

# Systematic identification of gene annotation errors in the widely used yeast mutation collections

Taly Ben-Shitrit<sup>1,6</sup>, Nir Yosef<sup>2,3,6</sup>, Keren Shemesh<sup>1</sup>, Roded Sharan<sup>4</sup>, Eytan Ruppin<sup>4,5</sup> & Martin Kupiec<sup>1</sup>

**The baker's yeast mutation collections are extensively used genetic resources that are the basis for many genome-wide screens and new technologies. Anecdotal evidence has previously pointed to the putative existence of a neighboring gene effect (NGE) in these collections. NGE occurs when the phenotype of a strain carrying a particular perturbed gene is due to the lack of proper function of its adjacent gene. Here we performed a large-scale study of NGEs, presenting a network-based algorithm for detecting NGEs and validating software predictions using complementation experiments. We applied our approach to four datasets uncovering a similar magnitude of NGE in each (7–15%). These results have important consequences for systems biology, as the mutation collections are extensively used in almost every aspect of the field, from genetic network analysis to functional gene annotation.**

The yeast *Saccharomyces cerevisiae* has long served as a powerful genetic system, more recently because of the availability of systematic genome-wide mutant collections. These collections have been used as a basic resource to screen for specific phenotypes<sup>1</sup>, study gene expression profiles under genetic perturbation<sup>2</sup> and discover genetic interactions<sup>3</sup>, among other uses. The basic yeast deletion library<sup>4</sup> includes ~4,700 haploid strains, each carrying a different deleted nonessential gene. Each gene in the collection has been deleted by replacing the open reading frame (ORF) with a selectable marker<sup>4</sup>, a process that should not influence neighboring genes. However, anecdotal observations and some published studies<sup>1,5–7</sup> indicate that the phenotype of a particular strain can actually be due to the effect that the deletion has on an adjacent gene. The extent of this problem, the NGE, has not been systematically explored to date to our knowledge.

Identifying the causal gene in a deletion analysis is of paramount importance. Yeast deletion libraries have been a prime workhorse for our understanding of yeast gene function (and other organisms by extrapolation), and wrong identification leads to erroneous gene annotations<sup>8</sup>. Furthermore, common global analyses such as protein-protein interaction network inferences<sup>9</sup> may be biased by NGEs, as misidentified genes distort

results and incorrectly link unrelated genes. Finally, as most current studies of genetic interactions<sup>3</sup> or the effects of particular mutants on global transcription (measured by DNA microarray hybridization or RNA sequencing<sup>10</sup>) are usually carried out in yeast deletion strains, incorrectly attributed effects are likely to have a strong influence on our understanding of genome function. Hypomorphic mutations in essential genes, including the genome-wide 'decreased abundance by mRNA perturbation' (DAmP) and temperature-sensitive collections<sup>11–13</sup> may also be susceptible to NGEs.

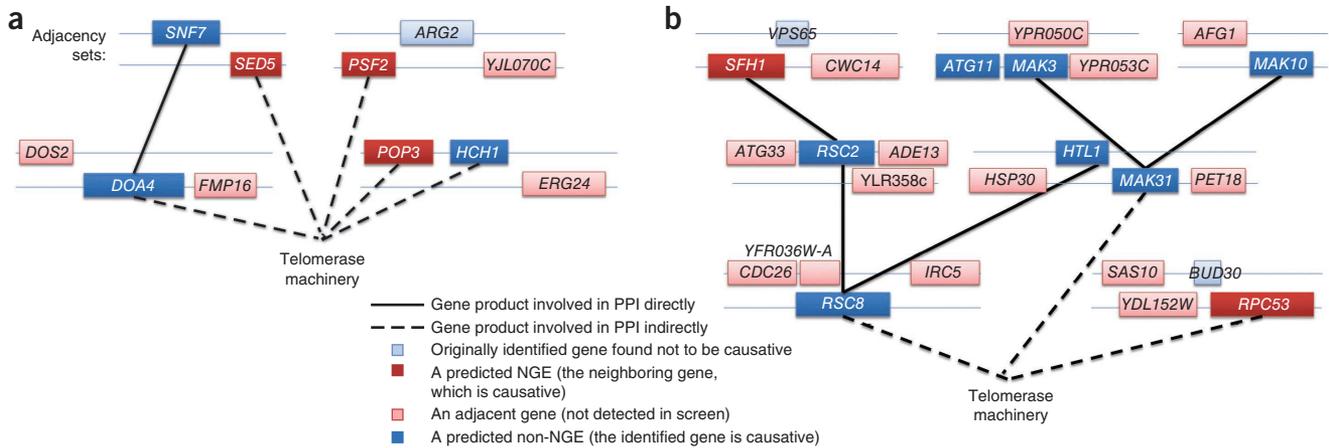
A difficulty with the NGE is that it is not straightforward to detect. For example, the ORF *YDL162C* has been identified in a screen for genes that affect genome stability<sup>6</sup>; deletion of this ORF was shown to affect the expression of its neighbor, *CDC9*, which encodes DNA ligase. One would expect the two genes to always appear together in any genetic screen. Remarkably, this is not always true, as most screens are not carried out in an exhaustive fashion. In addition, the causative gene may be essential or absent from the yeast collection for technical reasons. In the *CDC9* example, the gene identified in the screen is a dubious ORF, whereas its neighbor is a well-studied gene with a known function related to the phenotype that was screened for. In cases of unknown or ambiguous function, however, NGEs may remain undetected.

Here we show that NGE is a widespread phenomenon, affecting about 10% of all genes, and is thus likely to substantially distort our current perception of gene function and interactions. We present a network-based algorithm for identifying NGE cases and estimate its extent based on a number of datasets, and we validate predictions using complementation experiments.

## RESULTS

We reasoned that an automated method to pinpoint the true causative genes in a given genetic screen must rely on data that do not depend on the yeast mutant libraries and thus are not also prone to the NGE. We based our approach on protein-protein interaction (PPI) information, which has been shown by us and others to be a valuable resource for uncovering cellular phenotypes<sup>14–19</sup>.

<sup>1</sup>Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel. <sup>2</sup>Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. <sup>3</sup>Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. <sup>4</sup>School of Computer Science, Tel Aviv University, Ramat Aviv, Israel. <sup>5</sup>Sackler School of Medicine, Tel Aviv University, Ramat Aviv, Israel. <sup>6</sup>These authors contributed equally to this work. Correspondence should be addressed to M.K. (martin@post.tau.ac.il).



**Figure 1** | Examples of the NIRVANA method. (a,b) The genomic location of genes involved in TLM, shown in their adjacency sets. ORFs are ordered from right to left in top strands and left to right in bottom strands. NIRVANA chooses, from each set, the gene whose product is most likely to be linked to the telomere machinery (the anchor point). In **a**, *ARG2* was identified in the screen, but NIRVANA predicted *PSF2* as the causative gene. Neighboring genes may be chosen as TLM genes (for example, *SNF7* (identified in the screen) and *SED5* (not identified); *HCH1* (identified) and *POP3* (not identified)). In **b**, *SFH1* and *RPC53*, not found in the TLM screen, were predicted to be the real effectors (instead of the identified *VPS65* and *BUD30*). *HTL1*, *MAK31*, *RSC2* and *RSC8*, all identified in the screen, were predicted to be true effectors. Neighboring *HTL1* and *MAK31* were both chosen as TLM genes by the algorithm. Both genes encode proteins that act in a complex whose other subunits were also found in the screens: *HTL1* is connected to subunits of the RSC chromatin-remodeling complex; *MAK31* is connected to components of the NatC complex.

For each gene perturbed in a particular yeast strain, we define the genes that could be affected by its deletion, its ‘adjacency set’, as those genes totally or partially included in an interval starting 600 base pairs (bp) upstream of the gene and ending 600 bp downstream of it (Online Methods and Fig. 1). These sets usually contain the deleted gene and two flanking ORFs, although a few contain three, one or no ORFs. We devised a ‘reverse-engineering’ algorithm (NGE inference via a network-based approach; NIRVANA), to assign the causative gene(s) in each set by seeking the most probable PPI network that underlies the phenotype in question<sup>16</sup>.

Our algorithm requires that one or more proteins directly responsible for the phenotype are known; these serve as ‘end nodes’ in the network (the ‘anchor set’). From every adjacency set, the algorithm chooses one or more genes that are the most likely to be connected to the anchor set via the PPI network. The information in this network is expected to flow from the different telomere-length maintenance (TLM) proteins to the anchor. The rationale behind NIRVANA is that the true causative gene(s) in any set will contribute to a ‘better’ overall PPI subnetwork, accounting for the end-node phenotype. Two principles of a better subnetwork are used to guide node selection: a selected node should be close, or have the fewest possible steps, to the anchor set (local factor) and it should be close to other selected nodes (global factor). In addition, we positively weighted proximity to the deleted ORF using a Bayesian method (physical distance-based factor).

The NIRVANA software and user manual are available as **Supplementary Software**, and updates will be available at <http://www.cs.tau.ac.il/~bnet/NIRVANA/>. We explored the scope of NGEs in datasets from two phenotypic screens: the TLM system<sup>15,16,20,21</sup> and the response to the drug rapamycin<sup>22,23</sup>.

### Manual analysis of NGE in the TLM dataset

Telomere length is maintained by a complex balance between positive and negative signals. Previous high-throughput

measurements of telomere length in mutants with changes in non-essential<sup>20,21</sup> and essential genes<sup>15</sup> have resulted in a collection of 385 genes (TLM set; **Supplementary Table 1**). The corresponding adjacency sets contained 787 genes with an average of 2.17 members per set (note that some of the adjacency sets may overlap; **Supplementary Table 1** and Fig. 1).

As a first estimate of the prevalence of NGE in the TLM set, we manually examined each adjacency set. We used three literature-based decision rules to decide, where possible, which genes (one or more) are likely to affect telomere length: (i) proteins annotated as telomere-interacting, (ii) proteins shown to be involved in TLM in small-scale studies or (iii) proteins that participate in a complex for which more than half the subunits have been found in TLM screens (**Supplementary Table 2**).

We considered as NGEs all the cases in which one or more genes in an adjacency set satisfied at least one rule (120 adjacency sets; **Supplementary Tables 1–3**). Remarkably, in 25 of these sets (20.8%) the only gene marked as causative was actually the neighbor of the original gene discovered in the TLM screen. In four additional cases (3.3%), the rules resulted in selection of both the original gene and its neighbor. We observed this non-negligible amount of suspected NGEs (29/120 NGEs or 24.1%) irrespective of the decision rule (rule i, 22.2%; rule ii, 21.2%; and rule iii, 25.7%).

### Analysis of NGE in the TLM data by NIRVANA

Our heuristic manual curation procedure can approximate the prevalence of NGE in the TLM set but depends on substantial information from the literature and, hence, has limited applicability for other datasets. Seeking a more general and unbiased procedure, we applied NIRVANA to infer the most probable PPI network that connects the adjacency sets to the telomerase machinery. As the anchor set we considered ten telomere-binding proteins including telomerase subunits and telomerase-interacting proteins (accessory factors and exonucleases)<sup>16</sup>.

**Table 1** | NGE and Non-NGE predictions in the TLM dataset

Original gene	NIRVANA prediction	Manual assessment	Confirmed by complementation	Algorithm success rate
<i>APN1</i>	<i>APN1</i> and <i>RAD27</i>	<i>RAD27</i>	<i>RAD27</i>	Confirmed NGE 5/7 = 71.4%
<i>FYV12</i>	<i>RPS30B</i> and <i>SER1</i>	–	<i>RPS30B</i>	
<i>HUR1</i>	<i>PMR1</i> and <i>SUA5</i>	<i>SUA5</i>	<i>PMR1</i>	
<i>LST7</i>	<i>RSC1</i>	<i>RSC1</i>	<i>LST7</i> and <i>RSC1</i>	
<i>MRM2</i>	<i>MRM2</i>	–	<i>SEC27</i>	
<i>RPS17A</i>	<i>RPS17A</i>	–	<i>NSE5</i>	
<i>RRP8</i>	<i>RRP8</i> and <i>STN1</i>	<i>STN1</i>	<i>STN1</i>	
<i>ASC1</i>	<i>ASC1</i>	<i>SPC24</i>	<i>ASC1</i>	Confirmed non-NGE 10/11 = 90.9%
<i>BUD23</i>	<i>BUD23</i>	–	<i>BUD23</i>	
<i>ELG1</i>	<i>ELG1</i>	<i>ELG1</i>	<i>ELG1</i>	
<i>GCV3</i>	<i>PTA1</i>	–	<i>GCV3</i>	
<i>HIT1</i>	<i>HIT1</i>	–	<i>HIT1</i>	
<i>LDB7</i>	<i>LDB7</i>	<i>LDB7</i>	<i>LDB7</i>	
<i>MAK31</i>	<i>MAK31</i>	<i>MAK31</i>	<i>MAK31</i>	
<i>MET7</i>	<i>MET7</i>	–	<i>MET7</i>	
<i>SIW14</i>	<i>SIW14</i>	–	<i>SIW14</i>	
<i>SSN8</i>	<i>SSN8</i>	<i>SSN8</i>	<i>SSN8</i>	
<i>YOR1</i>	<i>YOR1</i>	–	<i>YOR1</i>	

–, not predicted by the manual analysis.

We constructed the network model such that at least one representative gene product was chosen from each adjacency set and connects to (directly or via contact with other proteins) at least one member of the anchor set (Fig. 1). Genes encoding proteins included in the network were predicted to be true effectors (Supplementary Table 1). In sets with more than one suitable candidate (where several genes contribute similarly to the overall likelihood of the network), the algorithm can choose more than one node (Fig. 1).

Benchmarking the algorithm against the manual annotation of NGEs listed in Supplementary Table 2, we achieved an accuracy of 84.4% in predicting the selected gene (versus a random expectation of  $55.9\% \pm 5.15\%$  ( $\pm$ s.d.) accuracy; Online Methods). Overall, our algorithm estimated that 11.6% of the TLM dataset suffers from NGE (37 out of 319 cases; Supplementary Table 1 and Online Methods). In seven (17.9%) of the NGE cases, the predicted gene had been identified in one of the original TLM screens (thus, only one of two adjacent TLM genes is the causative one). In 17 additional NGE cases (45.9%) the predicted gene was essential and was not detected in our screen of the DAMP collection<sup>15</sup>. This is not unexpected as hypomorphic alleles do not always have phenotypes. In the remaining cases, the causative gene could either have been absent from the deletion collection or did not give a clear TLM phenotype. Thus, the TLM set has a non-negligible number of cases in which the gene adjacent to a deleted gene is the causative one.

### Experimental validation of NGE in the TLM dataset

To validate our predictions we carried out complementation tests. Each deletion strain was transformed with single-copy plasmids carrying either the deleted gene, its neighbor, or no gene at all (vector). After propagation of the transformants for more than 150 generations, their telomere lengths were compared. We randomly chose and tested 9 cases from our literature-based decision rules (Table 1). We observe a very high concurrence between the experimental results and the rules' predictions

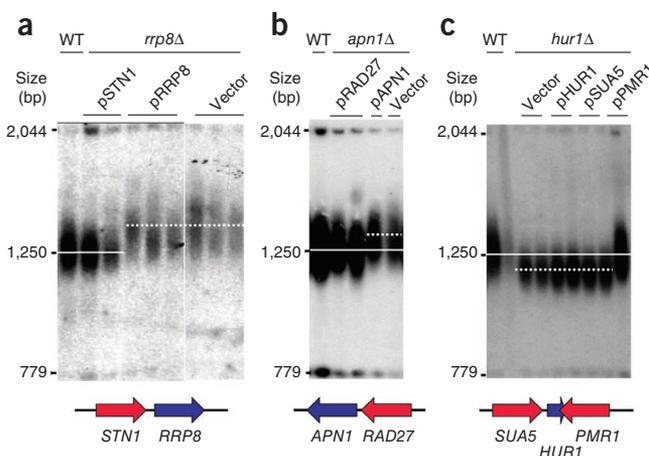
(7/9). In the two cases in which there was disagreement, NIRVANA's prediction coincided with the experimental result. For nine additional genes that cannot be evaluated using information from the literature NIRVANA achieved a match of 66.7% (6/9). Overall, our complementation experiments matched the algorithm's predictions in 83.4% of the cases (15/18) (Table 1). Three examples of NGEs are shown in Figure 2; in each case, the telomere length phenotype could be complemented with a plasmid bearing the neighboring gene, and not with one bearing the deleted gene.

Using complementation tests we also confirmed 11 cases in which the deleted gene is indeed responsible for the observed phenotype (that is, there was no NGE); 10 were in agreement with the algorithm's predictions (Table 1 and Fig. 3). The few instances in which the algorithm mispredicted the causative gene were likely due to poor representation of one of the genes of the adjacency set in the original PPI network.

### Analysis of NGE in the rapamycin response dataset

Rapamycin is an immunosuppressive and anticancer drug that inhibits the conserved serine threonine kinase target of rapamycin (TOR). Yeast contains two TOR homologs, TOR1 and TOR2 (ref. 24), which form two distinct complexes, TORC1 and TORC2. TORC1, which may contain TOR1 or TOR2, controls cell growth and is inhibited by rapamycin<sup>24</sup>. In contrast, TORC2, believed to be rapamycin-insensitive, is composed solely of TOR2 and has a role in cytoskeleton organization<sup>24</sup>. Two genome-wide screens for altered response to rapamycin have been carried out using the yeast deletion collections<sup>22,23</sup>, in both of which 347 genes had been identified (Supplementary Table 4). The corresponding adjacency sets contain 782 genes with an average of 2.43 members per set (Supplementary Table 4).

Using the heuristic decision rules described above to annotate the altered response to rapamycin adjacency sets again predicted



**Figure 2** | Examples of NGEs predicted by NIRVANA, which we confirmed by complementation tests. (a–c) Strains from the deletion library were transformed with plasmids carrying the deleted strain, its neighbor(s) or no gene at all (vector). Genes originally identified in the screens are colored blue and neighboring genes are in red. Telomeric Southern blots show the terminal chromosomal XhoI fragment and two size markers. Solid and dashed white lines mark the telomere size of wild-type (WT) and deletion strains, respectively. Deletion of *RRP8* altered the activity of adjacent *STN1* (a). Deletion of *APN1* affected neighboring *RAD27* (b). Deletion of *HUR1* affected neighboring *PMR1* (c).

**Table 2** | Predictions in the response to rapamycin dataset

Original gene	NIRVANA prediction	Manual assessment	Confirmed by complementation	Algorithm success rate
<i>AIM26</i>	<i>UGP1</i>	–	<i>UGP1</i>	Confirmed
<i>FYV5</i>	<i>KRR1</i>	–	<i>FYV5</i> and <i>KRR1</i>	NGE 3/3
<i>YEL045C</i>	<i>GLY1</i>	–	<i>GLY1</i>	= 100%
<i>SER1</i>	<i>SER1</i>	–	<i>SER1</i>	
<i>LST7</i>	<i>RSC1</i>	–	<i>LST7</i>	
<i>YGL211W</i>	<i>YGL211W</i> and <i>VAM7</i>	<i>YGL211W</i> and <i>VAM7</i>	<i>YGL211W</i>	Confirmed
<i>BUD23</i>	<i>BUD23</i>	–	<i>BUD23</i>	non-NGE 4/6
<i>YPR084W</i>	<i>YPR084W</i>	–	<i>YPR084W</i>	= 66.6%
<i>OPI9</i>	<i>VRP1</i>	–	<i>OPI9</i>	

–, not predicted by the manual analysis.

a non-negligible NGE (19/88 = 21.5%; **Supplementary Table 5**). For a more systematic identification of NGEs, we applied NIRVANA using TOR1, TOR2 and FPR1 (the direct target of rapamycin) as anchor points. Comparing the algorithm to the manual annotations, we obtained an overall accuracy of 80.85% in predicting the gene or genes annotated as causal (versus a random expectation of  $51.4\% \pm 7.0\%$ ; Online Methods). Similarly to the data obtained with the TLM dataset, NIRVANA estimated the amount of NGEs in the response to rapamycin data to be 10.0% (31/310 of adjacency sets; **Supplementary Table 4**).

### Validation of NGE in the rapamycin dataset

We experimentally tested seven NGE cases in which a neighboring gene predicted by NIRVANA did not come up in any of the rapamycin screens. In five cases, mutations in the predicted causative genes conferred rapamycin sensitivity (**Fig. 4a**). For example, deletion of the uncharacterized gene *HHY1* confers sensitivity to rapamycin<sup>22</sup>. Our algorithm predicted that deletion of *HHY1* affects the adjacent ORE, *PCM1* (*YEL058W*), encoding an essential N-acetylglucosamine-phosphate mutase; indeed, a hypomorphic *pcm1* DAMP allele had severe rapamycin sensitivity (**Fig. 4a**).

To validate our predictions, we conducted complementation experiments as above, testing the transformants on plates containing rapamycin (**Table 2** and **Fig. 4b**). Algorithmic predictions matched experimental results in seven of nine cases (77%).

### Analysis of NGEs in additional datasets

To test the generality of our method, we applied NIRVANA to two additional datasets. The first was a set of 191 mutants that affect sensitivity to overexpression of the mutant topoisomerase-1 allele *top1-T722A* (ref. 25). Using the Top1 protein as an anchor point, NIRVANA detected 30 cases of NGEs (15.7%). The second was a set of 138 genes that when deleted show hypersensitivity to the anticancer drug 5-fluorouracil<sup>26</sup>. Currently, the only known direct targets of this drug are the exosome complex and the Cdc21 thymidylate synthase. Using this partial list as anchors, the NIRVANA algorithm predicted 7.2% NGE in this screen (10 cases). This number is expected to be larger when additional anchors are introduced.

### Potential mechanisms for NGE

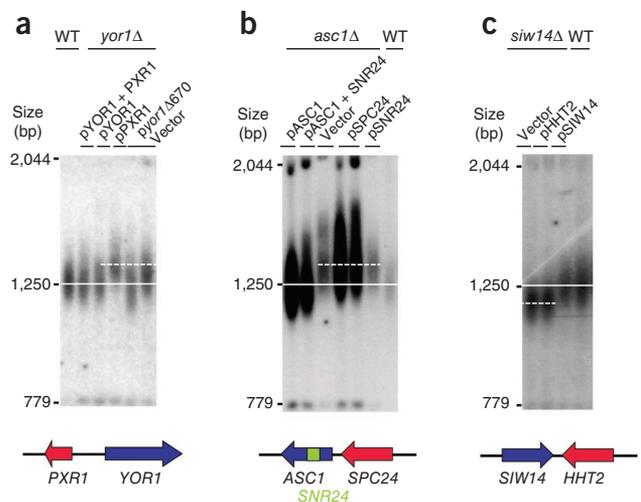
To probe the possible mechanisms by which NGE may occur, we explored the characteristics of the combined NGE list from the

TLM and response to rapamycin datasets (a total of 73 pairs of a deleted gene and an affected gene; **Supplementary Table 6**). First, we examined the relative orientation of the gene pairs, considering three possible options: converging, diverging and tandem (**Supplementary Table 6**). We found no significant (chi-squared  $P = 0.97$ ) preferences among NGE cases compared to the entire set of adjacent genes in the genome.

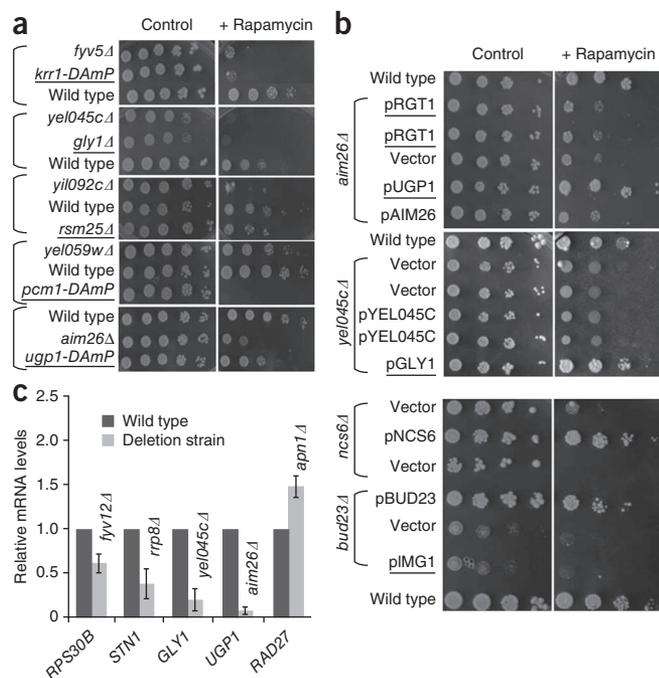
To test whether the effects occur at the level of transcription, we quantified the adjacent transcript in the deletion and wild-type strain in our validated NGE cases. Quantitative reverse-transcriptase PCR showed a substantial change in the steady-state mRNA amount for five of the nine cases tested (**Supplementary Table 7** and **Fig. 4c**; in the other four cases the change was mild). For example, there was a reduction of about 40% of the mRNA level of *STN1* in the *rrp8Δ* strain compared with the wild type (**Fig. 4c**). Thus, in a majority of the cases the deletion affected the steady-state amount of the neighboring gene.

Next, we examined whether the transcripts of the interacting pairs overlapped<sup>27,28</sup>. We found overlap in 44 out of 73 NGE cases, suggesting that the deletion of one gene may affect either the production or stability of its neighboring transcript. Other possible explanations for NGEs include the existence of short unannotated transcripts (for example, ref. 28) or the influence of the strong promoter present in the *KanMX* cassette commonly used to create the deletion collections<sup>4</sup> on the neighboring gene.

In both the TLM and rapamycin datasets there were cases in which two neighboring genes were predicted to be true effectors. In these cases, it is possible that both genes regulate each other through noncoding antisense sequences<sup>28</sup>. In support of this idea, only the presence of both the original gene and its neighbor succeeded in fully complementing the phenotype in two cases (*LST7* and its neighbor *RSC1* in the TLM dataset and *FYV5* and its neighbor *KRR1* in the rapamycin dataset).



**Figure 3** | Examples of predicted non-NGE confirmed by complementation tests. (**a–c**) Strains from the deletion library were transformed with plasmids carrying the deleted strain, its neighbor(s) or no gene at all (vector). Genes originally identified in the screens are colored in blue and neighboring genes are in red. Solid and dashed white lines mark the telomere size of wild-type (WT) and deletion strains, respectively. Wild-type *YOR1*, but not a mutant allele or its neighbor *PXR1* restored wild-type telomere length (**a**). *ASC1*, but neither the neighboring *SPC24* nor *SNR24* (located inside an *ASC1* intron and encoding a small nucleolar RNA), complemented the telomere length of an *asc1* strain (**b**). *SIW14* was a true TLM gene (**c**).



**Figure 4** | Validation of predicted NGE cases in the altered response to rapamycin network. **(a)** Tenfold serial dilutions of yeast haploid strains containing the DAmP allele or deletion of indicated genes and a wild-type (WT) strain were plated onto medium with or without 8 nM rapamycin. Predicted NGE cases (underlined) that had sensitivity to rapamycin similarly to the deletion mutant initially identified. **(b)** Tenfold serial dilutions of yeast mutant strains with plasmids carrying the deleted gene, its neighbor(s) or no gene at all (vector) plated onto medium with or without 8 nM rapamycin. Neighbor genes are underlined. **(c)** Comparison of mRNA levels of the true causative gene in its adjacent (original) gene deletion strain. Expression of the indicated verified effector genes in the wild type and the original (non-causative) deletion was examined by real-time PCR. Values shown are relative mRNA levels after normalization to the *ACT1* mRNA level and are averages  $\pm$  s.d. of at least three determinations. Only significant changes are shown ( $P < 0.04$ ).

## DISCUSSION

NGEs in yeast mutant collection data may introduce pervasive misannotations and misinterpretations. If the deleted gene is a known component of a certain pathway, the process studied is linked to that pathway. In addition, its neighbor (the true causative gene) is undetected. Thus, even a few NGEs likely affect our understanding of the ways in which genes, proteins and pathways are linked. An NGE rate of 1 in 10 (roughly what we observed in our analysis) may lead to substantial changes in the topology of the global genetic interaction map<sup>3</sup> (as currently connected nodes should disconnect and new connections should be created).

NIRVANA pinpointed the correct causative genes, but the problem still exists that more than one gene is affected in a single deletion strain. For proven cases of NGEs, possible solutions are to create new alleles that are simple insertions to disrupt the ORF, smaller deletions (that are likely to be less disruptive of neighbors) or, best of all, point mutations.

NIRVANA uses a large database of PPIs to uncover NGEs in any cellular system where there is knowledge about the network's 'anchor', that is, the protein(s) that are direct mediators of the phenotype studied. It will contribute to the future analysis of additional cellular subsystems as more anchors are established.

But algorithms should also be developed to analyze cellular networks devoid of known anchor points. The acknowledgment of the wide scope of the NGE problem, together with the widespread application of NIRVANA and its successors, will be important for reanalyzing current large-scale datasets toward more accurate determination of gene interactions and gene annotations<sup>29</sup>.

## METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturemethods/>.

Note: Supplementary information is available on the Nature Methods website.

## ACKNOWLEDGMENTS

We thank current and past members of the Sharan, Rupp and Kupiec laboratories for helpful discussions, and D. Shore, R. Rolfes, B. Cairns, A. Johnson, S. Harashima, A.S. Bystrom, S. Moye-Rowley, M. Johnston, M.A. Clayton, G.H. Braus, B. Polevoda, R. Movva, E. Alani, J.P. Gélugne, V. Measday, D. Ramotar, A. Munn, E. Miller, B.C. Laurent, C. Brenner, M. Polymenis, J. Gerst, L. Aragon, D. Spatt, J. Boeke, G. Brown, C. Burd, M. Tamas, D.H. Wolf, Y. Hannun, Y. Ohsumi, Z. Ciesla, M. Choder, E. Bi, M. Bard, R. Schaffrath, H.U. Mosch, A. Amon, R. Parker, Y. Saeki, J. Kim, H.O. Park, E. Etsuchi, H. Nakatogana, B. Daignan-Fornier for providing plasmids. This work was supported by grants from the Israeli Ministry of Science and Technology and the Israel Cancer Foundation to M.K., and by a grant from the James McDonnell Fund to M.K., R.S. and E.R.; E.R. and R.S. were also supported by a Bikura grant from the Israel Science Foundation.

## AUTHOR CONTRIBUTIONS

T.B.-S., N.Y., R.S., E.R. and M.K. conceived and designed the experiments and wrote the paper. T.B.-S. and K.S. carried out the wet laboratory experiments. N.Y. analyzed the data obtained. R.S. and E.R. contributed equally to this project.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturemethods/>.

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

- Hillenmeyer, M.E. *et al.* The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362–365 (2008).
- Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
- Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
- Winzler, E.A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
- Tong, A.H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
- Pan, X. *et al.* A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* **124**, 1069–1081 (2006).
- Addinall, S.G. *et al.* A genomewide suppressor and enhancer analysis of *cdc13-1* reveals varied cellular processes influencing telomere capping in *Saccharomyces cerevisiae*. *Genetics* **180**, 2251–2266 (2008).
- Tian, W. *et al.* Combining guilt-by-association and guilt-by-profiling to predict *Saccharomyces cerevisiae* gene function. *Genome Biol.* **9** (suppl. 1), S7 (2008).
- Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566 (2005).
- van Wageningen, S. *et al.* Functional overlap and regulatory links shape genetic interactions between signaling pathways. *Cell* **143**, 991–1004 (2010).
- Breslow, D.K. *et al.* A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat. Methods* **5**, 711–718 (2008).
- Ben-Aroya, S. *et al.* Toward a comprehensive temperature-sensitive mutant repository of the essential genes of *Saccharomyces cerevisiae*. *Mol. Cell* **30**, 248–258 (2008).

13. Li, Z. *et al.* Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat. Biotechnol.* **29**, 361–367 (2011).
14. Shachar, R., Ungar, L., Kupiec, M., Ruppin, E. & Sharan, R. A systems-level approach to mapping the telomere length maintenance gene circuitry. *Mol. Syst. Biol.* **4**, 172 (2008).
15. Ungar, L. *et al.* A genome-wide screen for essential yeast genes that affect telomere length maintenance. *Nucleic Acids Res.* **37**, 3840–3849 (2009).
16. Yosef, N. *et al.* Toward accurate reconstruction of functional protein networks. *Mol. Syst. Biol.* **5**, 248 (2009).
17. Dittrich, M., Klau, G., Rosenwald, A., Dandekar, T. & Müller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* **24**, i223–i231 (2008).
18. Huang, S.S. & Fraenkel, E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci. Signal.* **2**, ra40 (2009).
19. Yosef, N. *et al.* ANAT—a software tool for reconstructing and analyzing functional networks of proteins. *Sci. Signal.* **4**, l1 (2011).
20. Askree, S.H. *et al.* A genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length. *Proc. Natl. Acad. Sci. USA* **101**, 8658–8663 (2004).
21. Gatzbonton, T. *et al.* Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet.* **2**, e35 (2006).
22. Parsons, A.B. *et al.* Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat. Biotechnol.* **22**, 62–69 (2004).
23. Chan, T.F., Carvalho, J., Riles, L. & Zheng, X.F. A chemical genomics approach toward understanding the global functions of the target of rapamycin protein (TOR). *Proc. Natl. Acad. Sci. USA* **97**, 13227–13232 (2000).
24. Crespo, J.L. & Hall, M.N. Elucidating TOR signaling and rapamycin action: lessons from *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **66**, 579–591 (2002).
25. Reid, R.J. *et al.* Selective ploidy ablation, a high-throughput plasmid transfer protocol, identifies new genes affecting topoisomerase I-induced DNA damage. *Genome Res.* **21**, 477–486 (2011).
26. Gustavsson, M. & Ronne, H. Evidence that tRNA modifying enzymes are important in vivo targets for 5-fluorouracil in yeast. *RNA* **14**, 666–674 (2008).
27. Tuller, T. *et al.* Higher-order genomic organization of cellular functions in yeast. *J. Comput. Biol.* **16**, 303–316 (2009).
28. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
29. Snyder, M. & Gallagher, J.E. Systems biology from a yeast omics perspective. *FEBS Lett.* **583**, 3895–3899 (2009).

## ONLINE METHODS

**Software.** NIRVANA is available as **Supplementary Software** and online at <http://www.cs.tau.ac.il/~bnet/NIRVANA/>.

**Adjacency sets.** We reasoned that NGE is most likely caused by functional interference with the promoter or the untranslated regions of the adjacent genes. According to recent observations, more than 50% of the intergenic regions in *S. cerevisiae* are smaller than 500 bp<sup>27</sup>. Thus, in defining the adjacency set, we established a threshold distance of 600 bp. Because the closer (in physical distance terms) a gene is to its neighbor, the higher are the chances that its deletion will create an NGE event, we used a simple Bayesian approach to assign a prior probability for a gene to be causal as a function of its distance from the deleted ORF. The PPI data were assigned confidence scores based on the experimental evidence available for each interaction using a logistic regression model adapted from ref. 30. For adjacency sets of the essential TLM genes (which belong to the DAmP collection, in which the 3' untranslated regions, rather than the whole gene, has been disrupted<sup>11</sup>), we only considered genes located downstream, because this is the only region likely to be affected by the gene disruption. Adjacency sets of ORFs encoding a protein that does not have any known physical interaction were excluded from the analysis.

**Protein-protein interaction data source.** PPI data was assembled from various sources, which include affinity purification of protein complexes<sup>31,32</sup> (using the spoke model), yeast two-hybrid experiments<sup>33,34</sup> and literature curation<sup>35,36</sup>.

**The NIRVANA algorithm.** We assigned every PPI (edge) with a confidence score (in the range (0,1)) based on the available experimental evidence for it, using a logistic regression model we have previously described<sup>16</sup>. We also computed for every member in an adjacency set a prior probability for it to be a true effector (see below). We added to the network an additional node labeled 'root' and added directed edges to this node from the set of anchor proteins (in the telomere case the anchor set includes ten proteins that directly bind to the telomere or that compose the telomerase machinery as in ref. 16; in the altered response to rapamycin (ARR) case, the anchor set included the genes *TOR1*, *TOR2* and *FPR1*). We then searched for a high-confidence network that connects the root node to at least one member of every adjacency set (note that the paths to the root must go through at least one member of the anchor set). The algorithm aims to optimize local and global features of the reconstructed network. The local criterion favors highly reliable pathways between the root and the selected representatives of each set, optimizing

$$F_L(H) = \sum_{v \in X(H)} \left( -\log(p(v)) + \sum_{e \in P_H(v, \text{root})} -\log(p(e)) \right) \quad (1)$$

where  $H$  is a reconstructed network;  $X(H)$  is the set of representatives from the different adjacency sets that appear in  $H$  (at least one per set);  $p(v)$  is the prior probability assigned to the representative  $v$ ;  $P_H(v, \text{root})$  is the shortest (highest likelihood) path between the representative  $v$  and the root node in  $H$ ; and  $p(e)$  is the confidence score assigned with an edge  $e$ . The global criterion

looks for a parsimonious network that connects the root to at least one representative of each set, optimizing

$$F_G(H) = \sum_{v \in X(H)} -\log(p(v)) + \sum_{e \in H} -\log(p(e)). \quad (2)$$

Considering only the global criterion, this is standard instance of the group Steiner tree problem in directed graphs: find the least heavy subgraph, rooted in the 'root' node that contains at least one representative from every group (adjacency set). As we also wanted to account for the local criterion, we used the algorithm and software presented in refs. 16,37, which allow for a joint and balanced optimization of both criteria. We adjusted the algorithm to handle groups of target nodes (adjacency sets) rather than single nodes using the standard reduction from group Steiner tree to directed Steiner tree<sup>38</sup>.

To avoid an arbitrary choice among equally good choices (that is, equally plausible candidates from an adjacency set), our implementation recorded multiple solutions as in ref. 16. This was done by 50,000 random shuffles of the order by which the yeast PPI is processed. The random ordering affects the way ties are being handled during the run of the algorithm, thus producing different solutions. The resulting output network of NIRVANA is the union of all solutions obtained.

**Fitting a prior probability for NGE.** We assigned each member of an adjacency set with a probability for it to be the causal gene that is conditioned on its distance from the ORF of the deleted gene. For each adjacency set we considered all triplets of the form <"member gene", "deleted gene", distance (bp)> (note that the deleted gene is also a member gene with a distance of zero). Taking all the adjacency sets together, we divided the pairs into bins according to their distance value with intervals of 50 bp. We then computed the probability ( $p$ ) for each bin using a simple Bayesian rule:  $p(\text{NGE} \mid \text{distance}) = p(\text{distance} \mid \text{NGE}) \times p(\text{NGE}) / p(\text{distance})$ . The values on the right side of the equation were computed based on an external data source: for the TLM set we used the manual decision rules of the ARR set to define the set of triples and vice versa. After the Bayesian computation, we smoothed the posterior so that it was monotonous decreasing with distance using the pool adjacent-violators algorithm.

**Evaluating prediction accuracy.** We computed NIRVANA's accuracy as the average of (percentage of correct predictions in adjacency sets annotated as NGE) and (percentage of correct predictions in adjacency sets annotated as non-NGE). We compared these results to a random expectation obtained by randomly choosing representatives from each adjacency set. In the random process we retained the number of adjacency sets in which more than one gene was selected (but we randomly selected the adjacency set in which more than one gene was picked). We repeated the random procedure 1,000 times and for each run computed a success rate. We reported the average random success rate plus or minus 1 s.d.

**Applying NIRVANA on randomized data.** To test the stability of the NGE estimation using NIRVANA, we repeated the analysis of the TLM and ARR sets with randomized data. To this end, we applied NIRVANA on 50 randomized inputs obtained from the original (TLM or ARR) adjacency sets. In each random

simulation we retained the original gene set and randomized the neighbors. This process retains the structure of the adjacency sets (number of neighbors in each set and their prior probabilities) and also possible overlaps between neighbor nodes in different adjacency sets.

We expect that in the randomized data, NIRVANA would still recognize the cases in which the originally deleted gene is not likely to be causative. Indeed, we found that in cases in which the deleted gene was causative (according to the decision rules) NIRVANA did not choose a random 'decoy' neighbor instead (happened on average only in 2.5% of the cases in TLM and 4.1% in ARR), testifying to the stability of the non-NGE predictions. Conversely, for cases in which the deleted gene was not the causative one, NIRVANA was forced to choose between false candidates. Consequently we expected a higher ratio of cases in which the deleted gene was not selected. Indeed the observed ratio was substantially higher (average of 17.7% in TLM and 18.7% in ARR) and reflected the prior probabilities for neighbor selection (random selection based on the prior probabilities yielded an expected rate of 20.7% in TLM and 26.3% in ARR).

To estimate the random expectation for exclusion of the deleted node based only on the prior probabilities, we randomly choose a representative from every adjacency set  $G$ , where the chance of a member  $k$  to be selected is

$$\frac{P_k}{\sum_{i \in G} P_i} \quad (3)$$

where  $p_i$  is the prior probability assigned to member  $i$ . As before, we retained the number of adjacency sets in which more than one gene is selected. This was done by removing the first selected gene from the set and recomputing the chances as above using only the remaining members.

**Confirmation of the causative open reading frame using a complementation test.** The selected ORFs were tested using the relevant strains from the *Saccharomyces* Genome Deletion Project<sup>4</sup> or were deleted in the S288c background. These strains were transformed with plasmids that contained the wild-type

copy of the deleted gene, the adjacent gene or no gene at all. (We list all plasmids and primers used for cloning in **Supplementary Tables 8 and 9**, respectively.) Transformant cells underwent at least six restreaks (~150 cell divisions) on the appropriate selective medium to prevent plasmid loss before measuring their telomere length.

Total genomic DNA was digested with XhoI and was analyzed by Southern blot with the use of a probe specific for subtelomeric repeats as described in ref. 15. PCR fragments containing telomeric sequences and a genomic region that hybridizes to bands (2,044 bp and 779 bp) were used as probes. Each strain and the isogenic wild-type controls were run in triplicate.

**RNA and real-time PCR.** Total cellular RNA was isolated from the wild type and deletions strains using MasterPure yeast RNA purification kit (Epicentre Biotechnologies). Reverse transcription was carried out using Superscript first strand synthesis system, followed by real-time quantitative PCR with primers specific for each ORF. RNA levels were determined relative to a control gene, *ACT1*. A list of all primers used is found in **Supplementary Table 9**. Statistical methods used in this study are unpaired two-tailed  $t$ -tests, assuming unequal variance.

30. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* **102**, 1974–1979 (2005).
31. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
32. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
33. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
34. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
35. Reguly, T. *et al.* Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5**, 11 (2006).
36. Breitkreutz, B.J. *et al.* The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* **36** Database issue, D637–D640 (2008).
37. Yosef, N. *et al.* ANAT: a tool for constructing and analyzing functional protein networks. *Sci. Signal.* **4**, pl1 (2011).
38. Charikar, M. *et al.* Approximation algorithms for directed Steiner problems. *J. Algorithms* **33**, 73–91 (1999).