bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Inferability of transcriptional networks from large scale gene deletion studies

C. F. Blum[1,2,3], N. Heramvand[1,2,3], A. S. Khonsari[1], & M. Kollmann[1]

[1] *Institute for Mathematical Modeling of Biological Systems, Heinrich Heine University of Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany*

[2] *Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Köln, Germany*

[3] *Equally contributing authors*

Generating a comprehensive map of molecular interactions in living cells is difficult and great efforts are undertaken to infer molecular interactions from large scale perturbation experiments. Here, we develop the analytical and numerical tools to quantify the fundamental limits for inferring transcriptional networks from gene knockout screens and introduce a network inference method that is unbiased and scalable to large network sizes. We show that it is possible to infer gene regulatory interactions with high statistical significance, even if prior knowledge about potential regulators is absent.

The functionality of a living cell is determined by the interplay of multiple molecular components that interact with each other. Generating a global map of these molecular interactions is an essential step to advance our understanding of the molecular mechanisms behind disease, development, and the reprogramming of organisms for biotechnological applications [1]. The current advances in gene editing methods [2] have scaled up the size of genome-wide single and double

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

knockout libraries, ranging from microbes [3,4] to higher eukaryotes [5] and open up a much more informative data source than inferring gene regulatory networks from unspecific perturbations, such as stress or changes in growth conditions [6]. However, the detection of direct interactions between two genes from association measures – for example the covariance between transcript levels – remains a highly non-trivial task, given the significant noise among biological replicates, the frequent case where the number of parameters exceeds the number of independent data points, and the high dimensionality of the inference problem.

The existence of a direct interaction between gene A as a source of regulation (source node) and gene B as a target of regulation (target node) can be detected if a significant part of the transcriptional activity of B can be explained by the transcriptional activity of A but not by the activities of the remaining genes in the network. Thus a necessary condition for identifiability or inferability of links is the knowledge about the information that can be transmitted by alternative routes in the network, which can be obtained by perturbing the involved nodes [7]. As most gene perturbation screens are incomplete – for example due to the fact that essential genes cannot be knocked out – we have in general the situation that a significant amount of interactions within an N-gene network are non-inferable, regardless of the amount of experimental replicates and the strength of perturbations. Direct interactions inferred from transcriptome data typically oversimplify the molecular complexity behind gene regulation, which frequently involves protein-protein interactions and modifications on protein or DNA level.

To calculate the maximum number of links that can be inferred from knockout screens in

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

absence of other constraints, we consider a directed network of N nodes, with node activities as observables and a fraction q of node activities strongly perturbed by external forces. Given that the perturbed nodes are randomly distributed within the network, we can analytically calculate the expected fraction of inferable links, $F(q)$, by a counting procedure illustrated in Fig. 1a (Online Methods and Supplementary Note 1).

As $F(q)$ is an upper bound for the expected number of directed links that can be inferred from stationary node activities, we now ask how this bound is related to the structural properties of the network. To compare different network architectures, it is useful to define the network inferability, I_F, as the area under the $F(q)$-curve, $I_F := \int_0^1 F(q)dq$. Comparison of I_F between two general classes of network structures with node degrees either power law or Poisson distributed shows that networks that are enriched with nodes of high out degree (outgoing hubs) are the most difficult ones to infer (Fig. 1b). Differences in inferability due to network structure are most predominant for networks with low mean degree and become less predominant with high mean degree (Fig. 1c). As our measure of inferablity, I_F, is essentially determined by the outdegree distribution the curve starts saturating for scale-free exponents $\gamma > 3$, as in this regime the variance of the number of links per node is essentially constant for increasing γ and fixed mean degree [8] (Fig. 1d). The network inferability, I_F, is asymptotically independent of network size (Fig. 1e). We further investigated the inferability of causal interactions in biological and social networks as a function of the mean degree (Fig. 1f). The decreasing trend can be explained by the higher number of alternative routes that come with a stronger connected network. The low inferability of gene regulatory networks can be attributed to master regulators that regulate a large fraction of the genome (hubs

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with high outdegree), whereas the comparatively high inferability of protein interaction networks is a consequence of the low number of different binding domains per protein and that only a fraction of the existing interactions have been identified due to limitations of experimental methods [9]. If we assume that the probability $P(k, l, m|k \to l)$ can be factorized, the resulting inferability measure, I_F^*, is simply a function of the outdegree distributions, $P(k)$ and $P(l)$. We observed that $I_F^* \approx I_F$ for all networks investigated in this work (Fig. 1f, inset). This result shows that for a large variety of networks structures the outdegree is the dominating factor that determines network inferability.

Inference of transcriptional networks on a genome scale is best realised by methods that are (i) asymptotically unbiased, (ii) scalable to large network sizes, (iii) sensitive to feed-forward loops [10], and (iv) can handle data sets with and without knowledge about which nodes are targeted by experimentally induced perturbations [7,11–15] (Supplementary Note 2). Inference methods for *directed* networks typically require individual perturbation of all nodes [7] or many perturbations of different strengths to compute conditional association measures [6,16] or conditional probabilities [17]. Generation of time course data seems to be the most natural way to infer directed networks by simply analysing the temporal ordering of the transcriptional activities [18,19]. However, this approach precludes the use of knockout experiments and requires fast acting perturbations in combination with monitoring node activities over time, which is experimentally demanding, especially if nodes respond on very different time scales [20].

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Inference is further complicated by the fact that transcriptome data contains a significant amount of stochastic variation between biological replicates despite pooling over millions of cells (Supplementary Fig. 1a). It is interesting to see that the variation across biological replicates for baker's yeast [3] is close to a normal distribution and follows almost exactly a t-distribution with 11 degrees of freedom over five standard deviations (Supplementary Fig. 1a, inset) and that biological noise is much larger than technical noise (Supplementary Fig. 1b). As biological noise may arise from subtle differences in growth conditions that induce changes in gene regulation, we expected to see significant cross-correlations among genes (Supplementary Fig. 1c), whose magnitude is much larger than expected by chance (Supplementary Fig. 1c, inset). These cross-correlations can be exploited for inferring the structure of undirected networks [11], if the driving noise is independent and identically distributed for all nodes (Supplementary Note 2). In contrast, technical noise not only reduces the statistical significance for detecting true interactions but can also induce a significant fraction of false positive interactions, especially if the interaction network under investigation is sparse. Such noise induced misclassification of links can be illustrated by a simple linear network $A \rightarrow B \rightarrow C$ for which standard inference methods interpret the information that A has about C erroneously as a direct link between A and C if the state of B is corrupted by measurement noise (Supplementary Fig. 1d). The reason is that a part of the correlation between A and C cannot be explained by B.

To make use of the rapidly growing amount of single-gene knockout screens for which transcriptome data is [3] or may become available [21,22], we developed a method to infer directed networks on a genome scale where the number of genetic perturbations is typically below the number

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of nodes or genes in the network (Online Methods). In brief, our method uses the concept of probabilistic principle component analysis [23] to compute partial response coefficients (PRC) that are asymptotically unbiased with respect to Gaussian measurement noise (Supplementary Fig. 1d). Additionally the algorithm provides a feature to identify non-inferable links, which are removed before statistical analysis. In absence of noise, our numerical method correctly predictes the fraction of links that are inferable, $F(q)$ (Supplementary Notes 1). To evaluate the performance of our method we generated two synthetic knockout data sets that closely resemble the gene regulatory network structure of baker's yeast, using the GeneNetWeaver software [24] that uses a hierarchical network structure and our own generative model that uses a scale free network structure (Supplementary Notes 3). We added Gaussian measurement noise to the synthetic data with a standard deviation of 10% the $\log_2$ fold-change in expression level for each perturbation for each gene. Residual bootstrapping among replicates was used to quantify the statistical significance of the inferred link strengths. In comparison with standard inference methods, such as partial correlations [11–13,25,26], our method shows a significantly higher performance in absence of any penalties that enforce sparse network structures (Fig. 2b, left panel). The improved performance of our method can be assigned to the fact that it is unbiased with respect to measurement noise (Online Methods).

To further improve the predictive power of our method we included the prior knowledge that transcriptional networks are highly sparse. Sparsity constraints are typically realized by penalising either the existence of links or the link strengths by adding appropriate cost functions, such as L^1-norm regularized regression (Lasso) [27]. Adding a cost function to the main objective comes with the problem to trade-off the log-likelihood against the number of links in the network whose

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

strength is allowed to be non-zero. In absence of experimentally verified interactions there is no obvious way how to determine a suitable regularization parameter that weights the likelihood against the cost function, which is one of the great weaknesses of such methods.

In our approach we reduce network complexity by assuming that functionally relevant information in molecular networks can only pass through nodes whose response to perturbations is significantly above their biological noise level. The individual noise levels can be estimated from natural variations between wild type experimental replicates (Supplementary Fig. 1). The significance level that removes nodes from the network with high noise-to-signal ratio can be set by the desired false discovery rate. It can be shown that removal of noisy nodes imposes a sparsity constraint on the inference problem (Online Methods). The different steps required to arrive at a list of significant links are illustrated in Fig. 2a. In the first step, genes are grouped in clusters that are co-expressed under all perturbations. These clusters are treated as single network nodes in the subsequent steps. In the second step, only those samples are extracted from the dataset that correspond to perturbation of a chosen gene – the source node – with no other genes perturbed (node 5 in Fig. 2a). From this reduced dataset, we identify all nodes in the network that change expression above a given significance level upon perturbing the source node. These significantly responding nodes define a subnetwork for each source node, which is typically much smaller in size than the complete network. In the third step, we collect all perturbation data from the complete dataset for all nodes that are part of the subnetwork. Before inferring a direct interaction that points from the source node to a given target node in the subnetwork, we remove the perturbation data of target node and all nodes that respond significantly to perturbations of the target node (green arrows in

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 2a). The second and third steps realise numerically the counting procedure of inferable links as illustrated in Fig. 1a. Significant links are identified by partial response coefficients in combination with residual bootstrapping over replicates (Online Methods). In the fourth step, we collect all clusters of co-expressed genes that contain two nodes in total with one of the nodes perturbed and check statistical significance of the directed link between them. In the fifth step, all significant links are collected in an edge list. We refer to these five steps as the *clustering method*. If we remove all links from the edge list that have more than one node as a source or more than one node as a target, we obtain an edge list that corresponds to links between single genes. This reduced edge list would also arise by skipping the clustering step and we refer to remaining inference steps that compute links between single genes as *subnetwork method*.

The Lasso method followed by bootstrapping has been benchmarked as one of the highest performing network inference methods for *in silico* generated expression data [28]. The receiver operating characteristic (ROC) curve of the subnetwork method shows better performance than the Lasso method (Fig. 2b, middle and right panel) after adjusting the regularization parameter of the Lasso method such that the area under the ROC curve is maximized. However, a significant performance boost for the Lasso method can be achieved by applying the second step of our method that removes noisy nodes, resulting in comparable performance of Lasso with the subnetwork method for the case that validation data exists such that the regularisation parameter can be determined (Fig. 2B, middle and right panel).

To get insight into the optimal experimental design for generating data for network inference,

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

we computed the fraction of correctly inferred links and compared them against the fraction of independently perturbed nodes for different numbers of experimental replicates. We compared three different variants of our approach: partial response coefficients (PRC), PRC together with subnetwork method, PRC together with clustering method (Fig. 4C). As all variants share PRC as underlying inference method (Online Methods), the observed strong increase in performance can be assigned to the sparsity constraint that comes with the subnetwork method or the clustering method. Due to this constraint, both the subnetwork method and the clustering method can have higher accuracy than the noise-free analytical solution, as the latter does not enforce sparse network structures. The results show that in presence of 10% measurement noise the amount of available replicates limits the true positive rate, even if 100% of nodes are perturbed. However, inference of more than 80% of the network can only be achieved if the number of replicates is sufficiently high.

To evaluate the performance of our approach on real data, we use one of the largest publicly available transcriptome data sets [3], compromising transcriptomes that cover 6170 genes for 1441 single gene knockouts. We use the galactose utilisation network as a gene regulatory example, which is one of the best characterised gene regulatory modules in yeast [29]. The regulatory mechanism of the GAL4 gene as a key regulator is shown in Fig. 4D, left panel. As information about phosphorylation and protein interaction is absent in expression data, the inferred network structure from transcriptome data with GAL4 and GAL80 perturbed is different from the known gene regulation. By sorting genes with respect to their number of statistically significant outgoing links, we can identify potential key regulators. Besides transcription factors, the regulators with highest statistical significance are factors involved in chromatin remodelling, signalling kinases,

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and genes involved in ubiquitination (Supplementary Tables 1-3). This result – although expected for eukaryotes – is inaccessible for inference methods that *a priori* fix known transcription factors as regulatory sources. However, as the 1441 knockout genes of this dataset compromise just 23% of the genes for which transcript levels have been measured, we can estimate from our simulations that we have inferred less than 10% of the direct interactions in the transcriptional network of yeast. However, the recently available double knockout screens in yeast [22] drastically increased the number of targeted perturbations, which can boost the prediction accuracy of network inference methods once transcriptome data for a fraction of these mutants become available.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

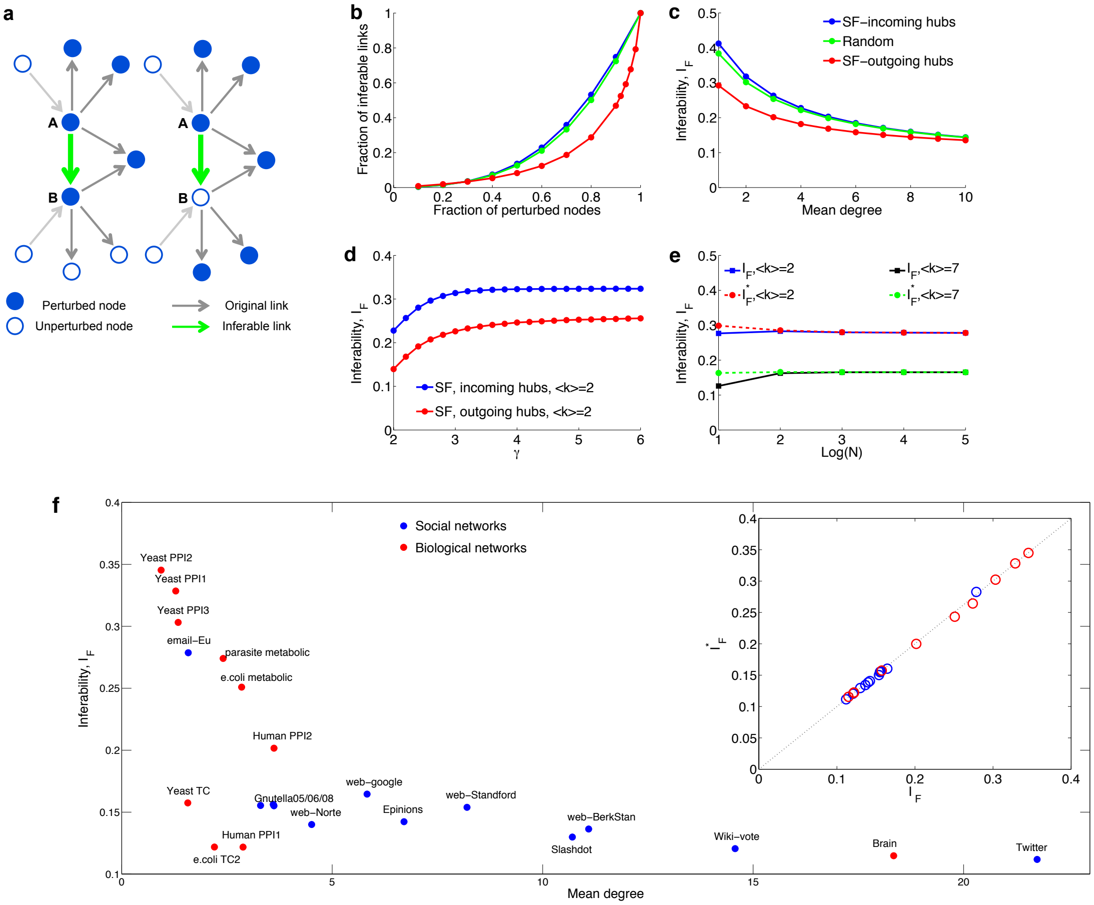
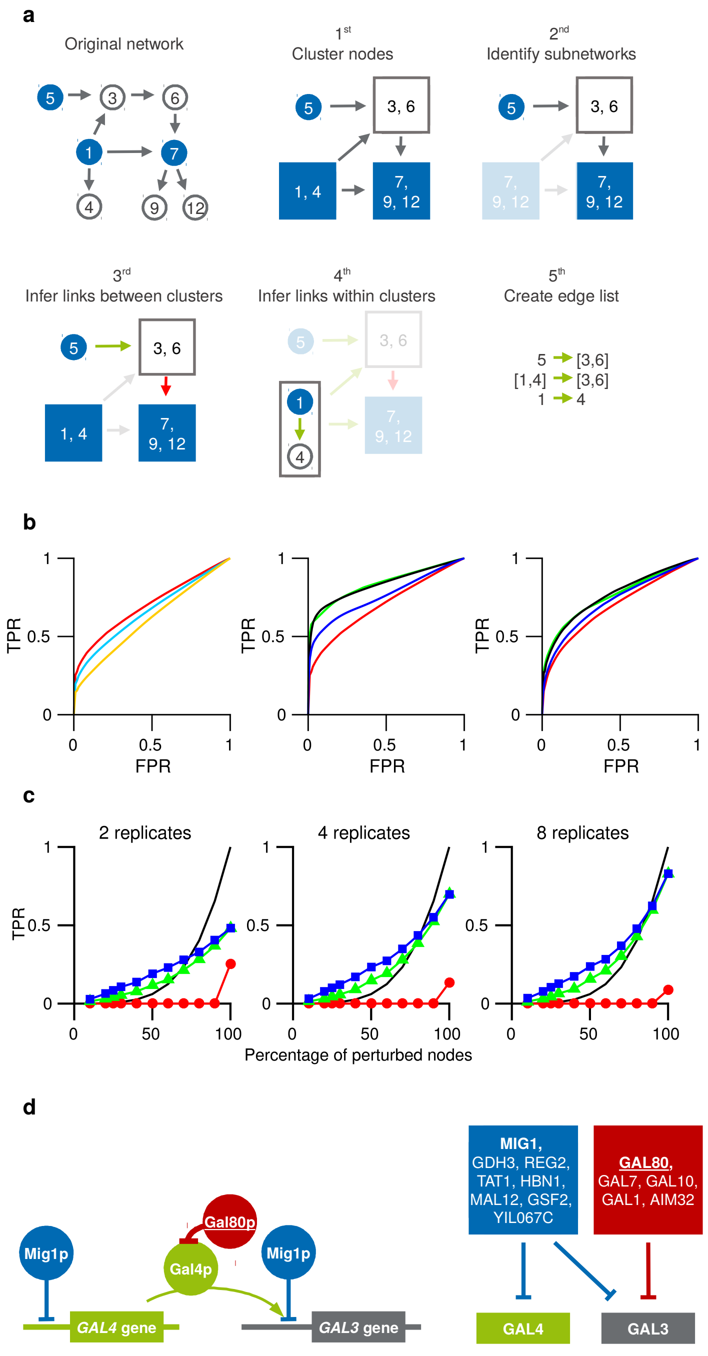


Figure 1 — Inferability as a function of network parameters. **(a)** Directed links are inferable if either all outgoing links of the source node are perturbed including the target node (left panel) or if all outgoing links of source node and target node are perturbed, with the target node not perturbed. **(b)** Fraction of inferable links against the fraction of perturbed nodes using three network types: (i) scale free network with exponent $\gamma = 2.5$ and mean degree $\langle k \rangle = 3$, where nodes of higher degree target nodes of lower degree (outgoing hubs), (ii) same network as (i) but with all link directions inverted (incoming hubs), or a network generated by random insertion of links with the same mean

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

degree as scale-free networks (random network). Colour coding as in (c). **(c)** Network inferability versus mean degree, using networks of (b). **(d)** Network inferability versus scaling exponent for two types of scale-free networks. **(e)** Asymptotic invariance of the two inferability measures introduced in the main text with respect to network size. **(d)** Network inferability as a function of mean degree for social and biological networks. Correlation between the two inferability measures introduced in the main text (inset).

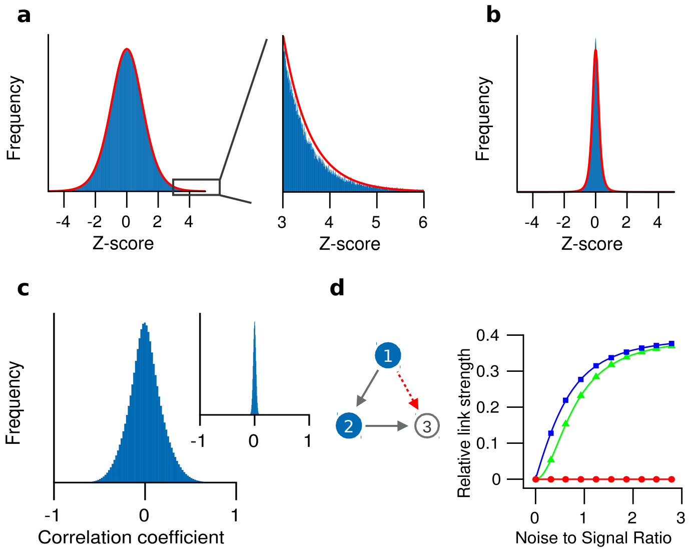
bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2 — Performance of our method. **a)** Flow-chart showing the algorithmic steps for network inference as explained in the main text. **b)** Receiver Operating Characteristic (ROC) curves for 300-node scale-free networks with additive Gaussian measurement noise of 10% of the expression level and 25% of the nodes perturbed. Data was generated using GeneNetWeaver (left and middle panel) as well as using scale-free network structure with mean degree of $\langle k \rangle = 2$ and scaling exponent $\gamma = 2.5$ (right panel, Supplementary Notes 3). Here, the true positive rate is computed with respect to the inferable links [30]. Performance of inference methods without sparsity constraints (left panel): PRC (red), partial correlations / linear regression (turquoise) and conditional mutual information (orange). Performance of inference methods with sparsity constraints (middle and right panel): PRC with subnetwork method (green) and Lasso (black) both applied to a subset of significantly responding nodes that were selected with 1% false discovery rate, Lasso regression applied to all 300 nodes (blue), and PRC from left panel (red) for comparison. **c)** True positives for the same scale-free network of (b) with 2, 4 and 8 experimental replicates with 5% false discovery rate for both significantly responding nodes and link strength: PRC (red), PRC with subnetwork method (green), PRC with subnetwork and clustering method (blue), and $F(q)$ (black line). **d)** The *S. cerevisiae* GAL network as an example for a gene regulatory network where phosphorylated Mig1 sets the basal expression levels of Gal4 and one of its many regulatory targets Gal3. Gal4 protein can activate Gal3 but is inactivated upon binding of Gal80 protein. The transcriptome dataset contains knockout mutants for Gal80 and Mig1 but not for the remaining Gal genes. A schematic representation of the key molecular mechanisms (left) and links inferred from transcriptome data [3] (right).

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 1 — Distribution of wild type expression levels for *S. cerevisiae* from 748 biological replicates. **(a)** Distribution of the relative expression, $log_2(r_i)$, with $r_i := x_i/x_i^{pool}$ and x_i the expression of gene i relative to x_i^{pool} followed by standardization of the $\log_2$ fold changes (z-score). The values x_i^{pool} have been obtained by first pooling the 748 biological replicates before measuring gene expression. The distribution is well described within 5 standard deviations by a t-distribution with 11 degrees of freedom (red line). **b)** Distribution as in (a) but now for differences among technical replicates. **c)** Correlation between gene expression levels is significantly higher than expected by chance (inset) **d)** Illustration of a noise induced false positive link (red arrow) as described in the main text. Data for a three-node network with two links was generated by

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

applying independent perturbations on node 1 and node 2. The link strength of the non-existing link from node 1 to node 3 relative to the existing link from node 2 to node 3 was inferred using three different methods (i) partial correlations (blue squares), (ii) conditional mutual information (green triangles), and PRC (red circles).

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Partial response coefficients (PRC). We aim at inferring direct interactions between N observable molecular components, such as transcripts or proteins, by measuring their copy numbers or concentrations, $\boldsymbol{y} \in \mathbb{R}^N$. We assume that the available dataset has been generated from P perturbation experiments, $\{\boldsymbol{y}_k\}_{k=1}^P$, which may also include experimental replicates. We further assume that the molecular targets of the perturbations are known, as it is the case for gene knockout or knockdown experiments. The elements of the interaction matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ define the strength of the directed interactions among the molecular components, for example A_{ij} quantifies the direct impact of component j on component i. Given the available experimental data, our aim is to correctly classify the off-diagonal elements of $\boldsymbol{A}$ as zero or nonzero to obtain the structural organization of the interaction network. We assume that the observed component abundance, $\boldsymbol{y}_k^{obs}$, differs from the true value, $\boldsymbol{y}_k$, by additive measurement noise, $\boldsymbol{\epsilon}(t)$, which is characterized by zero mean, $\mathbb{E}[\boldsymbol{\epsilon}(t)] = 0$, and variance, $\mathbb{E}[\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}(t)^T] = \sigma^2 \boldsymbol{I}_N$, with $\boldsymbol{I}_N$ the N dimensional identity matrix. We assume that the observed data can by described to sufficient accuracy by a linear stochastic process

$$\begin{aligned} \dot{\boldsymbol{y}}(t) &= \boldsymbol{A}\left(\boldsymbol{y}(t) - \boldsymbol{y}^{ref}\right) + \boldsymbol{B}\boldsymbol{u} + \boldsymbol{\eta}(t) \\ \boldsymbol{y}^{obs}(t) &= \boldsymbol{y}(t) + \boldsymbol{\epsilon}(t) \quad , \end{aligned} \tag{1}$$

with $\boldsymbol{A}$ negative definite to ensure stability. Equations of this type typically arise from linear expansion of a nonlinear model around a reference state, $\boldsymbol{y}^{ref}$. The perturbation vector $\boldsymbol{u}$ is time-independent and reflects perturbations that persist long enough to propagate through the network, such as mutations that affect gene activity. We further introduce with $\boldsymbol{B} \in \mathbb{R}^{N \times N}$, the associated standard deviations of the perturbative forces, $\boldsymbol{u}$, and assume that these forces are sampled from

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a standard normal distribution, with mean $\mathbb{E}[\boldsymbol{u}] = 0$ and covariance matrix $\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^T] = \boldsymbol{I}_N$. In general, only the positions of the non-zero elements of $\boldsymbol{B}$ are known from the experimental setup but their actual values are unknown. Fluctuating perturbations that are fast in comparison to the network response are represented by the stochastic vector $\boldsymbol{\eta}(t)$, which we model as white noise defined by $\mathbb{E}[\boldsymbol{\eta}(t)] = 0$ and $\mathbb{E}[\boldsymbol{\eta}(t)\boldsymbol{\eta}(t')] = \gamma\delta(t-t')\boldsymbol{I}_N$, with $\delta(t)$ the delta function and γ the fluctuation strength. The solution of Eq. (1) is given by

$$\boldsymbol{y}^{obs}(t) \quad = \quad \boldsymbol{y}^{ref} + \int_{t_0}^{t} e^{\boldsymbol{A}(t-t')}[\boldsymbol{B}\boldsymbol{u} + \boldsymbol{\eta}(t)]dt' + \boldsymbol{\epsilon}(t) \tag{2}$$

from which we obtain in the stationary regime, $t \gg t_0$, a relation between $\boldsymbol{A}$ and the covariance matrix of observed component abundances

$$\begin{aligned} \boldsymbol{C} \quad &:= \quad \mathbb{E}\left[(\boldsymbol{y}^{obs} - \boldsymbol{y}^{ref})(\boldsymbol{y}^{obs} - \boldsymbol{y}^{ref})^T\right] \\ &= \quad \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{A}^{-T} - \gamma\left(\boldsymbol{A} + \boldsymbol{A}^T\right)^{-1} + \sigma^2\boldsymbol{I}_N \end{aligned} \tag{3}$$

We exploit Eq. (3) to infer directed networks from correlation data. Here, we assume that component abundances are obtained from averaging over a large number of cells in the stationary growth phase. In this case, fast fluctuating perturbations that arise from thermal noise and can be observed only on single-cell level average out. To infer the interaction matrix, $\boldsymbol{A}$, we start with singular value decomposition of the matrix product $\boldsymbol{A}^{-1}\boldsymbol{B}$

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T := \boldsymbol{A}^{-1}\boldsymbol{B} \qquad \Rightarrow \qquad \boldsymbol{B} = \boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{4}$$

with $\boldsymbol{U}$ and $\boldsymbol{V}$ orthogonal matrices and $\boldsymbol{\Sigma}$ a diagonal matrix containing the singular values. The negative definite matrix $\boldsymbol{A}$ is invertible and hence has full rank. In the following, we show that it is possible to infer the directed link strength between a sender node j and a receiver node i, if

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

all direct perturbations on receiver node i are removed from the dataset and if a significant partial correlation between i and j exists. Removing the perturbation data for node i implies that the matrix $\boldsymbol{B}$ has at least one zero entry. As a consequence, $N_0 \geq 1$ singular values are zero – as in general not all nodes are perturbed – and the corresponding rows of $\boldsymbol{U}$ span the left nullspace of $\boldsymbol{A}^{-1}\boldsymbol{B}$. In the absence of fast fluctuating perturbations, $\gamma = 0$, we can rewrite the covariance matrix as

$$\begin{aligned} \boldsymbol{C} &= \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{A}^{-T} + \sigma^2\boldsymbol{I}_N && (5) \\ &= \boldsymbol{U}(\boldsymbol{\Sigma}^2 + \sigma^2\boldsymbol{I}_N)\boldsymbol{U}^T \quad . && (6) \end{aligned}$$

Assuming that the observed node activities follow a multivariate normal distribution, we can find estimates for the unknown orthogonal matrix $\boldsymbol{U}$, the singular values $\boldsymbol{\Sigma}$, and the observational noise σ by maximizing the log-likelihood function $\mathcal{L}$ under the constraint $\boldsymbol{U}_k^T\boldsymbol{U}_k = \boldsymbol{I}_N$, with

$$\begin{aligned} \mathcal{L} &:= \ln\prod_{n=1}^{P}\mathcal{N}(\boldsymbol{y}_n^{obs}|\boldsymbol{y}^{ref}, \boldsymbol{C}) + \sum_{k=1}^{N}\lambda_k(\boldsymbol{U}_k^T\boldsymbol{U}_k - 1) && (7) \\ &= -\frac{P}{2}\Big\{M\ln 2\pi + \ln|\boldsymbol{C}| + tr(\boldsymbol{C}^{-1}\boldsymbol{S})\Big\} + \sum_{k=1}^{N}\lambda_k(\boldsymbol{U}_k^T\boldsymbol{U}_k - 1) && (8) \end{aligned}$$

Here, $\boldsymbol{S} := \frac{1}{P}\sum_{n=1}^{P}(\boldsymbol{y}_n^{obs} - \boldsymbol{y}^{ref})(\boldsymbol{y}_n^{obs} - \boldsymbol{y}^{ref})^T$ and $\boldsymbol{y}^{ref} := \frac{1}{P}\sum_{n=1}^{P}\boldsymbol{y}_n^{obs}$ denote maximum likelihood estimates of the covariance matrix [31] and the expected node activities, λ_k is a Lagrange multiplier and $tr(.)$ denotes the trace of a matrix. In the following calculations, we substitute $\boldsymbol{S}$ by the unbiased sample covariance matrix, $\boldsymbol{S} \rightarrow P(P-1)^{-1}\boldsymbol{S}$. Note that $\boldsymbol{V}$ must disappear in the likelihood function as the covariance matrix of $\boldsymbol{u}$ is invariant under any orthogonal transformation $\boldsymbol{u} \rightarrow \boldsymbol{V}^T\boldsymbol{u}$.

The maximum of the log-likelihood function is determined by the conditions $d\mathcal{L}/d\boldsymbol{U}_k = 0$,

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$d\mathcal{L}/d\boldsymbol{\Sigma}_{kk} = 0$, and $d\mathcal{L}/d\sigma^2 = 0$, which results in

$$\boldsymbol{S}\boldsymbol{U}_k = \Lambda_k \boldsymbol{U}_k \quad \text{with} \quad \Lambda_1 \leq \Lambda_2 \leq ... \leq \Lambda_N \tag{9}$$

$$\sigma^2 = \frac{1}{N_0} \sum_{k=1}^{N_0} \Lambda_k \tag{10}$$

$$\Sigma_{kk} = \begin{cases} \sqrt{\Lambda_k - \sigma^2} & \text{for} \quad k > N_0 \\ 0 & \text{for} \quad k \leq N_0 \end{cases} \tag{11}$$

showing that maximum likelihood estimates of $\boldsymbol{U}$, σ^2, and $\boldsymbol{\Sigma}$ are uniquely determined by the sample covariance matrix $\boldsymbol{S}$. Our analysis is mathematically equivalent to the probabilistic interpretation of principle component analysis [23].

Solving the matrix equation, Eq. (4), for $\boldsymbol{A}$ gives

$$\boldsymbol{A} = (\boldsymbol{B}\boldsymbol{V}\boldsymbol{\Sigma}^+ + \boldsymbol{W}\boldsymbol{\Sigma}^0)\boldsymbol{U}^T \tag{12}$$

with $\boldsymbol{\Sigma}^+$ the pseudoinverse of $\boldsymbol{\Sigma}$. As the matrix $\boldsymbol{A}$ has full rank, we complement $\boldsymbol{\Sigma}^+$ with an unknown diagonal matrix $\boldsymbol{\Sigma}^0$ that has nonzero values where $\boldsymbol{\Sigma}^+$ has zero values and vice versa and complement $\boldsymbol{B}\boldsymbol{V}$ with an unknown orthogonal matrix $\boldsymbol{W}$. Note that by construction, $\boldsymbol{\Sigma}^+\boldsymbol{U}^T$ and $\boldsymbol{\Sigma}^+\boldsymbol{U}^T$ map from complementary subspaces and thereby ensure that $\boldsymbol{A}$ has full rank. The fact that $\boldsymbol{V}$, $\boldsymbol{W}$, and $\boldsymbol{\Sigma}^0$ cannot be determined from $\boldsymbol{S}$ shows that $\boldsymbol{A}$ cannot be computed from a single covariance matrix. A more general case arises when measurement noise is independent but not isotropic, $\sigma^2\boldsymbol{I} \to \sigma^2\boldsymbol{D}$, with $\boldsymbol{D} = diag(r_1, r_2, ..., r_N)$ a diagonal matrix with known positive elements that contains scaled noise variances, $r_i := \sigma_i^2/\sigma^2$, resulting in

$$\boldsymbol{C} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{A}^{-T} + \sigma^2\boldsymbol{D} \tag{13}$$

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A transformation to isotropic noise is possible by multiplying both sides of Eq. (13) by $\boldsymbol{D}^{-\frac{1}{2}}$, which changes the result Eq. (12) to

$$\boldsymbol{A} = (\boldsymbol{B}\boldsymbol{V}\boldsymbol{\Sigma}^{+} + \boldsymbol{W}\boldsymbol{\Sigma}^{0})\boldsymbol{U}^{T}\boldsymbol{D}^{-\frac{1}{2}} \tag{14}$$

with $\boldsymbol{U}$ the eigenvectors of $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$.

Case $\boldsymbol{N_0 = 1}$. We assume that the i-th node is the only unperturbed node in the network and consequently define $B_{il} = 0$ for all l. From Eq. (12) we obtain a unique solution for the i-th row of $\boldsymbol{A}$ relative to the diagonal element, A_{ii},

$$\frac{A_{ij}}{A_{ii}} = \frac{\sum_{k,l=1}^{N} B_{il}V_{lk}\Sigma_{kk} + \sum_{k=1}^{N} W_{ik}\Sigma_{kk}^{0}U_{kj}^{T}}{\sum_{k,l=1}^{N} B_{il}V_{lk}\Sigma_{kk} + \sum_{k=1}^{N} W_{ik}\Sigma_{kk}^{0}U_{ki}^{T}} = \left.\frac{U_{kj}^{T}}{U_{ki}^{T}}\right|_{k=1} = \frac{U_{j1}}{U_{i1}} \tag{15}$$

with U_{j1} the j-th element of the eigenvector of $\boldsymbol{S}$ that has the smallest eigenvalue. Note that the first term in the brackets vanishes as $B_{il} = 0$ and Σ_{11}^{0} is the only nonzero element of $\boldsymbol{\Sigma}^{0}$. The important point is that any dependency on σ has dropped out, which makes this method asymptotically unbiased with respect to measurement noise. The fact that we can determine the elements of the i-th row of $\boldsymbol{A}$ only relative to a reference value, A_{ii}, is rooted in fact that we have to determine the N parameters $\{A_{i1}, .., A_{ii}, .., A_{iN}\}$ from $N-1$ perturbations. As a consequence, the strengths of the links onto the target nodes cannot be compared directly if their restoring forces or degradation rates, A_{ii}, are different. Generally, only relative values of $\boldsymbol{A}$ can be determined, as the average perturbation strength on node i cannot be disentangled from its restoring force A_{ii} – a problem that is typically circumvented by defining $A_{ii} := 1$ for all i [7,12,14]. For the case that all nodes in the network are perturbed one-by-one, we can cycle through the network and remove the perturbations that act on the current receiver node, while keeping the perturbations on the remain-

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ing nodes. By computing the N corresponding covariance matrices and their eigenvectors, we can infer the complete network structure from Eq. (15) if the data quality is sufficiently high.

Case $N_0 > 1$. For all i that are not perturbed or whose perturbation data has to be removed from the dataset we get from Eq. (12)

$$\frac{A_{ij}}{A_{ii}} = \frac{\sum_{k=1}^{N_0} W_{ik} \Sigma^0_{kk} U^T_{kj}}{\sum_{k=1}^{N_0} W_{ik} \Sigma^0_{kk} U^T_{ki}} \tag{16}$$

Non-unique solutions of Eq. (16) can arise if a given fraction of the variance of the receiver node i can be explained by more than one sender node, for example, when a perturbed node j targets two nodes with index i and l. In this case it is unclear from the node activity data whether i is affected directly by j or indirectly through l, or by a combination of both routes. If the node l is unperturbed, we can use the simple criteria shown in Fig. 1a to decide whether the link from j to i is inferable or not. If node l is weakly perturbed, a statistical criteria is needed to decide about inferability, which can be computed numerically as follows: To find out whether j transmits a significant amount of information to i that is not in l, we first remove the perturbed node j from the network and determine the link strengths $\boldsymbol{A}'$ for the remaining network of size $N-1$. To construct a possible realisation of $\boldsymbol{A}'$ we set in Eq. (16) the nonzero values of $\boldsymbol{\Sigma}^0$ to unity and use $\boldsymbol{W} = \boldsymbol{U}$ to arrive at the expression

$$\frac{A'_{il}}{A'_{ii}} = \frac{\sum_{k=1}^{N_0} U'_{ik} U'_{lk}}{\sum_{k=1}^{N_0} U'_{ik} U'_{ik}} \tag{17}$$

with $\boldsymbol{U}'$ determined from the sample covariance matrix with the i-th column and i-th row removed. Using the inferred link strength from Eq. (17) we can rewrite Eq. (3) as a two-node residual inference problem between j and i, where we obtain a lower bound for link strength from node j to i

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

by using the variation of i that could not be explained by $\boldsymbol{A}'$. Defining by $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{B}}$ and $\tilde{\boldsymbol{D}}$ the 2×2 analogs to the full problem we obtain

$$\tilde{\boldsymbol{C}} = \tilde{\boldsymbol{A}}^{-1} \boldsymbol{B}\boldsymbol{B}^T \tilde{\boldsymbol{A}}^{-1} + \sigma^2 \tilde{\boldsymbol{D}} \tag{18}$$

with $\tilde{\boldsymbol{C}}$ the covariance matrix of the vector $\tilde{\boldsymbol{y}}^{obs} = (y_j^{obs}, y_i^{obs} + \sum_{l \neq \{i,j\}} A'_{il} (A'_{ii})^{-1} y_l^{obs})^T$ and $\tilde{D}_{11} = r_j$, $\tilde{D}_{22} = r_i + \sum_{l \neq \{i,j\}} A'^2_{il} A'^{-2}_{ii} r_l$, using the scaled variances $r_i = \sigma_i^2 / \sigma^2$. Note that $A_{ii} < 0$ for all i as these elements represent sufficiently strong restoring forces that ensure negative definiteness of $\boldsymbol{A}$ and that we have $0 = A'_{ii} y_i^{obs} + \sum_{l \neq i} A'_{il} y_l^{obs}$ from Eq. (1) in the stationary case. An estimate for the minimum relative link strength from node j to node i can be calculated from Eq. (14) and is given by

$$\frac{\tilde{A}_{12}}{\tilde{A}_{11}} = \frac{\tilde{U}_{21} \tilde{D}_{22}^{-1/2}}{\tilde{U}_{11} \tilde{D}_{11}^{-1/2}} \tag{19}$$

Eq. (19) can be considered as an asymptotically unbiased response coefficient between node 1 as target node and node 2 as source node, as any dependency on σ^2 has dropped out. An estimate for the maximum relative link strength from node j to node i follows from Eq. (19) with the off-diagonal elements of $\boldsymbol{A}'$ set to zero. We classify a link as non-inferable if there exists (i) a significant difference between the minimum und maximum estimated link strength and (ii) a minimum link strength that is not significantly different from noise.

Computational complexity of PRC. The computational cost for computing partial response coefficients scales as $\mathcal{O}(N_{sub}^3)$, with N_{sub} the size of the subnetwork under consideration. However, as we infer directed networks, we first have to remove the perturbations on each target node before

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

its incoming links can be inferred. The cycling through up to $N_{sub} - 1$ perturbed target nodes increases the computational complexity to $\mathcal{O}(N_{sub}^4)$ in the worst case. As we have generated a subnetwork for each significantly varying node and used residual bootstrapping to infer statistically significant links, the total computational complexity is given by $\mathcal{O}(N_{boot} N_{sig} \langle N_{sub}^4 \rangle)$, where $\langle . \rangle$ denotes averaging over all subnetworks and N_{boot} the number of bootstrap samples. If the travelling distance of perturbations (correlation length) in the network is significantly shorter than the network diameter, such that $N_{sub}/N \to 0$ in the limit of large networks, $N \to \infty$, the computational complexity scales linear with network size. In contrast, using Lasso to infer directed links requires $\mathcal{O}(N_{boot} N_{sig}^4)$ operations, as the more efficient Graphical Lasso method [32] is only applicable to undirected networks. Whether our method is computationally more efficient than Lasso depends on the inference problem. However, for the scale free networks investigated in this work our method required significantly less computational time than inference via Lasso using parallel computing.

Fraction of non-inferable links. Inferability of a directed link between source and target node requires that the remaining network may not contain the same information that is transmitted between them. A sufficient condition is that all information that the remaining network receives from the source node is destroyed by sufficiently strong perturbations. If the target node is not perturbed, information from the source node my reach the remaining network through the target node. In this case also the targets of the target node must be perturbed (Fig. 1a). Counting network motifs that satisfy these conditions gives the number of inferable links. If the network size, N, is significantly larger than the number of outgoing links for both the source and target nodes, we can approximate

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the fraction of inferable links, $F(q)$, by the expression (Supplementary Note 1)

$$F(q) \approx \sum_{k=1}\sum_{l=0}\sum_{m=0}^{\min(k-1,l)} \left[q^{k+1} + (1-q)\, q^{z}\right] P(k,l,m|k \to l)$$

Here, $P(k,l,m|k \to l)$ is the conditional probability of finding two connected nodes in the directed network, where the source node has $k \geq 1$ outgoing links, the target node has $l \geq 0$ outgoing links, and both share m nodes as common targets of their outgoing links. The first term in the brackets corresponds to the case that independent perturbation data for node B exists (Fig. 1a, left panel) and the second term to the case where independent perturbation data for node B is absent (Fig. 1a, right panel). In the calculation of $F(q)$ we assumed that the links in the network are identical with respect to the information they can transmit.

Data preparation Kemmeren et al.[3] provided a transcriptome data set of a Saccharomyces cerevisiae genome-wide knockout library (with mutant strains isogenic to S288c). This data set comprises transcript levels of 6170 genes for 1484 deletion mutants. The data is presented as the logarithm of the fluorescence intensity ratios (M-values) of transcripts relative to their average abundance across a large number of wild type replicates, resulting in logarithmic fold changes of mutant/wild type gene expression levels compared to a wild type reference level. Kemmeren et al. also used a dye swap setup for several experiments to average out the effect of a possible dye bias. Their chip design measures most of the genes twice per biological sample, thus allowing to estimate the technical variance. The preprocessing of the data is described in Kemmeren et al.[3], Supplementary Information.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Residual Bootstrapping We make use of the 748 measured wild type experimental replicates to determine the natural variation among biological replicates, $\delta_{in} := log_2(r_{in}) - \langle log_2(r_{in})\rangle_n$, with $r_{in} := x_{in}/x_i^{pool}$, x_{in} the expression of gene i in wild type replicate n, x_i^{pool} the expression level of gene i measured after pooling over wild type replicates, and $\langle . \rangle_n$ denoting the average over replicates. To generate the bootstrap samples we randomly select 200 different δ_{in} from the replicates for each gene i, and add these values to the log fold-changes of the perturbed expression levels, $\langle log_2(r_{in}^{pert})\rangle_n$, with $r_{in}^{pert} := x_{in}^{pert}/x_i^{pool}$ and the average is taken over the two replicates for each knockout.

Sparsity constraints by removing noisy nodes As network inference typically comes with an insufficient amount of independent perturbations and experimental replicates we run into the problem of overfitting the data. In this case, noisy information from many network nodes is collected to explain the response of a given target node. L^1-norm regularized regression (Lasso) systematically removes many links, where each link explains only a small part of the variation of the target node, in favour of few links, where each link contributes significantly. In our approach we remove noisy nodes and thus their potential outgoing links, where the critical noise level is determined by independent biological noise. Which nodes are removed depends on the *source node* that is perturbed. In the absence of noise, our algorithm removes weakly responding nodes from the network. We thereby assume that the existence of many indirect interactions between source and target node by first distributing the signal of the source node among many weakly responding nodes and then collecting these week signals to generate a significantly responding target node is much less likely than the existence of a single direct interaction. However, in the noise-free case we run into the

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

same problem as Lasso to determine the right cut-off (regularization parameter).

Synthetic Data Synthetic data was generated using our own model and GeneNetWeaver [24] – an open access software that has been designed for benchmarking network inference methods. With GeneNetWeaver, networks were generated from a model that closely resembles the structure of the yeast regulatory network [24], and steady state levels of node activities were computed using ordinary differential equations (ODEs). In our model, we first generated scale-free networks with an exponent of 2.5 and an average degree of 2. Then, we solved a system of ordinary differential equations with non-linear regulatory interactions between nodes to obtain steady state values of node activities, e.g. transcript levels. For both models, logarithmic fold changes of node activities were calculated (transcriptional levels upon perturbation relative to wild levels), and gaussian measurement noise was added.

1. Smanski, M. J. *et al.* Synthetic biology to access and expand nature's chemical diversity. *Nature Rev. Microbiol.* **14**, 135–149 (2016).

2. Esvelt, K. M. & Wang, H. H. Genome-scale engineering for systems and synthetic biology. *Mol. Syst. Biol.* **9**, 641–641 (2014).

3. Kemmeren, P. *et al.* Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. *Cell* **157**, 740–752 (2014).

4. Costanzo, M. *et al.* The Genetic Landscape of a Cell. *Science* **327**, 425–431 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).

6. Meinshausen, N. *et al.* Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci.* **113**, 7361–7368 (2016).

7. Kholodenko, B. N. *et al.* Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci.* **99**, 12841–12846 (2002).

8. Newman, M. E. J. *Networks* (Oxford University Press, 2010).

9. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601–1–1257601–7 (2015).

10. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nature Methods* **9**, 796–804 (2012).

11. Feizi, S., Marbach, D., Médard, M. & Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature Biotechnol.* **31**, 726–733 (2013).

12. Barzel, B. & Barabsi, A.-L. Network link prediction by global silencing of indirect correlations. *Nature Biotechnol.* **31**, 720–725 (2013).

13. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci.* **106**, 67–72 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

14. Bastiaens, P. *et al.* Silence on the relevant literature and errors in implementation. *Nature Biotechnol.* **33**, 336–339 (2015).

15. Bansal, M., Belcastro, V., Ambesi-Imiombato, A. & di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **3**, 1–10 (2007).

16. Hill, S. M. *et al.* Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods* **13**, 310–318 (2016).

17. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comp. Biol.* **7**, 601–620 (2000).

18. Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **37**, 424 (1969).

19. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **85**, 461–464 (2000).

20. Oates, C. & Mukherjee, S. Network Inference and Biological Dynamics. *Ann. Appl. Stat.* **6**, 1209–1235 (2012).

21. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).

22. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, 1381–1394 (2016).

23. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *J. R. Stat. Soc.* **61**, 611–622 (1999).

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

24. Schaffter, T., Marbach, D. & Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* **27**, 2263–2270 (2011).

25. Omranian, N., Eloundou-Mbebi, J. M. O., Mueller-Roeber, B. & Nikoloski, Z. Gene regulatory network inference using fused LASSO on multiple data sets. *Sci. Rep.* **6**, 1–14 (2016).

26. Alipanahi, B. & Frey, B. J. Network Cleanup. *Nature Biotechnol.* **31**, 714–715 (2013).

27. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–19 (2010).

28. Al-Momani, M. Shrinkage and Penalty Estimation Strategies in Some Spatial Models. *Electronic Theses and Dissertations* 1–239 (2013).

29. Stockwell, S. R., Landry, C. R. & Rifkin, S. A. The yeast galactose network as a quantitative model for cellular memory. *Mol. BioSyst.* **11**, 28–37 (2015).

30. Siegenthaler, C. & Gunawan, R. Assessment of Network Inference Methods: How to Cope with an Underdetermined Problem. *PLoS ONE* **9**, e90481 (2014).

31. Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Molec. Biol.* **4**, 1–32 (2005).

32. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements We acknowledge help from Patrick Kemmeren in interpreting the yeast transcriptome data, the High Performance Computing Plattform of the Heinrich-Heine University, and DFG funding by SPP 1395 and the Cluster of Excellence in Plant Sciences (grant no. EXC 1028).

Author contributions statements A.S.K. and M.K. conceived the PRC method. N.H. and M.K. developed the counting procedure for inferable links and the Inferability measure. C.F.B. and M.K. developed the clustering and subnetwork methods and analysed the transcriptome data. N.H. wrote Supplementary Note 1. M.K. wrote Supplementary Note 2. C.F.B. wrote Supplementary Note 3.

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

Correspondence Correspondence and requests for materials should be addressed to M.K. (email: markus.kollmann@hhu.de).

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Note 1

1 Counting the number of inferable links with respect to the number of perturbed nodes

Structural features of a network allows us to determine the expected fraction of inferable links analytically if the number of randomly perturbed nodes (N_p) is known and measurement noise is absent.
Assume a subnetwork where a source node (A) targets a node (B), given that the out-degree of node A is k, the out-degree of node B is l, and A and B have c nodes as common targets. We aim to infer the link from node A to node B (See figure 1).

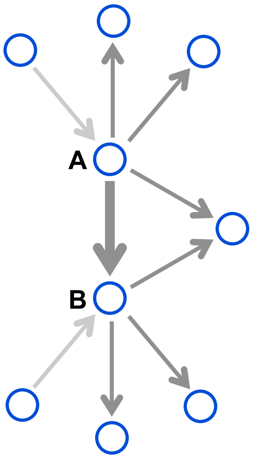


Fig. 1: Example of a directed subnetwork. We would like to infer the link from node A to node B (shown in bold), where A has out-degree 4, B has out-degree 3, and A and B share 1 common target.

We denote a direct link from source node (A) to target node (B) as inferable if there exits a detectable amount of mutual information between A and B that

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cannot be transmitted by any alternative route through the network. This requires that at least one nodes of each alternative route is perturbed and thereby part of the transmitted information is destroyed. To meet this requirement, either one of the following conditions must be fulfilled (Fig. 1a, main text):

1. All nodes that are targeted by A are perturbed, including node B.
2. Except B, all nodes that are targeted by A and B are perturbed.

By collecting all subnetworks that fulfil conditions 1 or 2 we can calculate the average fraction of inferable links, $F(N_p)$, using $N' := N-k-1$ and $N'_p := N_p-k-1$

$$F(N_p) = \sum_k \frac{\begin{pmatrix} N' \\ N'_p \end{pmatrix}}{\begin{pmatrix} N \\ N_p \end{pmatrix}} P(k|A \to B) + \sum_{k,l,m} \frac{\begin{pmatrix} N' - l + c \\ N'_p - l + c + 1 \end{pmatrix}}{\begin{pmatrix} N \\ N_p \end{pmatrix}} P(k,l,c|A \to B) \quad (1)$$

Here, N is network size, N_p is number of perturbed nodes, k is the out-degree of the source node (A), l is the out-degree of the target node (B), c is the number of common nodes targeted by A and B. We further defined by $P(k|A \to B)$ the conditional probability that for any two connected nodes, source node (A) has out-degree k and $P(k,l,c|A \to B)$ is conditional probability that for any two connected nodes, source node (A) has out-degree k , target node (B) has out-degree l, and A and B target c common nodes. The first term in the numerator counts motifs that fulfil condition 1, and the second term in the numerator counts motifs that fulfil condition 2. The term in the denominator counts all possible network motifs, when N_p nodes of the network are perturbed. Note that measurement noise can decrease the fraction of inferable links due to distortion of information flow. Thus equation 1 calculates the expected maximum fraction of inferable links.

Assuming that the network size, N, and the number of perturbed nodes, N_p, are much larger than the out-degrees, $N, N_p \gg k, l$, we can apply Stirling's approximation and simplify Eq. (1)

$$F(q) \approx \sum_{k,l,c} \left[q^{k+1} + (1-q)q^{k+l-c}\right] P(k,l,c|A \to B) \quad (2)$$

where q denotes fraction of perturbed nodes, $q = N_p/N$.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2 Network's Inferability

According to Eq. (2), networks with different structural features have different $F(q)$ curves (Fig. 1b, main text). We therefore define a inferability measure, I_F, as the area under the curve of F(q) that reflects how difficult it is to infer links for a given network structure

$$\begin{aligned} I_F &= \int_0^1 F(q)dq \\ &= \sum_{k,l,c} \left[\frac{1}{k+2} + \frac{1}{k+l-c+1} - \frac{1}{k+l-c+2} \right] P(k,l,c|A \to B) \end{aligned} \qquad (3)$$

Not that I_F is independent of network size. For a sufficient large networks $N \gg 1$ and when feed forward loops rare, we can approximate the joint probability by

$$\begin{aligned} I_F &\approx I_F^* \\ &:= \sum_{k,l,c} \left[\frac{1}{k+2} + \frac{1}{k+l+1} - \frac{1}{k+l+2} \right] P(k|A \to B)P(l|A \to B) \end{aligned} \qquad (4)$$

According to Eqs. (3) and (4), inferability mainly depends on the out-degree of nodes. Consequently, networks consisting of nodes with high out-degree have low inferability (I_F) and are the most difficult ones to infer.

3 Comparison of $F(N_p)$ with the inference algorithm

We compared the fraction of inferable links determined by the analytical formula, $F(N_p)$, and the average number of inferable links classified by our inference algorithm in the limit of low measurement noise. In order to calculate the average number of inferable links form the inference algorithm, first we consider a specific network type. For each number of perturbed nodes, N_p, we first randomly identify nodes that will be perturbed and subsequently generate node activity data. For each random sample of N_p nodes, we calculate the fraction of inferable links and finally average over all configurations of perturbed nodes. Figure 2 compares the results of the analytical and numerical approaches for three different types of networks. We consider the two extreme examples of scale-free networks; a scale-free network where hubs are targets of links (Fig. 2a), a scale-free network where hubs are souces of links (Fig. 2b) and a random network (Fig. 2c). In all three network types the network sizes are $N = 100$ and averaging runs over 300 different configurations of N_p randomly selected nodes that are perturbed individually by knockouts.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

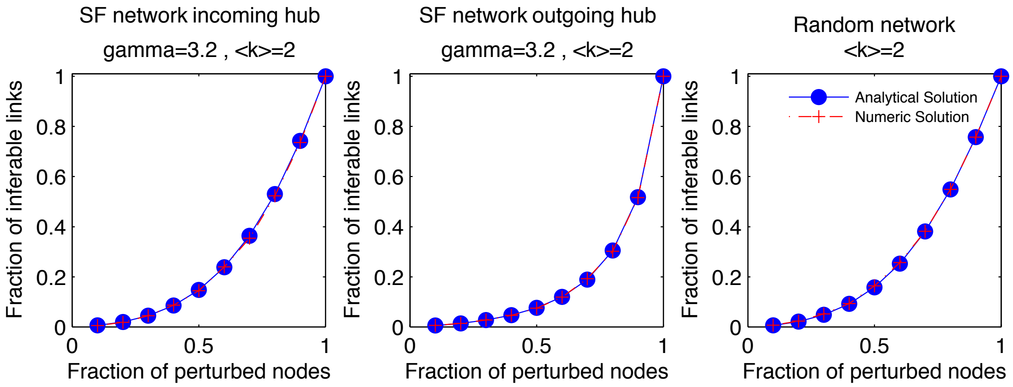


Fig. 2: Comparison of the analytical formula, Eq. (1) and the inference algorithm by calculating the number of inferable links with respect to the number of perturbed nodes, N_p.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Notes 2

Interpretation of network inference algorithms that are related to partial correlations

Algorithms based on partial correlations [1, 2] – a method designed to discriminate between direct and indirect interactions – show comparatively low performance for predicting direct correlations between transcript levels using transcriptome data [3] but excellent performance for predicting contact points of protein structures [4, 1]. To understand the difference in performance of partial correlations, we investigate Eq. (3) of the Online Methods section for the special case where (i) all perturbations are independent and fast fluctuating, $\mathbf{B} = \mathbf{0}$, (ii) all links are bidirectional (undirected network), $\boldsymbol{A} = \boldsymbol{A}^T$, and (iii) measurement noise is absent, $\sigma = 0$. In this case the elements of interaction matrix, $\boldsymbol{A}$, correspond to partial correlations for $\gamma = 2$

$$\boldsymbol{C} = \boldsymbol{A}^{-1} \qquad \Leftrightarrow \qquad \boldsymbol{A} = \boldsymbol{C}^{-1} \tag{1}$$

The constraints imposed by the conditions (i)-(iii) explain the poor performance of partial correlations in inferring undirected links for gene regulatory networks [3], where most of data is generated by perturbations that persist on long time scales, measurement noise is significant, and the network is directed. In contrast, conditions (i)-(iii) are satisfied to very good approximation for identifying physical contacts between amino acids in folded proteins [4], where single nucleotide substitutions occur on much shorter time scales than significant changes in protein structure, links are undirected, and measurement noise from sequencing errors make no relevant contribution.

In contrast to undirected networks, directed links can be inferred in absence of measurement noise using a response matrix method, if static perturbations are applied individually to all nodes in the network [5]. Given that fast fluctuating perturbations are absent, we can calculate from Eq. (2) of the Online Methods section the response of the i-th node, δy_i, with respect to a perturbation that acts exclusively on the j-th node by $\delta y_i = -A^{-1}_{ij}\boldsymbol{B}_j\boldsymbol{u}$, with $\delta y_i := y_i - y_i^{ref}$ and $\boldsymbol{B}_j$ the

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

j-th row of $\boldsymbol{B}$. Using the expression $\delta y_j = -A^{-1}_{jj}\boldsymbol{B}_j\boldsymbol{u}$ to substitute $\boldsymbol{B}_j\boldsymbol{u}$, we arrive at the relation

$$G_{ij} := \frac{\delta y_i}{\delta y_j} = \frac{A^{-1}_{ij}}{A^{-1}_{jj}} \qquad \Rightarrow \qquad \boldsymbol{G} = \boldsymbol{A}^{-1}[diag\boldsymbol{A}^{-1}]^{-1} \tag{2}$$

with $diag\boldsymbol{X}$ a matrix with elements $\{X_{11}, X_{22}, ...\}$ on the main diagonal and zero elements otherwise. Here, $\boldsymbol{G}_{ij}$ denotes the linear response coefficient of the network, describing the change in activity of node i if node j changes its value in response to a perturbation that acts exclusively on node j. After inverting both sides of Eq. (2) and setting the restoring forces (degradation rates) to unity, $A_{jj} = -1$ for all j, we arrive at the published result of Ref. [5]

$$\boldsymbol{A} = [diag\boldsymbol{A}^{-1}]^{-1}\boldsymbol{G}^{-1} = -[diag\,\boldsymbol{G}^{-1}]^{-1}\boldsymbol{G}^{-1} \quad . \tag{3}$$

In the derivation we used that by definition $diag\boldsymbol{A} = -\boldsymbol{I} = [diag\boldsymbol{A}^{-1}]^{-1}diag\,\boldsymbol{G}^{-1}$, with $\boldsymbol{I}$ the identity matrix.

The lack of a unique relation between $\boldsymbol{G}$ and $\boldsymbol{A}$ is rooted in the fact that the $N(N+1)$ parameters of $\boldsymbol{A}$ and $\boldsymbol{Bu}$ cannot be uniquely identified from the $N \times N$ measured node activities, even for the rare case that it was possible to perturb all nodes in the network one-by-one. This problem does not change if more perturbations are carried out as $\boldsymbol{Bu}$ takes different values for different perturbations. The result that direct interactions can be computed by inverting the linear response matrix, Eq. (3), has been derived several times, with either fixing the restoring force, $\boldsymbol{A} = -\boldsymbol{I}$, [5, 2] or by computing direct interactions from $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{G}^{-1}$, [1] (the latter reference defines an unusual response matrix with zero elements on the main diagonal by $\boldsymbol{G}_{obs} := \boldsymbol{G} - \boldsymbol{I}$).

Substantial confusion arrises if $\boldsymbol{G}$ is substituted by $\boldsymbol{C}$ in Eq. (3) for the reason that $\boldsymbol{C}$ can be computed using perturbative forces that act simultaneously on many nodes [2, 1], whereas computation of $\boldsymbol{G}$ requires that all nodes are perturbed individually [5]. Despite the fact that Eq. (1) and Eq. (3) look similar, the invalid operation $\boldsymbol{G} \rightarrow \boldsymbol{C}$ implies that we apply Eq. (1), which only holds for fast fluctuating perturbations, to static perturbations for which Eq. (3) has been derived. As expected, this invalid substitution gives poor performance if a substantial fraction of the dataset comes from slowly varying perturbations [3] and good performance if the conditions (i)-(iii) apply [1].

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] Feizi, S., Marbach, D., Médard, M. & Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature Biotechnol.* **31**, 726–733 (2013).

[2] Barzel, B. & Barabsi, A.-L. Network link prediction by global silencing of indirect correlations. *Nature Biotechnol.* **31**, 720–725 (2013).

[3] Bastiaens, P. *et al.* Silence on the relevant literature and errors in implementation. *Nature Biotechnol.* **33**, 336–339 (2015).

[4] Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci.* **106**, 67–72 (2009).

[5] Kholodenko, B. N. *et al.* Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci.* **99**, 12841–12846 (2002).

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Note 3

1 Input data to network inference algorithm

1.1 Saccharomyces cerevisiae genome-wide knockout library

1.1.1 Description of the data set

Kemmeren et al. [1] provided a transcriptome data set of a *Saccharomyces cerevisiae* genome-wide knockout library (with mutant strains isogenic to S288c [2]). This data set, hereafter simply referred to as the yeast deletome data, comprises of transcript levels of 6170 genes for 1484 deletion mutants (hereafter referred to as perturbation experiments). The data of 1441 of these perturbation experiments can be used for network inference as the transcript levels of the deleted genes was measured with the microarray chip design. The data is presented as the logarithm of the fluorescence intensity ratios of red and green labelled microarray targets (M-values): $M = \left(\frac{R}{G}\right)$. For the regular experimental setup, DNA from the deletion mutants was labelled red, and DNA that was pooled from several wild type batches was labelled green. The aim of pooling the wild type DNA was to average out biological fluctuations, so that this DNA pool could be used to define a reference gene expression level. This allows the interpretation of the M-values as logarithmic fold changes of mutant gene expression levels compared to a wild type reference. Kemmeren et al. also used a dye swap setup for several experiments to average out the effect of a possible dye bias. The chip design used by Kemmeren et al. measures each gene twice per biological sample, thus allowing estimation of the technical variance.

1.1.2 Preprocessing of data

We used the preprocessed data provided by Kemmeren et al. for our analyses. The preprocessing steps are described elsewhere [1]. We re-arranged the original preprocessed data layout to make it compatible with our network inference algorithm. Specifically, M-values corresponding to dye-swap experiments were multiplied by -1 and only knock-out experiments corresponding to genes that were also included in the chip design were kept in the data set.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1.1.3 Distribution of biological and technical variance (suppl. figure 1a and suppl. figure 1b)

Since there is insufficient information about the biological processes leading to variation among biological replicates (biological noise), we assume that the process is equal for all genes. By bringing the wild type data to unit variance for each gene, we show that the variation among the wild type data logarithmic fold changes is approximately Gaussian distributed (suppl. figure 1a). By taking the mean logarithmic fold change of each biological replicate (wild type data only) and then calculating the differences of the technical replicates to this mean, the distribution of the technical variance can be found (suppl. figure 1b).

1.1.4 Correlation among wild type experiment M-values (suppl. figure 1c)

The histograms shown in suppl. figure 1c were created from the upper triangular correlation matrix (without diagonal elements) of wild type experiment M-values. The inserted figure was created by breaking up correlations between genes via random shuffling of the experiments.

1.2 Simulating Gene Expression data

To estimate how well network inference works on data which is similar to the data set provided by Kemmeren et al. [1], the input data to the inference algorithms was generated as described below. In short, steady state solutions of ordinary differential equations were used to generate absolute gene expression levels, which were then turned into logarithmic fold changes, and measurement noise was added.

1.2.1 Using scale free networks and steady state solutions of ordinary differential equations to simulate absolute gene expression levels

We chose scale free network models with exponent 2.5 to simulate a comparable network structure to yeast; the average degree was set to 2. Up to a rounding error, 80 % of links were set as inhibiting. Non-linear regulatory interactions were simulated by solving ordinary differential equations. First, wild type expression levels were calculated by solving a system of equations of the form shown below. The Hill-coefficients h_{ki} were sampled from a uniform distribution between 1 and 2. The constant K_{ki} was set to 0.5. The linear degradation term λ_i was set to 2 to assure stability. The coefficients a_{ki} and b_{ki} were set to 1 and -1 respectively for inhibiting links and to 0 and 1 respectively for activating links.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\dot{y}_i = \sum_{k\ \in\{1,\ ...,\ N\}/i} \left(a_{\text{ki}} + b_{\text{ki}} \frac{y_k^{h_{\text{ki}}}}{K_{\text{ki}}^{h_{\text{ki}}} + \ y_k^{h_{\text{ki}}}} \right) + \ u_i - \ \lambda_i y_i$$

Here, u_i denotes the basal gene expression rate. To compute the response y_{ij} of a gene i to a knock-out of gene j, the above shown system of ODEs was modified by setting the basal expression rate for the knocked out gene to zero and by removing all links onto and away from the gene j.

1.2.2 Using GeneNetWeaver to simulate absolute gene expression data

For better comparability with other publications, gene expression data was also generated with the program “GeneNetWeaver” [3]. This data was only used for comparing inference methods using ROC curves. From the “gold standard” yeast network in the program, random subnetworks were extracted with at least 100 regulators (random vertex seed; greedy neighbor selection). Then, the “Generate Kinetic Model” option was used with removal of auto-regulatory interactions to allow the generation of data sets. Data sets were generated with the following options: deterministic (ODEs) model, knock-out & wild type experiments, no time series, no noise added.

1.2.3 Logarithmic fold changes (M-values)

To simulate logarithmic fold changes that were similar to the M-values (logarithmic fold changes) of the yeast deletome data set described above, the M-values of yeast knock-out mutants were investigated for the knocked-out genes. Due to measurement noise, these values never reach negative infinity as would be expected in the absence of technical noise. The median minimal absolute fluorescence intensity averaged over all knocked out genes was estimated to be $2^{-2.5}$. This value was added to all simulated absolute gene expression level values before calculating the logarithmic fold changes. That is, the gene expression response of a gene i towards a perturbation of a gene j, expressed as logarithmic fold change, was calculated as $M_{ji} = log_2\left(\frac{y_{ji}+2^{-2.5}}{y_i+2^{-2.5}}\right)$, with y_{ji} the absolute expression level of gene i when gene j is perturbed, and y_i the expression level of gene i for the non-perturbed (wild type) state, as defined above. Replicates were produced through duplication of the logarithmic fold changes for each perturbation experiment.

1.2.4 Simulating measurement noise

Although the reference node activities (wild type data logarithmic fold changes) in the data set provided by Kemmeren et al. [1] are correlated, we did not simulate correlated measurement node activities. This is because we were lacking

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

information about the processes that generate biological variance as well as the actual network structure of *S. cerevisiae*. Because the biological noise dominated the technical noise, we did not distinguish both noise types in our simulations. Rather, we simulated noise by adding normal random variables to all simulated logarithmic fold changes.

1.2.5 Signal-to-noise ratio

The signal-to-noise ratio was calculated from the covariance matrix C by eigenvalue decomposition and analysis of the sorted eigenvalues. That is, the covariance matrix was decomposed into $C = U\Sigma U^T$, where Σ is a diagonal matrix that has its eigenvalues sorted in ascending order ($\sigma_{11} < \ldots < \sigma_{\mathrm{NN}}$) on its diagonal. Then, the signal-to-noise ratio (SNR) was calculated as $SNR = \sqrt{\frac{s-n}{n}}$, with the signal $s = \frac{1}{N_0}\sum_{k=1}^{N_0}\Sigma_{\mathrm{kk}}$ and the noise $n = \frac{1}{N-N_0}\sum_{k=N_0+1}^{N}\Sigma_{\mathrm{kk}}$. Here, N_0 denotes the number of perturbed nodes, and N is the total number of nodes.

2 Workflow

2.1 Data normalization

All node activities x were centered to the average of the corresponding reference node activities (wild type data logarithmic fold changes) $\bar{x}$, such that the average reference node activities became zero. That is, with data for N reference node activities and P perturbation node activities, for a gene i, we have for an arbitrary normalized node activity: $x_{\mathrm{ni}}^{\mathrm{norm}} = x_{\mathrm{ni}} - \bar{x}_i$, with n between 1 and $N + P$ and $\bar{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_{\mathrm{ki}}$.

2.2 Identification of significantly affected nodes

To identify which nodes were significantly affected by a perturbation, it was checked whether the node activities were significantly different from the background noise. This was done by comparing node activities (perturbation experiment logarithmic fold changes) to the corresponding reference node activity variance (wild type data logarithmic fold change variance) using t-tests. Because we had found that the biological variance was larger than the technical noise, only the number of biological replicates was used to estimate the degrees of freedom. To account for the multiple hypothesis testing, we adjusted the overall false detection rate (FDR) [4]. For the analysis of the *Saccharomyces cerevisiae* genome-wide knockout library, we followed the recommendation by Kemmeren et al. (supplementing information) to exclude several genes from further analysis. Furthermore,

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

we excluded all Pseudogenes and dubious ORFs listed on yeastgenome.org [5]. Because we had found that the distribution of normalized wild type experiment logarithmic fold changes can be well described by a t-distribution with 11 degrees of freedom, this number was used as the degrees of freedom for the t-tests instead of the number of biological wild type experiments. Although the microarray design used by Kemmeren et al. could be used to disentangle technical and biological noise, the biological noise clearly dominated the technical noise. Hence, we did not disentangle the variances for our analyses but rather used the overall variance. Nodes that were significantly affected in at least one perturbation experiment were kept in the network; all other nodes were removed from further analysis.

2.3 Identification of significant links

Inferable links were tested for significance using z-tests. The estimators of the link strength's variances were found by bootstrapping over node activity data (described below). The false discovery rate (FDR) was adjusted to 5 % to account for multiple hypothesis testing [4].

2.4 Bootstrapping

Residual bootstrapping was used (this was done to get a speed-up compared to parametric bootstrapping, in which residuals are drawn de novo from a distribution rather than used multiple times but in randomized order). First, residuals were calculated from the reference node activity data (wild type data logarithmic fold changes) and then added to the average node activity values of each perturbation experiment. For this addition step, we followed two approaches as explained in the following.

2.4.1 Utilizing correlations among reference node activities for analysis of yeast deletome data

Because we had found that wild type experiment logarithmic fold changes are highly correlated in the yeast deletome data, we sought to maintain the correlations among the reference node activity residuals to improve inference performance.

2.4.2 Minimizing the effect of random correlations

Because we did not simulate correlated measurement noise in our simulated data, the effect of random correlations among the reference node activities was sought to be minimized. This was achieved by shuffling the the order of reference experiments among the nodes.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.4.3 Residual bootstrapping procedure illustrated

Both bootstrapping approaches are illustrated in the following example, in which it is described how residual bootstrapping is performed on the node activity data for one perturbation experiment. (Residual bootstrapping of reference node activity data is analogous). The node activity data for N reference experiments of a two-node network with nodes i and j is represented as column vectors $\mathbf{x}_i^{WT} = \left(x_{i,1}^{WT},\ \ldots,\ x_{i,N}^{WT}\right)^T$ and $\mathbf{x}_j^{WT} = \left(x_{j,1}^{WT},\ \ldots,\ x_{j,N}^{WT}\right)^T$, and node activity data for a single perturbation experiment with two replicates is represented as $\mathbf{x}_i^{KO} = \left(x_{i,1}^{KO},\ x_{i,2}^{KO}\right)^T$and $\mathbf{x}_j^{KO} = \left(x_{j,1}^{KO},\ x_{j,2}^{KO}\right)^T$. The residuals for the nodes are calculated to be $\mathbf{r}_i{=}(r_{i,1},\ldots,r_{i,N})^T = \bar{x}_i^{WT} {-} \mathbf{x}_i^{WT}$ and $\mathbf{r}_j{=}(r_{j,1},\ldots,r_{j,N})^T = \bar{x}_j^{WT} {-} \mathbf{x}_j^{WT}$, with $\bar{x}_j^{WT}$ and $\bar{x}_i^{WT}$ the mean of the corresponding reference node activities. To leave correlations among reference node activities intact, we have the following expression for an arbitrary bootstrap sample for the perturbation node activities: $\mathbf{x}_i^{KO,boot} = \left(\bar{x}_i^{KO}, \bar{x}_i^{KO}\right)^T - (r_{i,p}, r_{i,q})^T$ and $\mathbf{x}_j^{KO,boot} = \left(\bar{x}_j^{KO},\ \bar{x}_j^{KO}\right)^T - (r_{j,p}, r_{j,q})^T$, with p and q being random (possibly equal) integers between 1 and N, and $\bar{x}_i^{KO}$and $\bar{x}_j^{KO}$the average node activities of the perturbation experiment. To destroy correlations among reference node activity residuals, the residuals are shuffled randomly across bootstrap samples. We have the following expression for an arbitrary bootstrap sample for the perturbation node activities: $\mathbf{x}_i^{KO,boot} = \left(\bar{x}_i^{KO}, \bar{x}_i^{KO}\right)^T - (r_{i,p}, r_{i,q})^T$ and $\mathbf{x}_j^{KO,boot} = \left(\bar{x}_j^{KO},\ \bar{x}_j^{KO}\right)^T - (r_{j,u}, r_{j,v})^T$, with $p,\ q, u$ and v being random (possibly equal) integers between 1 and N.

2.5 Clustering

Nodes were grouped into clusters if they were not sufficiently linearly independent. The following steps were followed:

1. All node activities were normalized to unit reference node activity (wild type data logarithmic fold change) variance.

2. A $(P \times N)$ node activity matrix was formed by merging experimental replicates (here, P denotes the number of experiments and N denotes the number of genes). Replicates (biological and technical) were merged by averaging and multiplication by the square root of the number of biological replicates.

3. For each pair of nodes k and l, a test statistic T_{kl} was compared to a 5 % FDR cutoff. The test statistic was calculated as the square of the smaller singular value of the $P \times 2$ matrix consisting of those two columns of the node activity matrix which corresponded to the nodes k and l. For simplicity, we assumed that the reference node activities were Gaussian distributed and

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that the node interactions were sufficiently linear. Then, under the Null Hypothesis of linear dependence, the test statistic approximately follows a Chi-square distribution with P degrees of freedom (the expected value of the $\mathcal{X}_P^2$ distribution over-estimates the expected value of the small eigenvalue, but this is negligible if sufficiently high confidence is claimed to reject the null hypothesis). An FDR of 5 % was chosen to define a significance cutoff.

4. Clusters were finally formed by grouping unperturbed nodes with perturbed nodes on which they were linearly dependent. If an unperturbed node was linearly dependent on multiple perturbed nodes, this node was grouped with the perturbed node onto which it was the most linearly dependent. Remaining unperturbed nodes were grouped with other unperturbed nodes if they were not sufficiently linearly independent.

5. For each cluster, either the perturbed node, or, if no perturbed node existed in the cluster, the node with the largest signal-to-noise ratio was chosen to be the cluster representative, and the corresponding node activity data was used for further analysis.

Links between clusters were found by inferring links between cluster representatives. Links within clusters can only be inferred within two-node clusters if there is exactly one perturbed node and if the unperturbed node is the only node affected by the perturbed node. Note that all perturbation experiment node activities were used to create the $(P \times N)$ node activity matrix and that no bootstrapping was done.

2.6 Subnetwork method

To infer a link from a node j onto a target node i using the subnetwork method, the network is temporarily limited to only those nodes that are significantly affected when node j is perturbed (the cutoffs used here were the same ones that were used for identifying the significantly affected nodes). This means that the inference of several links is skipped, and the corresponding link strengths are set to zero and are not included in the calculation of the FDR to determine a cutoff for link significance.

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3 Applications

3.1 Analysis of yeast deletome data

3.1.1 Inference of GAL network

The network was limited to all genes that could in principle be affected by perturbations of the following nodes: GAL3, GAL4, GAL80 and MIG1. A false discovery rate of 10^{-3} was used as a threshold for the identification of significantly affected nodes. With this threshold, 44 nodes remained in the network, that were then clustered. Only clusters that contained at least one of the above mentioned nodes were kept for the graph shown in figure 2d.

3.1.2 Lists of inferred links (consensus_stringent.xlsx and consensus_less_stringent.xlsx)

To generate a list of the most significant links inferred from the whole yeast deletome data, a consensus list was created that reflects the links inferred from the data by using two different methods of bootstrapping (see also the section on bootstrapping). The reason for this is that we are lacking a sophisticated model describing the process that generates biological noise. A consensus list should thus represent a more reliable model for link inference, as only the most significant links that were found by all methods survive the selection process. We created two different consensus lists (consensus_stringent.xlsx and consensus_less_stringent.xlsx), each corresponding to a certain cutoff to select significantly affected nodes. That is, for the most stringent cutoff, the network size was smaller than for the less stringent cutoff because less nodes were found to be significantly affected by perturbations. The cutoffs correspond to false discovery rates of 10^{-8} and 10^{-6}. The covariance matrix was calculated from all available node activity data for the significantly affected nodes. The following methods for residual bootstrapping were used (see also the section on bootstrapping) to create two separate link lists:

1. Correlations between residuals were left intact.
2. Correlations between residuals were destroyed.

The lists were merged by keeping only links that appeared in both lists and by then keeping the (maximally 500) most significant links with the highest average Z-score. Gene descriptions from the Saccharomyces Genome Database [5] were added to the list to allow easier investigation. We found that, among the most significant links, there are links whose source and target nodes are either adjacent on the genome or which are paralogs. These links are most likely false positives

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that reflect inspecificities of the gene deletion process used to generate the yeast deletion collection [2].

3.1.3 List of inferred hub nodes (consensus_stringent_hubs.xlsx)

From the consensus link list with the stringent cutoff for the selection of significantly affected nodes, we selected the 10 nodes with the highest number of significant outgoing links (significance cutoff: 5 % false discovery rate). The mean and median link strengths of the outgoing links were calculated for each of these 10 hub nodes. Gene descriptions from the Saccharomyces Genome Database [5] were added.

3.2 Simulations

3.2.1 Effect of noise-induced bias on relative link strength (suppl. figure 1d)

To show that network inference methods are biased towards measurement noise, we simulated node activity data for the limiting case of infinitely many replicates. In that case, all parameters are estimated perfectly because uncertainty resulting from a limited number of observations are averaged out. The covariance matrix of node activities was simulated according to the following formula:

$$\mathrm{C}_{\text{exact}} = \left(A^T\right)^{-1}\left(B^T B\right)(A)^{-1} + \sigma^2 I_N.$$

Where B is a diagonal matrix with element $b_{\text{ii}} = 1$ if node i is perturbed. The signal-to-noise ratio was adjusted by changing the value of σ^2.

3.2.2 Comparison of simple, clustering and subnetwork methods (figure 2c)

The Inferability curves represent averages over 4 different networks and 12 perturbation samples. The network size was 60 nodes. Gene expression data was simulated from scale free networks and a non-linear model (as described below).

3.2.3 Receiver Operating Characteristic (ROC) curves (figure 2b)

Simulation setup The ROC curves were averaged (as described below) over 24 perturbation samples and 4 network structures. The network size was 300 nodes, and 25 % of nodes were randomly perturbed in each perturbation sample. The noise-to-signal ratio was adjusted to 10 %. Data was generated with both GeneNetWeaver [3] and scale-free networks with a non-linear model to generate

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

stationary data. For each perturbation sample, only data corresponding to significantly perturbed nodes was used to calculate the covariance matrices (1 % FDR), the only exception being Lasso regression applied to the data of all nodes. All methods (except Lasso applied to the data of all nodes) received almost the same covariance matrix C as input: because of the randomness of bootstrap sampling, small numerical differences between the covariance matrices that PRC and the other methods received may have occurred. Binary classification was done solely based on links pointing away from perturbed nodes because only those links are potentially inferable. This is similar to the method described by Siegenthaler and Gunawan [6]. To infer a link onto a certain node, the data corresponding to a perturbation of that node was removed prior to calculating the covariance matrix. Because both Lasso regression and the subnetwork method set certain link strengths to exactly zero, it would be impossible to smoothly reach a false positive rate of 1. This is why we set all link test statistics that were set to zero to random values smaller than the smallest non-zero link test statistic. When creating the ROC curve, this procedure corresponds to random guessing for all links that were set to exactly zero.

ROC curve averaging In the following, the false positive rate is denoted by FPR and the true positive rate is denoted by TPR. To generate an averaged ROC curve, the tuples (FPR, TPR) of the individual ROC curves were assigned to 50 bins according to the FPR of each tuple. Bins were filled up in a way such that approximately equally many tuples were assigned to all bins. Then, for each bin, the FPR and TPR was calculated. These 50 averaged tuples (FPR, TPR) were used for the averaged ROC curve.

ROC curves for Subnetwork method and Lasso We generated 10 different ROC curves for the subnetwork method, each curve corresponding to a different cutoff for the selection of significantly affected nodes (based on which the subnetworks were created). For Lasso, we generated ROC curves for 10 different regularization coefficients. To model a smooth transition between the cutoffs or regularization coefficients, we interpolated (cubic spline) between the ROC curves in a way such that the resulting hybrid curve had a possibly larger area under the curve than the individual ROC curves. The 10 different cutoffs for selection of significant nodes correspond to the following FDRs: 1.00, 0.89, 0.78, 0.67, 0.56, 0.45, 0.34, 0.23, 0.12, and 0.01. The 10 different regularization coefficients were: 0, 0.001, 0.0018, 0.0032, 0.0056, 0.010, 0.0178, 0.0316, 0.0562, and 0.1000.

Comparison of non-regularized inference methods Test-statistics T_{ji} corresponding to a link from a node j to a node i were calculated according to the

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

following formulas. The standard deviations of the test statistics were estimated via residual bootstrap.

1. Regression: $T_{\text{ji}} = \frac{\left|G_{\text{ji}}\right|}{\text{std}\left(G_{\text{ji}}\right)}$, where $G_{\text{ji}} = \frac{(C)^{-1}_{\text{ji}}}{(C)^{-1}_{\text{ii}}}$.
2. Conditional Mutual Information: $T_{\text{ji}} = \frac{\left|G_{\text{ji}}\right|}{\text{std}\left(G_{\text{ji}}\right)}$, with $G_{\text{ji}} = \log_2\left(\frac{\det(A) \times \det(B)}{\det(C) \times \det(D)}\right)$. Here, A, B and D are covariance matrices with rows and columns of certain nodes removed. A: node j removed. B: node i removed. D: nodes i and j removed.
3. PRC: $T_{\text{ji}} = \frac{\left|G_{\text{ji}}\right|}{\text{std}\left(G_{\text{ji}}\right)}$, where $G_{\text{ji}} = \frac{\sum_{k=1}^{N_0} U'_{\text{ik}}U'_{\text{jk}}}{\sum_{k=1}^{N_0} U'_{\text{ik}}U'_{\text{ik}}}$, where N_0 is the number of unperturbed nodes and U_i is the i$^{\text{th}}$ eigenvector of the covariance matrix. Note that links were not identified as artifacts.

Comparison of regularized inference methods For a fair comparison, we used 10 different regularization coefficients to generate ROC curves for Lasso regression, keeping only the ROC curve corresponding to the regularization coefficient that gave the best performance. The following is a brief derivation of the formulas used for regularized regression using the L$_1$ and L$_2$ norms. The simulation setup is described afterwards. Consider the following equation, which is essentially equation 1 of the main paper except that perturbations are not restricted to single nodes, which is expressed through the perturbation strength matrix U:

$$\begin{aligned} \dot{Y} &= YA^T + U \\ Y^{\text{obs}} &= Y + \epsilon \end{aligned}$$

For simplicity, we assumed here that the reference state is 0. Here, $Y$are the true node activities that are masked by measurement noise ϵ to yield observed node activities Y^{obs}, and the matrix A^Tdenotes the link strengths. In the steady state and under the assumption that the measurement uncertainties ϵ are Gaussian distributed, a maximum a posteriori (MAP) function of the parameters A^T and U given the observed data Y^{obs} can be defined. The logarithm of this function is:

$$a\left(A^T, U|Y^{\text{obs}}\right) = K + \sum_{i=1}^{N}\sum_{n\in S_i}\left[\left(\sum_{j=1}^{N} Y^{\text{obs}}_{\text{nj}} A^T_{\text{ji}}\right) + U_{\text{ni}}\right]^2 \frac{1}{\sigma^2_{\text{ii}}} \qquad (1)$$

$$+ r^2 \sum_{i}^{N}\left(A^T_{\text{ii}} + \mu_i\right)^2 + v^2 \sum_{i}^{N}\sum_{n\in T} U^2_{\text{ni}} \qquad (2)$$

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Here, K is a normalization constant from the distributions that vanishes upon differentiation. The experiment indices n run over the set of all experiments except the ones in which node i is perturbed, S_i. N is the number of nodes; P is the number of experiments; $Y_{\mathrm{nj}}^{\mathrm{obs}}$ is the matrix of node activities (each row corresponds to a sample and each column corresponds to a node); σ_{ii}^2 is a variance coefficient and r^2, v^2 are regularization coefficients that correspond to weights of the prior distributions. Note that the first prior distribution is only defined over the diagonal elements of the network matrix A. It can be shown that the following estimator maximizes the MAP function:

$$\left(\hat{A}^T\right)_{\mathrm{ji}} = \frac{\left(\Phi^T\Phi\right)_{\mathrm{ji}}^{-1}}{\left(\Phi^T\Phi\right)_{\mathrm{ii}}^{-1}}\mu_i.$$

Here, Φ denotes the observed node activities with samples corresponding to a perturbation of node i removed (the number of rows of Φ is less than or equal to the number of rows of Y^{obs}, depending on whether node i is perturbed). This means that the data may need to be re-organized to infer links onto each node. To arrive at this result, one needs to maximize $a\left(A^T, U|Y^{\mathrm{obs}}\right)$ with respect to A. This is possible because the diagonal elements of the network matrix A are regularized, which does not allow for arbitrary combinations of either A or U. The following steps lead to the above stated solution:

1. Set the derivative of the log-MAP function w.r.t. U_{ni} to zero: $\frac{\delta a\left(A^T, U|Y^{\mathrm{obs}}\right)}{\delta U_{\mathrm{ni}}} = 0.$

2. Solve for U_{ni} to obtain the MAP estimator $\hat{U}_{\mathrm{ni}}$. Plug this expression into formula for log-MAP function.

3. Set this new expression of the log-MAP function w.r.t. $A_{\mathrm{ji}}^T$to zero: $\frac{\delta a\left(A^T, U|Y^{\mathrm{obs}}\right)}{\delta A_{\mathrm{ji}}^T} = 0.$

4. Solve for A_{ji}^T to obtain the MAP estimator$\hat{A}_{\mathrm{ji}}^T$. To achieve this, one has to apply the Sherman-Morrison-Woodbury formula and let the regularization coefficient r^2 go to infinity, which essentially sets all $\mathrm{A}_{\mathrm{ii}}^T$ equal μ_i.

Note that, for $\mu_i = -1$, the set of parameters A_{ji}^T (except for $\mathrm{A}_{\mathrm{ii}}^T$, which become equal to μ_i) are simply the ordinary least squares (OLS) estimators that minimize the following expression:

$$F = \|\Omega\beta - \Phi_i\|_2^2$$

Here, Ω is a matrix that is equal to Φ except that it is missing column i, and β is a vector that is equal to A_i^T except that it is missing the element equal

bioRxiv preprint doi: https://doi.org/10.1101/082925; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to A^T_{ii}. The L_1 Norm ("Lasso" regularization) and L_2 Norm ("Tikhonov-Miller" regularization) impose further constraints on this extreme value problem. The corresponding formulas are:

$$F_{L1} = \|\Omega\beta - \Phi_i\|_2^2 + \gamma_{L_1}\left\|A_i^T\right\|_1$$

and

$$F_{L2} = \|\Omega\beta - \Phi_i\|_2^2 + \gamma_{L_2}\left\|A_i^T\right\|_2^2$$

for the L_1 and L_2 Norm, respectively.

4 Software

For all analyses and simulations, we used MATLAB and Statistics Toolbox and Parallel Computing Toolbox Release R2014b, The MathWorks, Inc., Natick, Massachusetts, United States. Our MATLAB code is available on demand.

References

[1] Kemmeren, P. *et al.* Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. *Cell* **157**, 740–752 (2014).

[2] Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387–391 (2002).

[3] Schaffter, T., Marbach, D. & Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* **27**, 2263–2270 (2011).

[4] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 289–300 (1995).

[5] Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).

[6] Siegenthaler, C. & Gunawan, R. Assessment of Network Inference Methods: How to Cope with an Underdetermined Problem. *PLoS ONE* **9**, e90481 (2014).