactions of size  $y$ , where there are a total possible  $k$  interactions, for which the reference has  $m$  positive interactions. Given these parameters, the probability is given by the hypergeometric probability function. Interactions with the same score (common, for instance, for correlations based on few data points and for the MIC) are all given the probability and precision for the set of predicted interactions that includes all interactions with this same score. *P* values are calculated for undirected interactions, and self-connections are excluded from the analysis. This means that for a data set of  $m$  rows (or variables), the total number of possible interactions equals  $m \times (m - 1)/2$ . See <http://www.his2graph.net/> for a fast implementation of the precision and *P*-value calculation.

For comparison of the different inference methods, we compared their average statistics of the highest 300 to the highest 1 scoring interaction(s), which is calculated as the average of the precision and significance statistics of the top 300, top 299, top 298, ..., to the top 1 predicted interaction(s). This average top 300 statistics favors results in which confirmed interactions are more abundant and ranked closer to 1. As discussed in the main text, these are both desirable properties for methods that infer interactions that require further experimental validation. The conclusions drawn from the global validation results shown in **Figure 2c** do not change when the top 1,000 predicted interactions are used.

**Human Protein Atlas data normalization.** Human Protein Atlas cancer data were downloaded from <http://www.proteinatlas.org/>. For each protein, the manually annotated labels “negative,” “weak,” “moderate” and “strong” were converted to the numerical values ranging from 1 to 4, respectively, and averaged over all samples (from different patients) per protein and cancer type. These values were next z-score normalized (subtract the mean of the set and divide by its s.d.) for each cancer type (column) first, and subsequently for each protein (row), to emphasize difference in protein levels among cancers.

**Cell lines.** Wild-type mouse embryonic fibroblasts (MEFs) (*PTK2*-WT) were obtained from ATCC, CRL-2645. MEFs knocked out for focal adhesion kinase gene (*PTK2*<sup>-/-</sup>) were obtained from ATCC, CRL-2644. MEFs rescued by overexpression of the focal adhesion kinase gene in a *PTK2*<sup>-/-</sup> background (*PTK2*-RESC) were obtained from D. Schlaepfer<sup>34</sup>, clone DA2.

**Microarray.** For dense condition  $8.2 \times 10^6$  cells were seeded, whereas for sparse condition  $0.41 \times 10^6$  cells were seeded. Cells were grown for 24 h, harvested by trypsinization and frozen at  $-80^\circ\text{C}$ . RNA preparation was done with the Qiagen RNeasy Mini Kit. The quality of the isolated RNA was determined (NanoDrop ND 1000 and Bioanalyzer 2100) and only the samples with a 260 nm/280 nm ratio between 1.8 and 2.1 and an RNA integrity number higher than 8 were further processed. Total RNA samples were reverse transcribed into double-stranded cDNA in presence of RNA poly(A) controls from the RNA Spike-In Kit (Agilent). The double-stranded cDNA were *in vitro* transcribed in the

presence of Cy3-labeled nucleotides using a Low Input Quick Amp Labeling Kit. The Cy3-cDNA was purified using a RNeasy Mini Kit, and only cDNA samples with a total cDNA yield higher than 2 µg and a dye incorporation rate between 8 pmol µg<sup>-1</sup> and 20 pmol µg<sup>-1</sup> were considered for hybridization. Cy3-labeled cRNA samples were randomly fragmented to 100–200 bp at 65 °C, and target cRNA samples were hybridized to Whole Mouse Genome 4x44k OligoMicroarrays for 17 h at 65 °C. Raw data processing was performed using the Agilent Scan Control and the Agilent Feature Extraction Software Version 10. Quality-control measures were considered before statistical analysis was performed. These included inspection of the array hybridization pattern, proper grid alignment, performance of the spike-in controls (linear dynamic range between five orders of magnitude) and number of green feature nonuniformity outliers (below 100 required for all samples).

**Phosphoproteomics analysis.** *PTK2*-WT, *PTK2*-KO and *PTK2*-RESC were grown in DMEM (+10% FCS), reaching a final concentration of approximately 5 × 10<sup>5</sup> cells per ml after 48 h, corresponding to a nonconfluent state. Cells were washed with 1× PBS and disrupted in lysis buffer (150 mM NaCl, 50 mM Tris-HCl, pH 7.2, 10 mM EDTA, 1 mM Na<sub>3</sub>VO<sub>3</sub>, 200 mM okadaic acid, 2 mM calyculin A, 1 mM PMSE, 0.1% Rapigest). The purification and enrichments of phosphorylated peptides were performed with TiO<sub>2</sub>, and mass spectrometry was performed in triplicate. Both purification and measurement were performed as

previously described<sup>35</sup>. Significance was calculated by comparing the triplicates with a two-tailed *t*-test.

**Western blot analysis.** *PTK2*-WT, *PTK2*-KO and *PTK2*-RESC were grown, reaching a final concentration of approximately 5 × 10<sup>5</sup> cells per ml after 48 h, corresponding to a nonconfluent state. Cells were washed with 1× PBS and disrupted in lysis buffer (0.5% sodium deoxycholate, 150 mM NaCl, 50 mM Tris-HCl, pH 7.2, 0.1% SDS, 1% Triton X-100, 0.2% NaN<sub>3</sub>), and 15 µg of each protein extract was separated using 10% PAGE. Separated proteins were then transferred onto a membrane (Immobilon-P, 0.45 µm, Millipore) using the humid chamber method. Membranes were saturated with 5% milk proteins in 1× TBS-T (1× TBS, 0.1% Tween) for 1 hour. Mouse anti-Lamp1 (#555798, BD Biosciences) and rabbit anti-tubulin (Ab6046, Abcam) primary antibodies as well as HRP-conjugated secondary anti-mouse (#170-6516, BioRad) and anti-rabbit (#170-6515, BioRad) antibodies were diluted at respectively 1:1,000, 1:2,000, 1:5,000 and 1:5,000 in the same buffer. Primary and secondary antibodies were applied for 90 min and 60 min, respectively. Signal was revealed with HRP substrate solution, imaged with a CCD camera and quantified with ImageJ. Owing to heavy glycosylation of Lamp1, our mouse anti-Lamp1 signal was over 90 kDa (predicted MW: 44 kDa), consistent with the product specifications for this antibody.

34. Sieg, D.J. *et al.* *EMBO J.* **17**, 5933–5947 (1998).

35. Bodenmiller, B., Mueller, L.N., Mueller, M., Domon, B. & Aebersold, R. *Nat. Methods* **4**, 231–237 (2007).