Published online 23 November 2015

Nucleic Acids Research, 2016, Vol. 44, No. 5 ***e50***
doi: 10.1093/nar/gkv1284

Systematic identification and correction of annotation errors in the genetic interaction map of *Saccharomyces cerevisiae*

Nir Atias[1], Martin Kupiec[2,*] and Roded Sharan[1,*]

[1]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel and [2]Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv 69978, Israel

Received July 22, 2015; Revised October 21, 2015; Accepted November 4, 2015

ABSTRACT

The yeast mutant collections are a fundamental tool in deciphering genomic organization and function. Over the last decade, they have been used for the systematic exploration of ∼6 000 000 double gene mutants, identifying and cataloging genetic interactions among them. Here we studied the extent to which these data are prone to neighboring gene effects (NGEs), a phenomenon by which the deletion of a gene affects the expression of adjacent genes along the genome. Analyzing ∼90,000 negative genetic interactions observed to date, we found that more than 10% of them are incorrectly annotated due to NGEs. We developed a novel algorithm, GINGER, to identify and correct erroneous interaction annotations. We validated the algorithm using a comparative analysis of interactions from *Schizosaccharomyces pombe*. We further showed that our predictions are significantly more concordant with diverse biological data compared to their mis-annotated counterparts. Our work uncovered about 9500 new genetic interactions in yeast.

INTRODUCTION

The yeast *Saccharomyces cerevisiae* has been a model organism for genetic studies since the 1950s (1). Among its obvious advantages are a rapid growth rate, easy-to-carry-out genetics and biochemical procedures, the ability to introduce foreign DNA and the ease with which genomic manipulations can be implemented. The construction of ordered mutant libraries jump-started the systems biology revolution by allowing the systematic analysis of a variety of cellular phenotypes. The non-essential deletion collection consists of ∼4700 yeast strains, each carrying a precise deletion of a single, non-essential gene, replaced by a marker gene that confers antibiotic resistance (2). Two additional collections include either hypomorphic (3) or temperature-sensitive (4,5) alleles of the remaining ∼1100 essential genes. Going beyond single genes, the development of the Synthetic Genetic Array (SGA) technology (6,7), allowed the large scale systematic surveys of double mutants. One of the most ambitious goals of these large-scale explorations is an investigation of all the possible genetic interactions (GIs) in yeast. In brief, each mutant is combined with all the others, and the growth rate of the double mutants is compared to that of the single mutants. A differential increase or decrease in fitness (measured by the size of the colonies formed) implies either a positive or a negative genetic interaction. As the genome of *S. cerevisiae* contains ∼6000 genes, the completion of this project will require the creation of ∼36 000 000 yeast strains. Up to now, ∼6 000 000 of these have been created and analyzed (8). The resulting database constitutes one of the most important resources for the analysis of genomic function.

One obvious assumption, when using these mutant collections, is that any phenotype observed in a particular strain stems from the specific mutation carried by the strain (either a perfect deletion, an hypomorphic or a temperature-sensitive allele). However, the apparent direct relation between the mutated gene and the phenotype is sometimes misleading (9–14). As previously demonstrated (15), the mutations may affect the expression of genes located next to the deleted genes along the genome. We refer to this effect as the neighboring gene effect (NGE). We have previously shown that NGE contaminates the results of systematic analyses in yeast in a non-trivial fashion: analysis of four different genetic screens found between 7 and 15% of NGE cases (15). The direct result of the NGE is a double mis-annotation: the correct gene mutation is annotated as having a phenotype it lacks, whereas the gene responsible for the phenotype (the adjacent gene) is not identified. It is therefore important to identify NGE cases and correct the annotation. Briefly, our previous approach selected between deleted genes and their neighbors by connecting them, via a protein–protein interaction (PPI) network with genes that

*To whom correspondence should be addressed. Tel: +972 3 6409031; Fax: +972 3 6409407; Email: martin@post.tau.ac.il
Correspondence may also be addressed to Roded Sharan. Tel: +972 3 640 7139; Fax: +972 3 640 9373; Email: roded@post.tau.ac.il

are known to be central to the studied phenotype. We reasoned that NGE may also result in mis-annotations among the large-scale GI network being created. The approach developed, however, cannot be applied to growth phenotype where the fitness of different mutants is affected via a variety of different biological mechanisms (16).

Here we study potential annotation errors in the largest database of genetic interaction data available to date (8,17), showing that systematic biases due to NGE greatly affect these data. We present the Genetic Interaction Neighboring Gene Effect Recovery (GINGER) algorithm to detect cases in which the neighboring gene, rather than the deleted one, should be considered the real genetic interactor. GINGER overcomes the need to specify central genes, by using data on physically interacting genes as partial evidence to the similarity between their genetic profiles. Subsequently, these local evidences are combined so that the overall evidence over the set of available interactions is maximized. Our analysis suggests significant NGE biases that cover more than 10% of the experimental results reported to date.

MATERIALS AND METHODS

S. cerevisiae genetic interaction data

Raw genetic interaction data for 6.6M double knockout experiments were downloaded from Costanzo *et al.* and filtered using the cutoffs recommended by the authors ($|\epsilon| > 0.08$ and P-value < 0.05) (8,17). Reciprocal interactions with ϵ having different signs or where one of the P-values did not meet the cutoff were removed. Focusing on negative interactions among non-essential genes, these thresholds yielded an initial set of 88 037 genetic interactions. We additionally used stringent cutoffs, as defined by Costanzo *et al.*, to create a 'strict' dataset ($\epsilon < -0.12$ or $\epsilon > 0.16$; P-value < 0.05).

Construction of a gold standard

Genetic interaction data for *Schizosaccharomyces pombe* were compiled from two recent studies (18,19). Low confidence interactions (S-score > -2.0) were not considered in our analysis. Reciprocal interactions with S-scores indicating different interaction type (negative versus positive) were additionally removed from the dataset, resulting in a set of 121 921 negative genetic interactions.

Information about orthologouos genes between *S. cerevisae* and *S. pombe* was downloaded from PomBase (20), discarding non-unique *S. cerevisae* to *S. pombe* mappings. Interactions that are conserved between *S. cerevisae* and *S. pombe* were considered as positive (non-NGE) cases. For the negative (NGE) cases, we compiled a list of suspected interactions; these are interactions that are not conserved but when one of the interactors was replaced with its neighbor the interaction was found in *S. pombe* but not in *S. cerevisiae*.

The training set was refined by removing interactions that were neither found in low-throughput experiments nor supported by the network. The set of low-throughput experiments was downloaded from BioGRID (21). Network support is measured by the number of incomplete bipartite graphs containing the suggested interaction as the missing edge. The bipartite graph motif was previously shown to be a good predictor for potential genetic interactions (22,23). Accordingly, we removed suspected interactions if they were supported by fewer bipartite motifs than the observed ones, and removed conserved interactions if they were supported by less than 5707 (top 25% of all conserved interactions) bipartite motifs. The low-throughput based lists contained 295 positive and 51 negative interactions. The network-support based lists contained 378 positive and 196 negative interactions. The combined gold standard contained 460 positive and 215 negative cases.

Optimal operating point

We calculated the optimal operating point on the receiver operating curve (ROC) according to two established methods. The Youden Index (24) specifies that the optimal point on the curve as the one with maximum distance from the expected performance on random data (where the false positive rate equals the true positive rate). A different criterion (25) denotes the optimal point as the one closest to the perfect classifier (where the true positive rate is one and the false positive rate is zero). Application of both methods indicated an optimal cutoff of 0.8.

Physical interactions data

We downloaded physical interaction data from BioGRID (21). We assigned confidence scores to edges based on the experimental evidence supporting them using a logistic regression model, as previously described (26). The choice of training sets for the logistic regression was also previously described (27). Briefly, 500 positive examples denoting high confidence interactions were extracted from KEGG (28). Similarly 500 negative interactions were defined as those whose end points are the most distant in the physical interaction network when that interaction was removed. The interaction data and scores are available in ANAT (29). In this study we removed edges with very low confidence (<0.4; see 'Parameters estimation' section). Overall our PPI network contained 14 843 interactions among 3826 proteins.

Genomic data

Genomic loci for *S. cerevisiae* were downloaded from the Saccharomyces Genome Database (30). Genomic loci for *S. pombe* were downloaded from NCBI Gene database (31). Genomic locations for genes were defined by their transcription start site (TSS) and transcription stop point (TSP). Genomic intervals were ordered so that every gene is associated with a genomic interval (s, e) such that $s = \min(\text{TSS}, \text{TSP})$ and $e = \max(\text{TSS}, \text{TSP})$, $s < e$ regardless of the strand on which the gene is located. Neighboring genes were defined as those whose genomic distance was less than 350 bp, the median intergenic distance. Formally, the distance between genes A and B ($s_A \leq s_B$) is given by: $s_B - s_A$ if $e_A \leq e_B$ or 0 otherwise. By this definition non-positive distances denote overlapping genes.

The GINGER algorithm

The GINGER algorithm integrates genetic interaction data, physical interaction data and data on the genomic co-

ordinates of genes; it aims to identify the set of interactions that are best supported by all the data sources. The first step in the algorithm is to supplement the set of 88 037 observed genetic interactions with ones that may arise due to NGE. Formally, denote by N(g) all the genes whose transcription starts or stops within 350 bp from the transcribed interval of g (see 'Genomic data' section). Then, the interaction A–B is supplemented with A′–B′ for every A′ in N(A) and B′ in N(B). Note that some of the interactions A′–B′ may also be present in the database (e.g. the observed interaction) and we therefore keep track of whether an interaction was observed in the experimental data or not. This step greatly increased the size of the database (∼4.2-fold; 376 133 interactions).

In the second step of the algorithm the information from all the data sources is encoded in a single 'support' network. In this network, nodes correspond to genetic interactions and edges connect interactions that are supported by the PPI data. Precisely, we searched the physical and genetic interaction data for a motif where physically interacting proteins A–B have a common genetic interactor C and added the nodes A–C and B–C to the support network, connected by an edge. Genetic interactions that were not supported by any physical interaction were removed, and the resulting network spanned 133 547 nodes. Next, we scored the edges of the support network in the following manner: edges connecting nodes that correspond to experimentally observed interactions were given a score of 1.0 (full support). In contrast, the score for edges that connect at least one putative interaction exponentially decays with the number of neighbors used in the putative interactions they connect. Formally, let $p_{A\text{-}C}$ denote the number of neighbors used to infer the interaction A–C ($p_{A-C} \in \{0, 1, 2\}$) then the score for the edge connecting A–C and B–C is given by: $\alpha^{p_{A-C}} \cdot \alpha^{p_{B-C}}$, where α is a parameter denoting the penalty for deviating from the experimental data.

In the final step of the algorithm we search the support network for a collection of nodes (genetic interactions) with maximum support (connected by high-scoring edges). We constrained the search procedure to select at most one candidate interaction from the set of candidates derived from each observed interaction. We formalize this optimization problem as an integer linear program. The program uses two sets of binary variables (i) Y(v) indicating which nodes (interactions) are selected by the algorithm, and (ii) X(e) indicating which edges are connecting the selected nodes. Let c_e denote the score associated with edge e then the objective is trivially given by maximizing $c_e^T \cdot X(e)$. The program uses two types of constraints. The first type, $X(e) < Y(u),\ Y(v) \quad \forall e = (u, v)$, ensures that edges may only be selected if both their corresponding nodes are selected. The second type, $\sum_{v \in N_{A-B}} Y(v) \leq 1 \quad \forall N_{A-B}$, ensures that at most one interaction is selected from the set of candidate interactions $N_{A\text{-}B}$ derived from the observed interactions A–B. The complete program is given below:

$$\begin{aligned} \max \quad & \sum_e c_e^T \cdot X(e) \\ \text{s.t.} \quad & X(e) \leq Y(u),\ Y(v) \quad \forall e = (u, v) \\ & \sum_{v \in N_{A-B}} Y(v) \leq 1 \quad \forall N_{A-B} \\ & X(e),\ Y(v) \in \{0, 1\} \quad \forall e, v \end{aligned}$$

The resulting program may admit multiple equally-good solutions. We therefore associate predictions with confidence scores by repeatedly perturbing the program and recording the number of optimal solutions containing each prediction. In detail, we solve the above program 100 times while adding a zero-mean Gaussian noise (with standard deviation controlled by the parameter β) to the coefficients c_e. The confidence score for a given interaction is the number of times it appears in the optimal solutions for the randomized programs.

Parameter estimation

The performance of the GINGER algorithm may be affected by several parameters. In particular, we studied the effect of the penalty for deviation from observed data (α), the amount of noise added to the objective (β) and the cutoff for filtering low-confidence PPI data. We followed a grid-search procedure exploring all the combinations for setting α in the range 0.1–1.0 in 0.1 steps, $\beta \in \{0.05, 0.1, 0.15\}$ and the threshold for PPIs in the range 0.2–0.7 in 0.1 steps. Top performance was achieved with PPI cutoff of 0.4, $\alpha = 0.2$ and $\beta = 0.1$. Notably, the performance was robust to small perturbations: the top three performing parameter sets differed only in the amount of noise used to evaluate the score the predictions (the parameter β). The top nine performing parameter sets additionally differed only in the penalty for deviation from the experimentally observed data, with α set to either 0.4 or 0.3. An implementation of the algorithm is freely available (File S2).

Co-expression data

Co-expression data were downloaded from COEXPRESSdb (32). As per the authors recommendations we only considered gene pairs with mutual rank ≤200.

Enrichment calculation

We downloaded the association of genes with Gene Ontology (GO) terms from the GO consortium website (33). We found that ∼10% of the associations were based solely on genetic interaction studies and therefore removed them from the analysis to avoid potential biases. Similarly, we removed annotations based on physical interactions or mutant phenotype data. Overall, we retained 15 325 annotations out of the 26 482 published ones. For a fair comparison, interactions were limited to those spanning common proteins. Enrichment scores were computed for genes pairs sharing the exact same category. We only focus on specific categories having no more than 30 annotations. The reported *P*-values were computed for the two-tailed Fisher Exact Test and were corrected for multiple hypotheses testing via false discovery rate (FDR).

RESULTS

Our hypothesis is that the deletion in the mutant strains may affect the behavior of neighboring genes, thus leading to incorrect association with a given phenotype. Under this working model, the phenotype in question—slower growth of a double knockout—is likely to be observed in the strain in which the true causative gene is mutated as well as in the strain in which its neighboring gene (genomic distance <350 bp) was knocked out. As a simple proof of principle, we scanned the available collection of negative genetic interactions for neighboring genes sharing common interactors. Indeed, we found that a very large fraction (>7.4%; 7270) of the reported set of negative interactions satisfied this criterion (Figure 1A). This observation was highly significant compared to the expected number of interactions with neighboring genes in a randomized network that preserves node degrees ($P < 0.001$; empiric P-value).

Algorithmic approach

Following previous works (22–23,34), we focus on the set of negative genetic interactions which covers the majority of the mapped interactions. When a double mutant shows a reduction in fitness, it is commonly concluded that the genes mutated exhibit a genetic interaction (8,35–36). As explained, the genetic interactions may be caused by either of the deleted genes or any of their neighbors, or even by interactions between the neighboring genes of each deletion. To identify and correct NGEs, we make use of PPI data, which does not depend on the yeast mutant libraries and thus is not prone to be affected by NGE. We rely on the empirical observation that proteins whose encoded genes have similar genetic interaction profiles tend to physically interact and *vice versa* (8,37) (see Figure 1B). This observation allows us to disambiguate candidate genetic interactions by preferring those that increase the genetic profile similarity of physically interacting proteins.

Our computational approach is outlined in Figure 2 (see 'Materials and Methods' section for details): First, the set of 88 037 experimentally observed genetic interactions is supplemented with additional putative interactions that include the neighboring genes (for a total of 376 133 putative interactions). Second, we integrate the physical and genetic interaction data into a network whose nodes correspond to the possible genetic interactions and edges connect the interactions that are further supported by the physical interaction data. Edges are additionally weighed to penalize deviations from the experimental data. Removing nodes that were not supported by the physical interaction data, leaves a total of 133 547 candidate genetic interactions. Finally, we search this network for a collection of genetic interactions (nodes) that are maximally supported by both genetic and physical interaction data. In our formulation we additionally require that no more than one interaction is chosen from a set of candidate interactions generated for a single experimental observation. We formalize the optimization problem as an integer linear program and solve it using a dedicated solver, obtaining a set of interaction predictions along with confidence scores ('Materials and Methods' section). Using GINGER we were able to generate confident predictions for 49 698 (56%) experimentally detected interactions.

Evaluation

In order to evaluate the predictions generated by GINGER, we compiled a gold standard containing both positive (true non-NGE) and negative (true NGE) examples. For the positive examples, we used the set of genetic interactions found to be conserved in budding yeast and in the fission yeast, *S. pombe*. Although *S. pombe* and *S. cerevisiae* have a similar number of genes (of which around two thirds are confirmed homologs), the genomic architecture (synteny) of these two organisms differs (38). Thus, neighboring genes are not shared between the two organisms, and conserved genetic relationships are likely to represent true interactions (39).

Compilation of a gold standard negative set is challenging as it requires the identification of true NGE cases. To this end, we first compiled a list of suspected NGEs by searching for *S. cerevisiae* interactions that were not con-

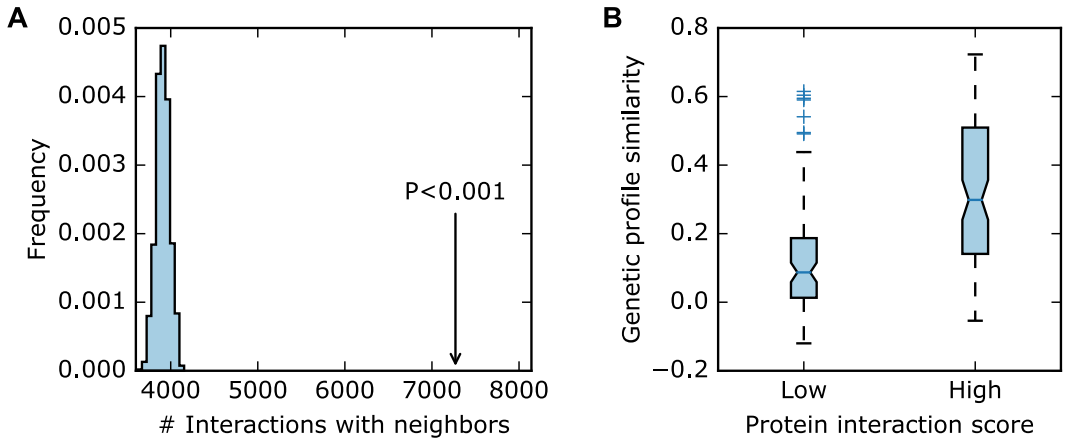

Figure 1. Genetic interaction and neighboring gene statistics. (**A**) Number of neighboring gene pairs that share an interactor compared to a random distribution (based on 1000 degree-preserving randomizations). (**B**) Proteins with high-confidence physical interaction score (≥0.5) are more correlated in their genetic interaction profile.

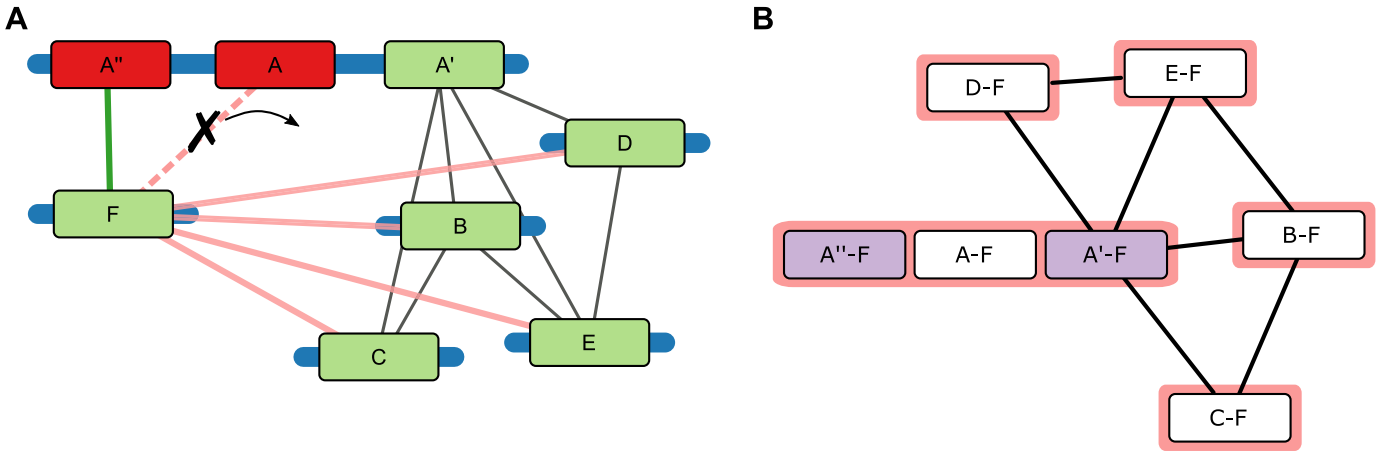

Figure 2. Overview of GINGER. (**A**) The input to the GINGER algorithm is data on genetic interactions (light red), physical interactions (gray) and genomic loci (blue lines denote DNA molecules). Genetic interactions are supplemented with candidate interactions to neighboring genes (green edges). (**B**) GINGER builds a 'support' network to disambiguate genetic interactions. In this network nodes represent both observed genetic interactions (white nodes) and candidate interactions (violet nodes). Nodes are further grouped (pink rectangles) based on the observed interactions (white nodes) that they might correct. Physically interacting genes having a common genetic interactor attest to similarity in the genetic profile and the genetic interactions are connected in the support network (e.g. the physical interaction B–C supports the genetic interactions B–F and C–F). GINGER aims to pick as many edges (support) as possible without touching more than one node in each group. In the example the interaction A′–F is preferred to either A–F or A′′–F (red nodes; panel A) and the decision is supported by the additional interactions with the complex A/B/C/D/E.

served in *S. pombe,* but for which a genetic interaction exists in *S. pombe* when one of the *S. cerevisiae* genes is replaced by its genomic neighbor (Figure 3A). This criterion is likely to introduce false positives because (i) more combinations involving neighbors exist, and (ii) the data on genetic interactions in *S. pombe* is incomplete. Therefore, for both positive and negative examples we retain only those interactions that, in addition, were either (i) confirmed in a low-throughput experiment, or (ii) have increased support in the genetic interaction network (6,8,23,37) (see 'Materials and Methods' section and Figure 3B). Overall our gold standard contained 460 positive examples and 215 negative examples.

Applying GINGER to the GI data, we observed that its predictions were highly concordant with interactions that were confirmed by low-throughput experiments, achieving an area under the ROC (AUC) of 0.94 (Figure 3C). When also taking into consideration suspected NGE cases that were supported by the topology of the network, the AUC, as expected, slightly decreased (to 0.83). In order to determine a cutoff for high-quality predictions, we calculated the optimal operating point according to the Youden Index (24). The optimal operating point corresponded to a threshold of 0.8. Under this cutoff, most of GINGER's predictions (32 269/49 698; 65%) were considered *confident*. The set of confident predictions contained 9438 predicted NGE cases (Supplementary Table S1) representing a substantial part (10.7%) of the original set of 88 037 negative genetic interactions and 29.2% of the predictions. In the following analyses we focus on these confident predictions.

GINGER uncovers non-trivial NGE cases

The yeast genome includes ∼800 open reading frames (ORFs) not likely to encode for proteins. These are called 'dubious ORFs', and many times (71%, 563/784) overlap other, confirmed ORFs. Since the deletion collection contains deletions of many of these dubious ORFs (which also inactivate the overlapping gene), one trivial possibility is that the potential NGE cases identified are mainly of this category. The information on whether an ORF is dubious or not was not available to GINGER and we use it as an additional measure of the algorithm's performance. Of the 9438 confident NGE predictions, only 1037 (10.9%) interactions (involving 153 genes) include a dubious ORF overlapping a confirmed gene. Reassuringly, the corrections suggested by GINGER significantly reduced the number of interactions involving dubious ORFs overlapping with confirmed genes to 224 ($P < 4.5E\text{-}24$; Wilcoxon) and the number of dubious ORFs they spanned to 50. Thus, GINGER recognizes and corrects most errors caused by the deletion of overlapping genes, and the majority of NGE events uncovered by GINGER do not stem from dubious ORFs.

Enrichment analysis

Negative genetic interactions usually indicate functional relationships between the interacting genes, and are often used to place two mutants in the same cellular process or pathway (37,40–41). In line with these previous approaches, we sought to assess the quality of our predictions in the context of current biological knowledge. To this end, we considered four sets of interactions: (i) *All* is the set of all experimentally observed interactions; (ii) *GINGER* is the set of 'corrected' interactions that GINGER suggests as an alternative to the experimentally observed ones that are predicted to be affected by NGE; (iii) *Predicted non-NGE* is the set of experimentally observed interactions that GINGER considers correct; and (iv) *Predicted NGE* is the set of experimentally observed interactions that GINGER considers as incorrect, i.e. NGE-affected. We measured the agreement with biological knowledge by calculating the enrichment (42) of each set of interactions in (i) the set of pairs of genes

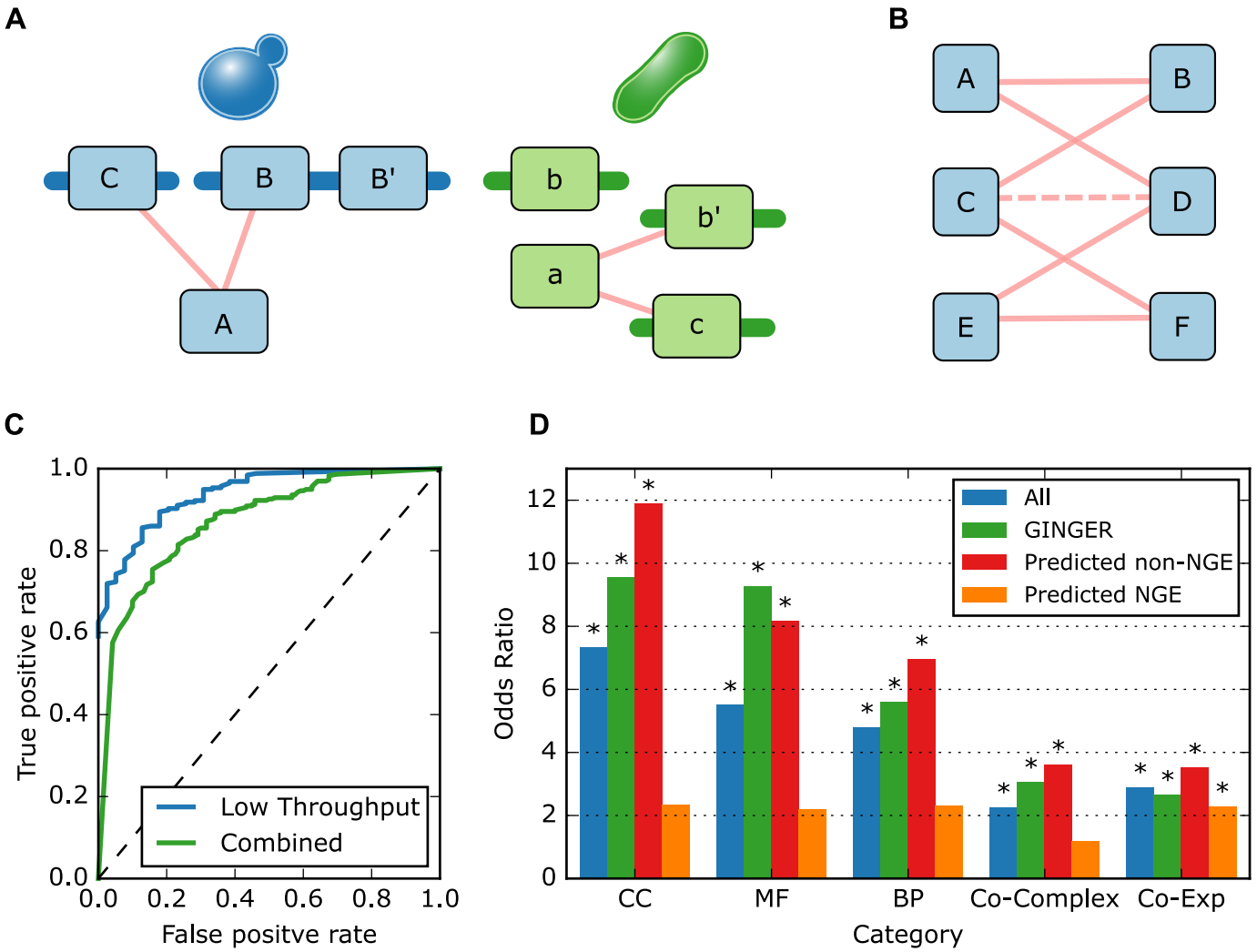

Figure 3. Evaluation of the GINGER algorithm. (**A**) To compile a gold standard for our predictions in *Saccharomyces cerevisiae* (blue) we used genetic interaction data from *Schizosaccharomyces pombe* (green). A conserved interaction (A-C/a-c) is more likely to be true and thus is considered a case where NGE does not occur. On the other hand interactions that are not conserved (A–B) but become conserved when considering a neighbor (A-B′/a-b′) are likely to represent NGE cases. We additionally require support from low throughput experiments or network topology to qualify such cases for the gold standard. (**B**) Network topology support. The interaction C–D is supported by two incomplete bipartite graphs, one spanning A, B, C and D, and the other spanning C, D, E and F. (**C**) Evaluation of GINGER's predictions versus the gold standard, using only low-throughput experiments as additional criterion for NGE cases yields AUC of 0.94, adding the motif-supported interactions reduced the AUC to 0.83. (**D**) Enrichment analysis of confident predictions for co-occurrence in GO categories from cellular component (CC), molecular function (MF) and biological process (BP). In addition, we test for enrichment for complex membership (Co-Complexes) and for co-expression (Co-Exp). Asterisks denote significance with FDR-corrected $P < 0.05$.

annotated with identical GO terms (33), (ii) membership in curated complexes (43) and (iii) pairs of co-expressed genes (32). We excluded GO annotations with evidence based on genetic interactions (see 'Materials and Methods' section).

The enrichment analysis highlights the utility of GINGER in capturing current biological knowledge. Specifically, the set of *Predicted NGE* interactions was not significantly enriched in most of the tested datasets, except for the set of co-expressed genes (Figure 3D). In sharp contrast, the sets derived from GINGER's predictions (*GINGER* and *Predicted non-NGE*) proved significant across the entire test set and achieved higher odds-ratio rates across almost all the tested categories. More importantly, the enrichment of GINGER's predictions resembled more closely the enrichment of the original set of interactions in the database whereas the *Predicted NGE* set did not. The largest differences in performance were noticed in enrichment in the various GO namespaces (Molecular Function: $P < 7.1E\text{-}9$, Cellular Component: $P < 5.3E\text{-}16$ and Biological Process: $P < 3.1E\text{-}3$. *P*-values are FDR corrected). A milder gap was noticed in the enrichment for membership in complexes which is probably due to negative interactions being less prevalent within complexes (44). Interestingly, we found that all the sets achieved near-similar performance for the enrichment in co-expression data. A possible explanation is that neighboring genes tend to co-express (45) (Odds ratio 41.4; $P = 0$; Two-tailed Fisher Exact Test).

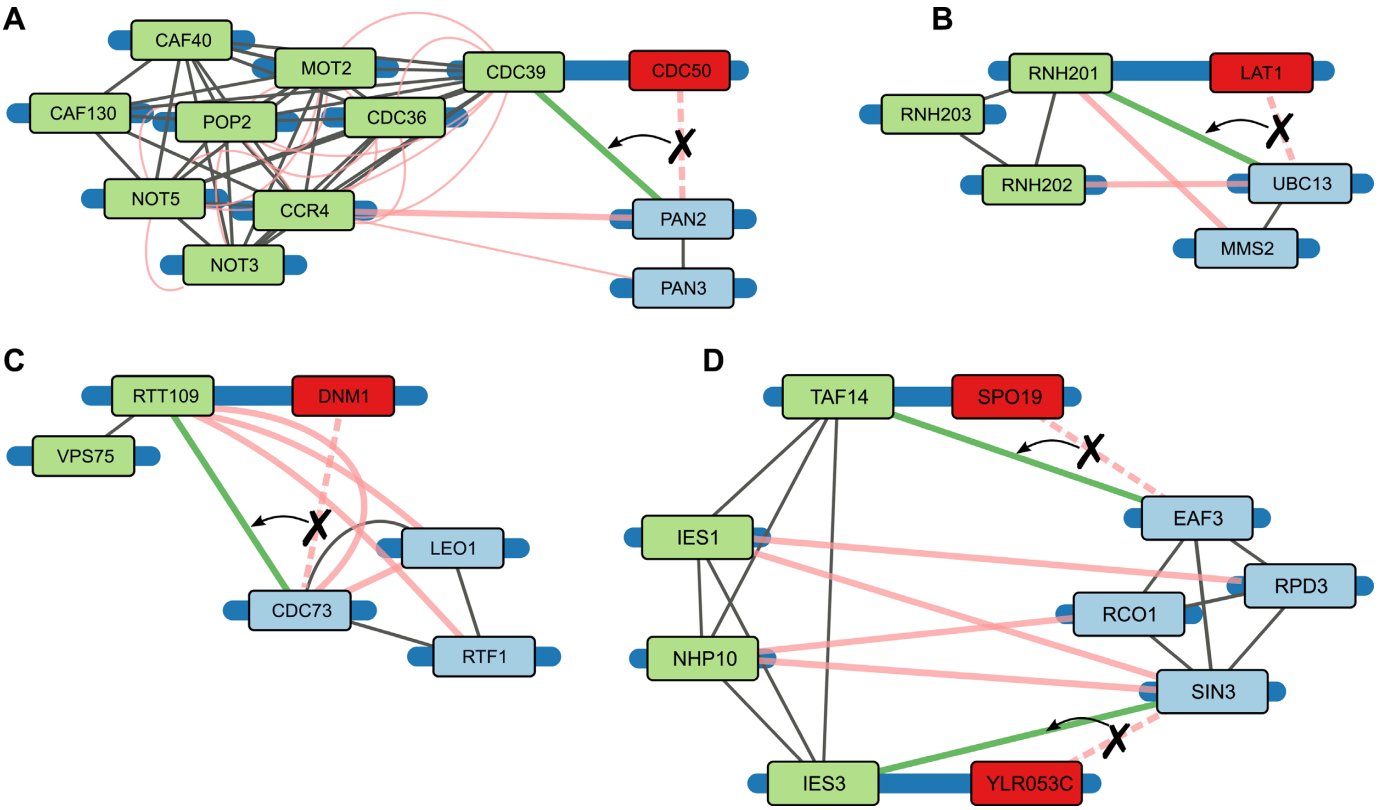

Figure 4. Examples of GINGER's novel predictions. (**A**) GINGER predicted that the observed interaction between *PAN2* and *CDC50* (red) is incorrect and should be attributed to NGE. Instead GINGER suggests *CDC39*, a member of the CCR4-NOT complex (green). Both CCR4-NOT and PAN complexes (blue) regulate mRNA levels while *CDC50* is an endosomal protein. (**B**) The observed interaction between *UBC13* and *LAT1* (red) is predicted to be incorrect. GINGER suggests *RNH201*, a member of RNase (green), as the correct interactor. RNase and Ubc13/Mms2 complexes (blue) provide alternative DNA repair mechanisms while *LAT1* is a mitochondrial gene. (**C**) GINGER replaced the observed interaction between *CDC73*, a member of the Paf1 complex (blue) and *DNM1* (red) with an interaction with *RTT109*. Rtt109 and Vps75 (green) form an Histone acetyltransferase that acetylates histone H3. Consistently, one of Paf1's roles is histone modification. Dnm1 is a mitochondrial protein. (**D**) GINGER replaces the observed interaction between *EAF3*, a component of Rpd3S histone deacetylase complex (blue) and *SPO19* (red), which encodes a meiosis-specific protein, with an interaction with *TAF14*, a subunit of Ino80 complex (green). Another NGE was detected between *SIN3* and the uncharacterized gene *YLR053C* (red). Edge color coding: gray denotes physical interaction. Red edges denote genetic interactions where wide edges are high-confidence DRYGIN interactions (available to GINGER) and thin edges are lower-confidence information from DRYGIN or synthetic lethal interactions in BioGRID, both not available to GINGER. Dark blue lines denote DNA

Analysis with high scoring genetic interactions

Costanzo *et al.* designate a subset of their genetic interaction database as being of exceptional quality, based on particularly high genetic interaction scores. We applied our algorithmic approach to this dataset of 47 427 interactions. Encouragingly, GINGER produced 35 313 (74%) confident interactions, markedly higher than the rate obtained for the weaker quality dataset (56%). Comparing them against our gold standard revealed that the quality of these predictions was indeed higher than the ones obtained from the weak dataset. In particular, agreement with low-throughput interactions (AUC: 0.94) was maintained and even increased (AUC 0.87) when the gold standard was supplemented with cases derived from topological information. Based on the Youden index 22 870 interactions were deemed in highly confidence, of them 4128 (18%) corresponded to NGE cases. Enrichment analysis applied to these predictions revealed that their performance was better than the performance of those derived from the weaker dataset. Our results are summarized in Supplementary Figure S3.

DISCUSSION

The genetic interaction map of *S. cerevisiae* constitutes an invaluable resource for all systems biology researchers. Here, for the first time, we systematically scan the largest genetic interaction dataset gathered to date for systematic errors. Our analysis suggests that a large fraction (>10%) of the annotations in this invaluable data source may be misleading and that the annotations should probably be attributed to genes in the vicinity of the deletion loci. In our study we developed a novel approach, GINGER, that not only detects the possible errors in annotations but also suggests the way to correct them. Overall, GINGER provided close to 9500 new genetic interaction pairs (Supplementary Table S1).

The success of GINGER in detection and correction of NGE biases is apparent not only through large-scale analysis but also in many specific cases where the observed interactions seem highly unlikely. For example, *CDC50* encodes an endosomal protein that interacts with phospholipid flippase Drs2p (46–48). In DRYGIN (17) *CDC50* is

annotated with high confidence as a genetic interactor of *PAN2*, a subunit of the Pan2p-Pan3p poly (A)-ribonuclease complex which regulates poly (A) tail length (Figure 4A). However this interaction between a membrane-bound protein and an mRNA controller seems unlikely. On the other hand, *CDC50* is located 319 bp downstream to *CDC39* (also known as *NOT1*) which GINGER predicts as the true interactor of *PAN2*. Indeed, *CDC39* is a member of the CCR4–NOT complex, which is involved in many mRNA regulation processes (49–51). These two complexes provide alternative mechanisms of mRNA de-adenylation and are important for mRNA decay (52,53). In addition to the functional support for the interaction, DRYGIN also detects additional interactions between the PAN and CCR4–NOT complexes: both PAN members positively interact with *POP2* and *PAN3* is annotated to interact with CCR4.

In another case (Figure 4B), *RNH201*, the catalytic subunit of the Ribonuclease H enzyme appears (in DRYGIN) as a genetic interactor of *LAT1*, a component of the pyruvate dehydrogenase complex, which catalyzes the oxidative decarboxylation of pyruvate to acetyl-CoA in the mitochondria. GINGER corrects this annotation, and connects *RNH201* to the *UBC13* gene, encoding, together with *MMS2*, an E2 ubiquitin-conjugating enzyme involved in the error-free DNA post-replication repair pathway (54). During DNA replication, ribonucleotides are incorporated into DNA; these are usually excised by the RNase H complex; in the absence of this activity, the RNA:DNA hybrid is cut by topoisomerase I, leading to increased mutations, and, sometimes, to replication stalling (55). Lately, it has been suggested that under these circumstances, Srs2, an helicase of the post-replication repair pathway may act (by ubiquitinating target proteins) in order to allow lesion bypass and a resumption of DNA replication (56). Thus, a genetic interaction between RNase H and the Ubc13/Mms2 complex is consistent with alternative mechanisms of DNA repair/bypass. An additional interaction between *RNH202* and Ubc13, and a weaker interaction between *RNH201* and Mms2, further support our hypothesis.

Similarly, whereas a genetic interaction was described between Cdc73, a member of the Paf1 complex (which affects, among other roles, histone modifications (57)) and Dnm1, a Dynamin-related GTPase involved in mitochondrial organization (58), GINGER detects interactions between *CDC73* and *RTT109*, *DNM1*'s neighboring gene (Figure 4C). *RTT109* encodes a protein that, together with Vps75 forms an Histone acetyltransferase that acetylates histone H3 at position K56 (59). Mutations in both histone-modifying proteins reduce the cell's fitness. This prediction is further supported by genetic interactions between *RTT109* and *RTF1* and *LEO1*, two additional components of the Paf1 complex.

As a final example, *EAF3*, a component of both the Rpd3S histone deacetylase complex and the NuA4 acetyltransferase complex (60,61) is annotated as having a genetic interaction with Spo19, a meiosis-specific prospore protein. GINGER corrects this annotation by determining that *EAF3* mutations have reduced fitness when combined with *SPO19*'s neighbor, *TAF14*, a non-essential subunit of the TFIID, TFIIF, INO80, SWI/SNF and NuA3 chromatin remodeling complexes (62). Interestingly, a second NGE example can be seen in Figure 4D: YLR053c, an uncharacterized gene proposed to interact with *SIN3*, another component of the Rpd3S complex, is a neighbor of *IES3*, part of the INO80 complex.

Our systematic analysis has highlighted a large number of interactions that may suffer from NGE. However, in this study we have only investigated the non-essential portion of the genetic map and focused on negative interactions. It is likely that the additional interactions in the map are similarly biased and additional study of these interactions is needed. To systematically study these interactions requires algorithms that take into account the heterogeneous nature of these data sources but, more importantly, an experimental effort should be made to establish a large gold-standard to measure against. We hope that this paper will help to raise the interest in the community to inspect more carefully the annotations and conclusions that are drawn from deletion-based experiments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Edmond J. Safra Center for Bioinformatics at Tel Aviv University (to N.A.); The Israel Science Foundation (to M.K.); I-CORE Program of the Planning and Budgeting Committee [757/12 to R.S.]. Funding for open access charge: Tel Aviv University.
Conflict of interest statement. None declared.

REFERENCES

1. Pomper,S. and Burkholder,P.R. (1949) Studies on the Biochemical Genetics of Yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **35**, 456–464.
2. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al.* (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
3. Breslow,D.K., Cameron,D.M., Collins,S.R., Schuldiner,M., Stewart-Ornstein,J., Newman,H.W., Braun,S., Madhani,H.D., Krogan,N.J. and Weissman,J.S. (2008) A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat. Methods*, **5**, 711–718.
4. Ben-Aroya,S., Coombes,C., Kwok,T., O'Donnell,K.A., Boeke,J.D. and Hieter,P. (2008) Toward a comprehensive temperature-sensitive mutant repository of the essential genes of Saccharomyces cerevisiae. *Mol. Cell*, **30**, 248–258.
5. Li,Z., Vizeacoumar,F.J., Bahr,S., Li,J., Warringer,J., Vizeacoumar,F.S., Min,R., Vandersluis,B., Bellay,J., Devit,M. *et al.* (2011) Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat. Biotechnol.*, **29**, 361–367.
6. Tong,A.H.Y., Lesage,G., Bader,G.D., Ding,H., Xu,H., Xin,X., Young,J., Berriz,G.F., Brost,R.L., Chang,M. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
7. Schuldiner,M., Collins,S.R., Weissman,J.S. and Krogan,N.J. (2006) Quantitative genetic analysis in Saccharomyces cerevisiae using epistatic miniarray profiles (E-MAPs) and its application to chromatin functions. *Methods*, **40**, 344–352.
8. Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L.Y., Toufighi,K., Mostafavi,S. *et al.* (2010) The genetic landscape of a Cell. *Science*, **327**, 425–431.
9. Gibney,P.A., Lu,C., Caudy,A.A., Hess,D.C. and Botstein,D. (2013) Yeast metabolic and signaling genes are required for heat-shock survival and have little overlap with the heat-induced genes. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E4393–E4402.

10. Lee,A.Y., St Onge,R.P., Proctor,M.J., Wallace,I.M., Nile,A.H., Spagnuolo,P.A., Jitkova,Y., Gronda,M., Wu,Y., Kim,M.K. *et al.* (2014) Mapping the cellular response to small molecules using chemogenomic fitness signatures. *Science*, **344**, 208–211.
11. Franzosa,E.A., Albanèse,V., Frydman,J., Xia,Y. and McClellan,A.J. (2011) Heterozygous yeast deletion collection screens reveal essential targets of Hsp90. *PLoS One*, **6**, e28211.
12. Zhao,Y., Du,J., Zhao,G. and Jiang,L. (2013) Activation of calcineurin is mainly responsible for the calcium sensitivity of gene deletion mutations in the genome of budding yeast. *Genomics*, **101**, 49–56.
13. Gaytán,B.D., Loguinov,A.V., Peñate,X., Lerot,J.-M., Chávez,S., Denslow,N.D. and Vulpe,C.D. (2013) A genome-wide screen identifies yeast genes required for tolerance to technical toxaphene, an organochlorinated pesticide mixture. *PLoS One*, **8**, e81253.
14. Soifer,I. and Barkai,N. (2014) Systematic identification of cell size regulators in budding yeast. *Mol. Syst. Biol.*, **10**, 761.
15. Ben-Shitrit,T., Yosef,N., Shemesh,K., Sharan,R., Ruppin,E. and Kupiec,M. (2012) Systematic identification of gene annotation errors in the widely used yeast mutation collections. *Nat. Methods*, **9**, 373–378.
16. Hillenmeyer,M.E., Fung,E., Wildenhain,J., Pierce,S.E., Hoon,S., Lee,W., Proctor,M., St Onge,R.P., Tyers,M., Koller,D. *et al.* (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, **320**, 362–365.
17. Koh,J.L.Y., Ding,H., Costanzo,M., Baryshnikova,A., Toufighi,K., Bader,G.D., Myers,C.L., Andrews,B.J. and Boone,C. (2010) DRYGIN: a database of quantitative genetic interaction networks in yeast. *Nucleic Acids Res.*, **38**, D502–D507.
18. Frost,A., Elgort,M.G., Brandman,O., Ives,C., Collins,S.R., Miller-Vedam,L., Weibezahn,J., Hein,M.Y., Poser,I., Mann,M. *et al.* (2012) Functional repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. *Cell*, **149**, 1339–1352.
19. Ryan,C.J., Roguev,A., Patrick,K., Xu,J., Jahari,H., Tong,Z., Beltrao,P., Shales,M., Qu,H., Collins,S.R. *et al.* (2012) Hierarchical modularity and the evolution of genetic interactomes across species. *Mol. Cell*, **46**, 691–704.
20. McDowall,M.D., Harris,M.A., Lock,A., Rutherford,K., Staines,D.M., Bähler,J., Kersey,P.J., Oliver,S.G. and Wood,V. (2015) PomBase 2015: updates to the fission yeast database. *Nucleic Acids Res.*, **43**, D656–D661.
21. Chatr-Aryamontri,A., Breitkreutz,B.-J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2014) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
22. Wong,S.L., Zhang,L.V., Tong,A.H.Y., Li,Z., Goldberg,D.S., King,O.D., Lesage,G., Vidal,M., Andrews,B., Bussey,H. *et al.* (2004) Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 15682–15687.
23. Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.*, **23**, 561–566.
24. Youden,W.J. (1950) Index for rating diagnostic tests. *Cancer*, **3**, 32–35.
25. Vermont,J., Bosson,J.L., François,P., Robert,C., Rueff,A. and Demongeot,J. (1991) Strategies for graphical threshold determination. *Comput. Methods Programs Biomed.*, **35**, 141–150.
26. Yosef,N., Kupiec,M., Ruppin,E. and Sharan,R. (2009) A complex-centric view of protein network evolution. *Nucleic Acids Res.*, **37**, e88.
27. Bader,J.S., Chaudhuri,A., Rothberg,J.M. and Chant,J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.*, **22**, 78–85.
28. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
29. Yosef,N., Zalckvar,E., Rubinstein,A.D., Homilius,M., Atias,N., Vardi,L., Berman,I., Zur,H., Kimchi,A., Ruppin,E. *et al.* (2011) ANAT: a tool for constructing and analyzing functional protein networks. *Sci. Signal.*, **4**, pl1.
30. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
31. Cooper,P.S., Lipshultz,D., Matten,W.T., McGinnis,S.D., Pechous,S., Romiti,M.L., Tao,T., Valjavec-Gratian,M. and Sayers,E.W. (2010) Education resources of the National Center for Biotechnology Information. *Brief. Bioinform.*, **11**, 563–569.
32. Okamura,Y., Aoki,Y., Obayashi,T., Tadaka,S., Ito,S., Narise,T. and Kinoshita,K. (2015) COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.*, **43**, D82–D86.
33. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
34. Ye,P., Peyser,B.D., Pan,X., Boeke,J.D., Spencer,F.A. and Bader,J.S. (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol. Syst. Biol.*, **1**, 2005.0026.
35. Schuldiner,M., Collins,S.R., Thompson,N.J., Denic,V., Bhamidipati,A., Punna,T., Ihmels,J., Andrews,B., Boone,C., Greenblatt,J.F. *et al.* (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, **123**, 507–519.
36. Costanzo,M., Baryshnikova,A., Myers,C.L., Andrews,B. and Boone,C. (2011) Charting the genetic interaction map of a cell. *Curr. Opin. Biotechnol.*, **22**, 66–74.
37. Bandyopadhyay,S., Kelley,R., Krogan,N.J. and Ideker,T. (2008) Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput. Biol.*, **4**, e1000065.
38. Wood,V. (2006) Schizosaccharomyces pombe comparative genomics; from sequence to systems. In: Sunnerhagen,P and Piskur,J (eds). *Comparative Genomics, Topics in Current Genetics*. Springer, Berlin Heidelberg, pp. 233–285.
39. Dixon,S.J., Fedyshyn,Y., Koh,J.L.Y., Prasad,T.S.K., Chahwan,C., Chua,G., Toufighi,K., Baryshnikova,A., Hayles,J., Hoe,K.-L. *et al.* (2008) Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 16653–16658.
40. Tong,A.H., Evangelista,M., Parsons,A.B., Xu,H., Bader,G.D., Pagé,N., Robinson,M., Raghibizadeh,S., Hogue,C.W., Bussey,H. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.
41. Ulitsky,I., Shlomi,T., Kupiec,M. and Shamir,R. (2008) From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. *Mol. Syst. Biol.*, **4**, 209–209.
42. Rolland,T., Taşan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R. *et al.* (2014) A Proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.
43. Pu,S., Wong,J., Turner,B., Cho,E. and Wodak,S.J. (2009) Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.*, **37**, 825–831.
44. Collins,S.R., Miller,K.M., Maas,N.L., Roguev,A., Fillingham,J., Chu,C.S., Schuldiner,M., Gebbia,M., Recht,J., Shales,M. *et al.* (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **446**, 806–810.
45. Tuller,T., Rubinstein,U., Bar,D., Gurevitch,M., Ruppin,E. and Kupiec,M. (2009) Higher-order genomic organization of cellular functions in yeast. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **16**, 303–316.
46. Saito,K., Fujimura-Kamada,K., Furuta,N., Kato,U., Umeda,M. and Tanaka,K. (2004) Cdc50p, a protein required for polarized growth, associates with the Drs2p P-type ATPase implicated in phospholipid translocation in Saccharomyces cerevisiae. *Mol. Biol. Cell*, **15**, 3418–3432.
47. Misu,K., Fujimura-Kamada,K., Ueda,T., Nakano,A., Katoh,H. and Tanaka,K. (2003) Cdc50p, a conserved endosomal membrane protein, controls polarized growth in Saccharomyces cerevisiae. *Mol. Biol. Cell*, **14**, 730–747.
48. Azouaoui,H., Montigny,C., Ash,M.-R., Fijalkowski,F., Jacquot,A., Grønberg,C., López-Marqués,R.L., Palmgren,M.G., Garrigos,M., le Maire,M. *et al.* (2014) A high-yield co-expression system for the purification of an intact Drs2p-Cdc50p lipid flippase complex, critically dependent on and stabilized by phosphatidylinositol-4-phosphate. *PLoS One*, **9**, e112176.
49. Collart,M.A. (2003) Global control of gene expression in yeast by the Ccr4-Not complex. *Gene*, **313**, 1–16.

50. Collart,M.A. and Panasenko,O.O. (2012) The Ccr4–not complex. *Gene*, **492**, 42–53.
51. Denis,C.L. and Chen,J. (2003) The CCR4-NOT complex plays diverse roles in mRNA metabolism. *Prog. Nucleic Acid Res. Mol. Biol.*, **73**, 221–250.
52. Wolf,J., Valkov,E., Allen,M.D., Meineke,B., Gordiyenko,Y., McLaughlin,S.H., Olsen,T.M., Robinson,C.V., Bycroft,M., Stewart,M. *et al.* (2014) Structural basis for Pan3 binding to Pan2 and its function in mRNA recruitment and deadenylation. *EMBO J.*, **33**, 1514–1526.
53. Basquin,J., Roudko,V.V., Rode,M., Basquin,C., Séraphin,B. and Conti,E. (2012) Architecture of the nuclease module of the yeast Ccr4-not complex: the Not1-Caf1-Ccr4 interaction. *Mol. Cell*, **48**, 207–218.
54. Liefshitz,B., Steinlauf,R., Friedl,A., Eckardt-Schupp,F. and Kupiec,M. (1998) Genetic interactions between mutants of the 'error-prone' repair group of Saccharomyces cerevisiae and their effect on recombination and mutagenesis. *Mutat. Res.*, **407**, 135–145.
55. Williams,J.S., Smith,D.J., Marjavaara,L., Lujan,S.A., Chabes,A. and Kunkel,T.A. (2013) Topoisomerase 1-mediated removal of ribonucleotides from nascent leading-strand DNA. *Mol. Cell*, **49**, 1010–1015.
56. Potenski,C.J., Niu,H., Sung,P. and Klein,H.L. (2014) Avoidance of ribonucleotide-induced mutations by RNase H2 and Srs2-Exo1 mechanisms. *Nature*, **511**, 251–254.
57. Lenstra,T.L., Benschop,J.J., Kim,T., Schulze,J.M., Brabers,N.A.C.H., Margaritis,T., van de Pasch,L.A.L., van Heesch,S.A.A.C., Brok,M.O., Groot Koerkamp,M.J.A. *et al.* (2011) The specificity and topology of chromatin interaction pathways in yeast. *Mol. Cell*, **42**, 536–549.
58. Gammie,A.E., Kurihara,L.J., Vallee,R.B. and Rose,M.D. (1995) DNM1, a dynamin-related gene, participates in endosomal trafficking in yeast. *J. Cell Biol.*, **130**, 553–566.
59. Schneider,J., Bajwa,P., Johnson,F.C., Bhaumik,S.R. and Shilatifard,A. (2006) Rtt109 is required for proper H3K56 acetylation: a chromatin mark associated with the elongating RNA polymerase II. *J. Biol. Chem.*, **281**, 37270–37274.
60. Carrozza,M.J., Li,B., Florens,L., Suganuma,T., Swanson,S.K., Lee,K.K., Shia,W.-J., Anderson,S., Yates,J., Washburn,M.P. *et al.* (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, **123**, 581–592.
61. Eisen,A., Utley,R.T., Nourani,A., Allard,S., Schmidt,P., Lane,W.S., Lucchesi,J.C. and Cote,J. (2001) The yeast NuA4 and Drosophila MSL complexes contain homologous subunits important for transcription regulation. *J. Biol. Chem.*, **276**, 3484–3491.
62. Kabani,M., Michot,K., Boschiero,C. and Werner,M. (2005) Anc1 interacts with the catalytic subunits of the general transcription factors TFIID and TFIIF, the chromatin remodeling complexes RSC and INO80, and the histone acetyltransferase complex NuA3. *Biochem. Biophys. Res. Commun.*, **332**, 398–403.