A gene ontology inferred from molecular networks

Janusz Dutkowski[1], Michael Kramer[1], Michal A Surma[2,3], Rama Balakrishnan[4], J Michael Cherry[4], Nevan J Krogan[3,5] & Trey Ideker[1,6]

Ontologies have proven very useful for capturing knowledge as a hierarchy of terms and their interrelationships. In biology a major challenge has been to construct ontologies of gene function given incomplete biological knowledge and inconsistencies in how this knowledge is manually curated. Here we show that large networks of gene and protein interactions in *Saccharomyces cerevisiae* can be used to infer an ontology whose coverage and power are equivalent to those of the manually curated Gene Ontology (GO). The network-extracted ontology (NeXO) contains 4,123 biological terms and 5,766 term-term relations, capturing 58% of known cellular components. We also explore robust NeXO terms and term relations that were initially not cataloged in GO, a number of which have now been added based on our analysis. Using quantitative genetic interaction profiling and chemogenomics, we find further support for many of the uncharacterized terms identified by NeXO, including multisubunit structures related to protein trafficking or mitochondrial function. This work enables a shift from using ontologies to evaluate data to using data to construct and evaluate ontologies.

Ontologies are central to many branches of biomedical research. In recent years, numerous ontologies have been developed to capture structured knowledge about taxonomy, anatomy and development, cellular and molecular function, bioactive compounds, and clinical diagnosis and disease, and other areas[1,2]. One very successful ontology is the GO, which aims to unify all knowledge about biological processes, cellular components and molecular functions through a hierarchy of biological terms which are, in turn, used to describe genes[3]. GO is widely used for systematic assessment of the key functions and processes enriched in a set of genes identified by genomic, transcriptomic or proteomic data. The development of GO was transformational because it provided a gold-standard reference of gene functions against which any data set could be assessed.

Historically, GO (and most other ontologies[1,2]) has been built and curated manually by teams of domain experts. However, as ontologies grow in size and complexity—GO currently represents a total of 34,765 terms and 64,635 hierarchical term-term relations annotating genes from >80 species—manual curation is encountering a series of hurdles that are becoming increasingly difficult to surmount[4,5]. First, despite stringent curation standards and the availability of advanced text-mining tools[6,7], it has been difficult to maintain consistency in how literature and domain expertise translate to terms and relations in GO. Second, there has been a strong bias in coverage within GO toward processes that are well-studied, and a corresponding lack of coverage of processes that have been more recently identified. Such problems are difficult to assess due to the lack of any definitive gold standard for the rigorous validation of GO.

One solution to these problems would be to systematically structure an ontology using large-scale data sets. Such data sets could be used not only to assign genes to existing terms[8] but also to directly infer new terms and their hierarchical relationships. Systematic inference of ontologies is an area of active research[9,10] but has not, to our knowledge, been applied to construct gene ontologies from omics data (although doing so has been suggested[4]). High-throughput measurements of genetic and protein interactions[11–15] and mRNA expression profiles[16], for example, are available and are already being used to build network maps of the cell[17–20]. Although these networks are often analyzed using hierarchical clustering methods[21–23], the full hierarchy is almost always reduced to a flat set of gene clusters in which one can test for enrichment of existing GO terms[24–26]. In a few cases, it has been shown that some clusters of interactions can be grouped to form larger clusters that represent higher-order biological units[27–32]. The key question, however, is whether the interaction networks can be used to systematically infer a hierarchy of clusters that is analogous to the complete hierarchy of terms represented by GO. If so, one enticing possibility is that the existing collection of high-throughput network maps for an organism could be analyzed to automatically (or semi-automatically) reconstruct and improve GO.

RESULTS

Gene networks embed hierarchical structure consistent with GO

To analyze the agreement between gene networks and GO, we focused on four fundamental types of large interaction networks: physical protein-protein interactions, genetic interactions (synthetic lethality and epistasis), co-expressed genes and an integrated functional network known as YeastNet[20] (Online Methods and **Supplementary Table 1**). Within each of these networks, we examined the interactions falling within and between existing GO terms. The interaction density of each term was computed as the fraction of gene pairs assigned to that term for which an interaction was present in the network

[1]Departments of Medicine and Bioengineering, University of California San Diego, La Jolla, California, USA. [2]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [3]Department of Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, California, USA. [4]Department of Genetics, Stanford University, Stanford, California, USA. [5]J. David Gladstone Institutes, San Francisco, California, USA. [6]Institute for Genomic Medicine, University of California San Diego, La Jolla, California, USA. Correspondence should be addressed to T.I. (tideker@ucsd.edu) or J.D. (janusz@ucsd.edu).

Received 27 June; accepted 22 November; published online 16 December 2012; doi:10.1038/nbt.2463

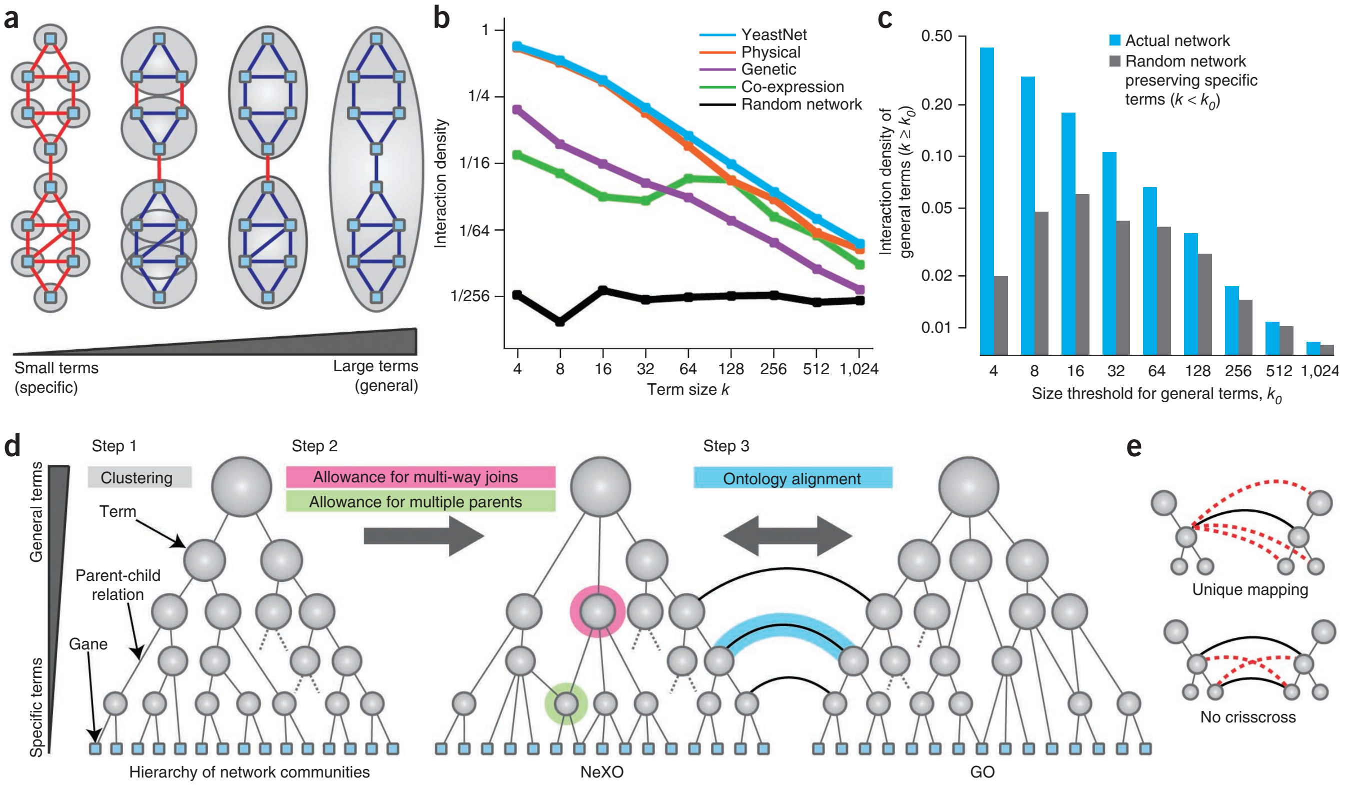


Figure 1 Automated assembly and alignment of gene ontologies. (**a**) Specific GO terms correspond to small gene communities in an interaction network, which are nested within larger communities corresponding to more general terms. Gray ovals represent terms, blue squares represent genes and links represent gene interactions that fall within (dark blue) or between (red) terms. (**b**) The density of interactions within terms in the input networks is plotted as a function of term size (term sizes are binned). Different colors represent each input network controlled for the number of interactions. (**c**) Network density of general terms of size $k \geq k_0$ in the YeastNet (blue bars) in comparison to a random network in which interactions falling within specific terms of size $k < k_0$ are preserved and others are shuffled (gray bars). The difference between the blue and gray bars represents the additional network density contributed by terms of size $\geq k_0$. (**d**,**e**) A data-driven ontology is created using a multistep procedure. First, probabilistic community detection within the input networks yields a binary tree in which nodes correspond to ontology terms and links correspond to parent-child term relations (dotted lines indicate additional branches). Second, unsupported terms in the tree are removed and substituted by multi-way joins, and additional parent-child relations are added based on network data. Third, the resulting ontology is aligned against the reference GO, in a way (**e**) that prohibits non-unique mappings and ancestor-descendant criss-crossing (indicated by dotted red lines).

(**Fig. 1a** and Online Methods). In all four networks, we found that the interaction density of GO terms was substantially greater than that expected for random networks (**Fig. 1b**), indicating general agreement between the network and GO. Although the highest interaction density was observed for specific GO terms (those with few annotated genes), elevated density was also observed for more general terms at all levels of the GO hierarchy. Notably, interaction density for general terms could not be explained simply by the density of more specific terms contained within them (**Fig. 1c**). Rather, dense patterns of network interactions span the specific GO terms that fall underneath the same general term, providing evidence that networks embed hierarchical information consistent with that captured by GO.

Inferring ontologies from networks

Motivated by these results, we developed a multistep automated system for the assembly of gene ontologies based on network data (Online Methods). First, biological networks are integrated and a hierarchy of network communities is identified based on a probabilistic model for community detection[32,33]. This approach seeks to construct a binary tree, or dendrogram, that maximizes the overall probability of the network data by hierarchically joining sets of genes with similar patterns of interactions (**Fig. 1d**, Step 1). These gene sets, represented by nodes in the tree, identify biological entities corresponding to terms in an ontology. Joining two sets, represented by connecting two nodes beneath a third, identifies specialized terms that are part of a more general term.

Although the binary tree enables a computationally tractable approximation of the term ‘hierarchy’, it artificially requires that every term (except those at the root and leaves) connect to exactly two specialized terms below it and a single, more general, term above it. However, many cellular processes, components and functions are composed of more than two parts and participate in multiple parent processes—types of relations that are well-represented by GO. Hence, we transformed the original binary tree to match this more flexible ontology structure. In particular, we identified binary joins in the tree that can be replaced by multiway joins to increase the overall probability score. In addition, the tree was supplemented with optional new connections from nodes to second parents when such relations were supported by the network data (**Fig. 1d**, Step 2).

To directly map and compare this computed ontology to manually curated ontologies such as GO, we had to develop an algorithm for alignment of gene ontologies (**Fig. 1d**, Step 3). Our approach extends general methods for ontology alignment from the computational and cognitive sciences[34], and matches terms between ontologies based on similar gene assignments and similar positions in

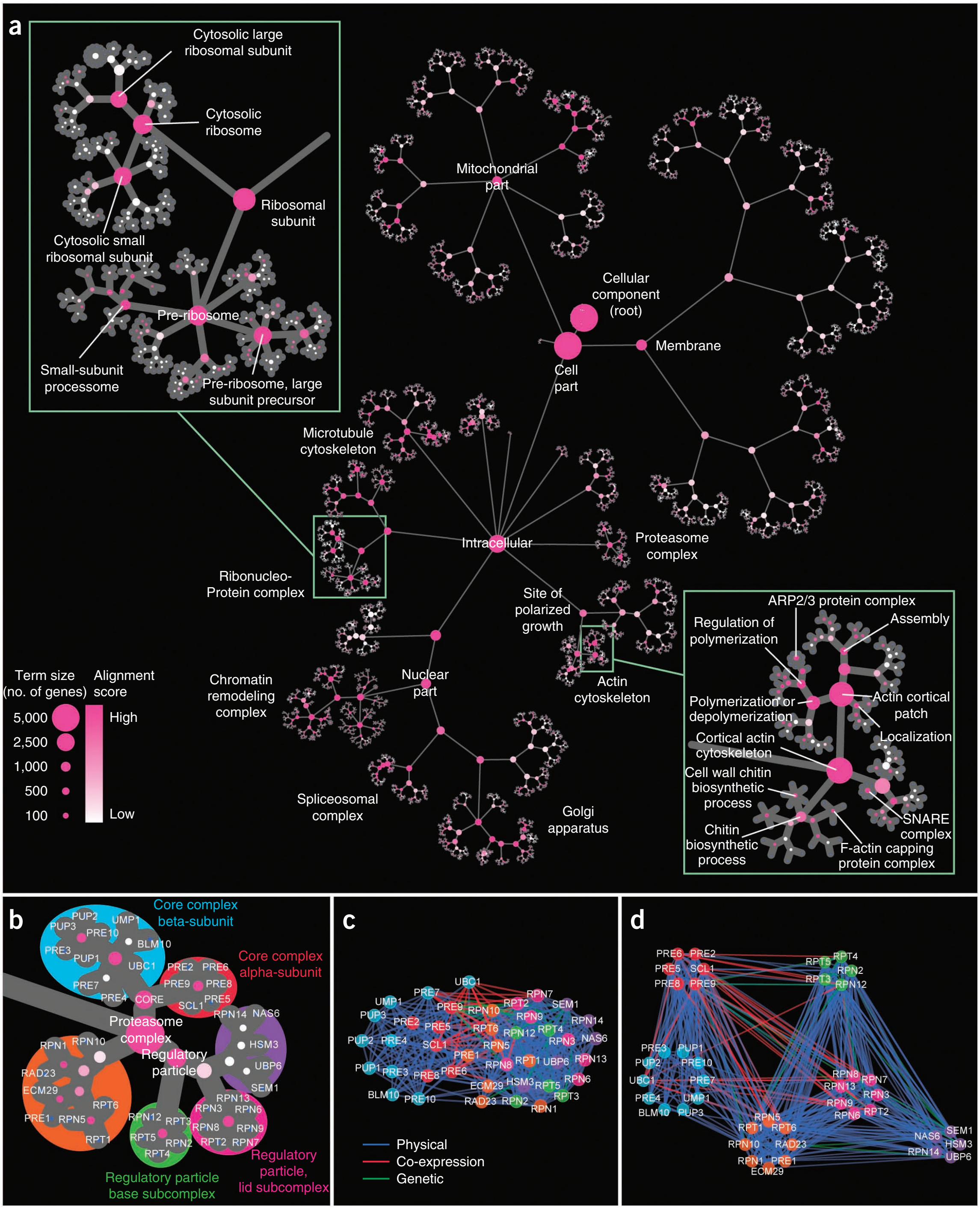


Figure 2 The NeXO ontology. (**a**) The ontology is shown as a tree, with nodes indicating terms and edges indicating hierarchical relations between terms. Node sizes indicate the number of genes assigned to a term. Node colors represent the degree of correspondence to a term in GO as determined by ontology alignment, with high-level alignments labeled. Insets show the hierarchical organization identified for the ribosome and actin cytoskeleton. (**b**–**d**) The subcomponents of the proteasome as identified by NeXO (**b**) are displayed by applying a force-directed network layout (**c**), as is commonly used, and a layout showing interactions falling within and between subcomponents (**d**).

the hierarchy. In contrast to simpler measures that rely strictly on gene assignment, for example, a hypergeometric test or gene set enrichment[35], term mappings are unique (each term is mapped to at most one term in the other ontology) and respect topological constraints (**Fig. 1e**). The effects of this alignment procedure are threefold: to allow for immediate transfer of term labels and definitions from GO, to identify novel terms not found in GO, and to identify consistent and conflicting term-term relations.

The yeast NeXO

We next applied this pipeline to assemble an ontology from the four large yeast networks we had previously obtained. The resulting

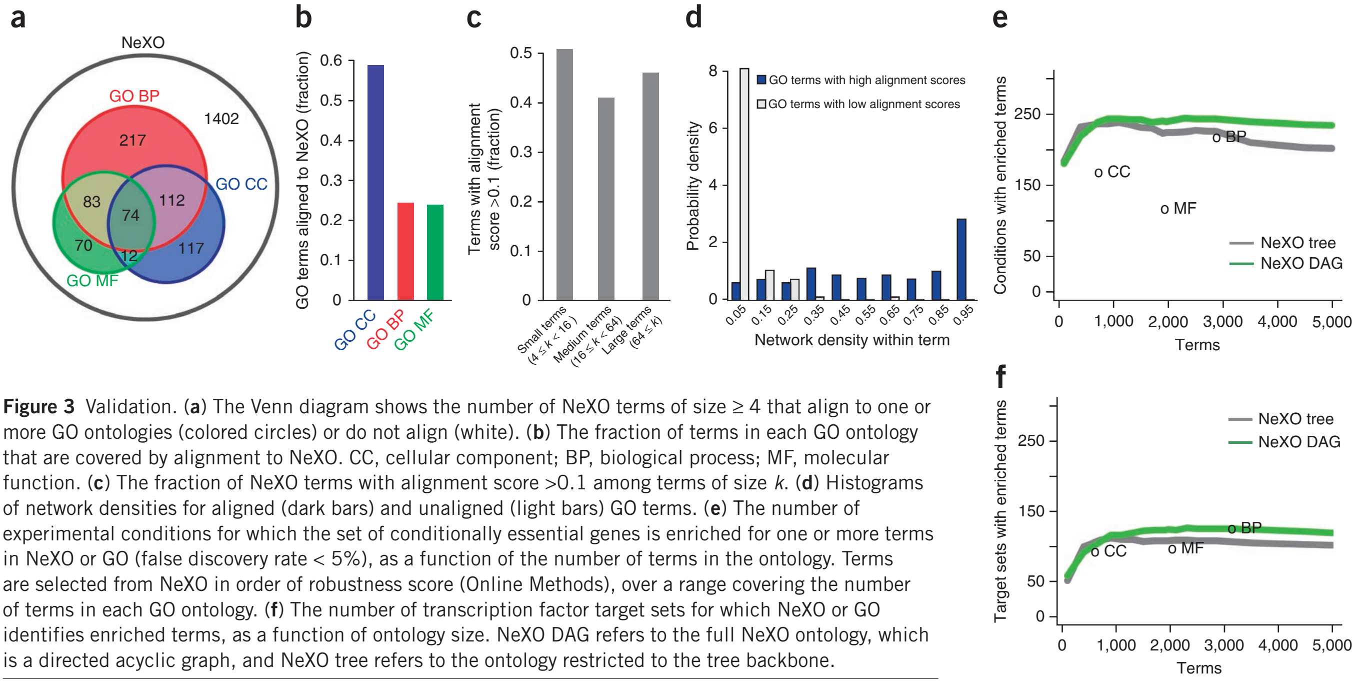


Figure 3 Validation. (**a**) The Venn diagram shows the number of NeXO terms of size ≥ 4 that align to one or more GO ontologies (colored circles) or do not align (white). (**b**) The fraction of terms in each GO ontology that are covered by alignment to NeXO. CC, cellular component; BP, biological process; MF, molecular function. (**c**) The fraction of NeXO terms with alignment score >0.1 among terms of size *k*. (**d**) Histograms of network densities for aligned (dark bars) and unaligned (light bars) GO terms. (**e**) The number of experimental conditions for which the set of conditionally essential genes is enriched for one or more terms in NeXO or GO (false discovery rate < 5%), as a function of the number of terms in the ontology. Terms are selected from NeXO in order of robustness score (Online Methods), over a range covering the number of terms in each GO ontology. (**f**) The number of transcription factor target sets for which NeXO or GO identifies enriched terms, as a function of ontology size. NeXO DAG refers to the full NeXO ontology, which is a directed acyclic graph, and NeXO tree refers to the ontology restricted to the tree backbone.

network-extracted ontology, NeXO, contained a total of 4,123 terms and 5,766 term-term relationships; the complete ontology is provided in both OBO (Open Biological and Biomedical Ontologies) and Cytoscape formats (**Supplementary Note 1** and **Supplementary Files 1**, **2**). NeXO's central tree structure (**Fig. 2a**) has three major branches, which correspond to the intracellular compartment, the membrane and the mitochondrion, respectively. Within the major branches, in particular the intracellular compartment, many subtrees align with major cellular components that are annotated in GO, such as the ribosome and actin cytoskeleton (**Fig. 2a**, insets), as well as the proteasome, chromatin remodeling complexes and the spliceosome. The NeXO hierarchy not only identifies these components but also captures their internal organization. For example, the proteasome consists of the core complex and regulatory particle, which in turn contain the alpha and beta subunits and the base and lid complexes, respectively, all of which are captured by the NeXO ontology (**Fig. 2b**). This hierarchical organization is determined completely by the network based on the density of interactions between and within the different proteasome components (**Fig. 2c**,**d**). Some of this structure is visualized in a force-directed layout of the raw interaction data (**Fig. 2c**), which broadly segregates the proteasome into two clusters of protein interactions, but the full structure becomes apparent only after hierarchical module detection and alignment against GO (**Fig. 2b**).

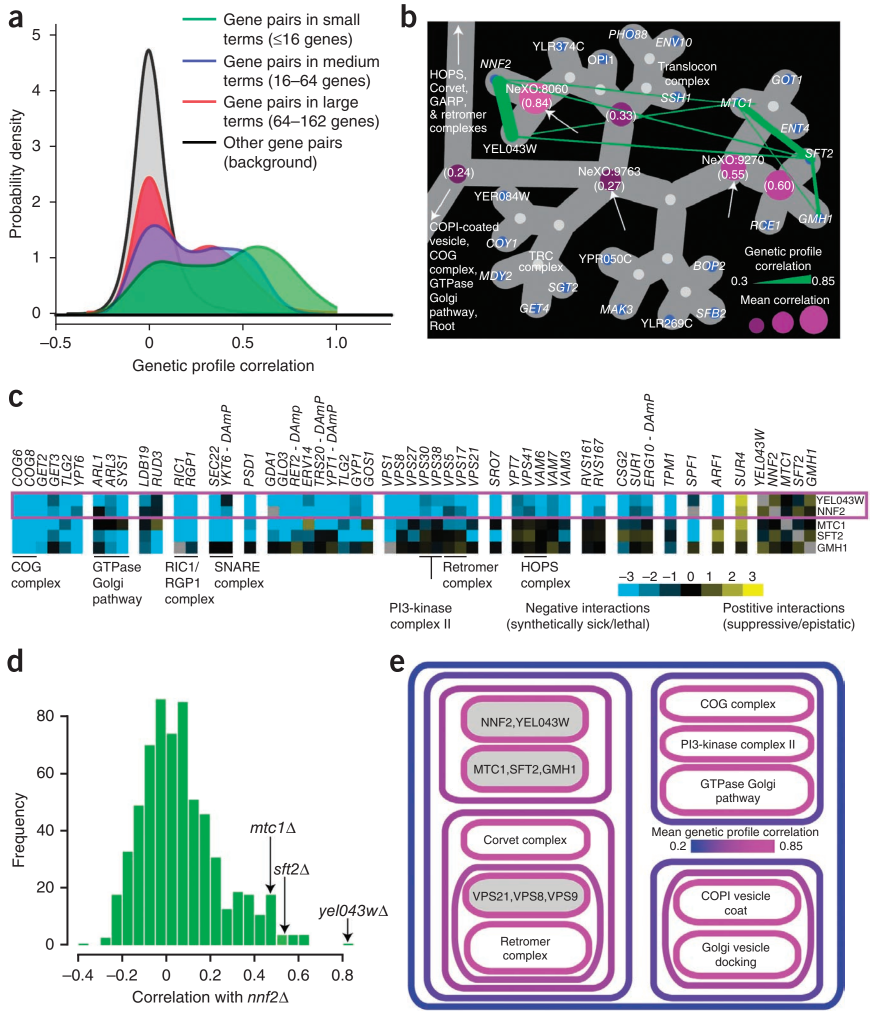
Based on the ontology alignment, we found that 33% of terms in NeXO map to terms in the three GO ontologies (Biological Process, Cellular Component, Molecular Function) with many terms mapping to more than one ontology (**Fig. 3a**). Conversely, NeXO captures nearly 60% of terms in the Cellular Component ontology and roughly a quarter of terms in the other two GO ontologies (**Fig. 3b**). Thus, NeXO largely represents an ontology of cellular components, which might indicate that these are the structures best highlighted by the input networks. In addition, we found that NeXO captures all levels of the GO hierarchy including both general and specific GO terms (**Fig. 3c**). Finally, genes in GO terms that align to NeXO were much more densely connected than genes in unaligned GO terms (**Fig. 3d**), verifying that NeXO tends to correctly identify GO terms if they have good network support. We considered that some of this observed correspondence between NeXO and GO might occur by construction, since cut-off thresholds for two of the input networks (YeastNet and co-expression) had been optimized to connect genes with similar GO Biological Process annotations. However, very similar alignment results were obtained when these two networks were removed and NeXO was built using the remaining protein-protein and genetic interaction networks that have not been influenced by GO in any explicit way (Online Methods and **Supplementary Fig. 1**). Moreover, the NeXO-GO alignment remained relatively stable over a wide range of thresholds for identifying protein-protein and genetic interactions (**Supplementary Fig. 2**). Finally, we found that the alignment results were stable when excluding gene-to-term associations in GO based on high-throughput interaction data (**Supplementary Fig. 3**).

To further validate NeXO in comparison to GO, we used both ontologies to perform a functional enrichment analysis of gene sets, the task for which GO is most often used[24,26]. To this end, we downloaded two large data sets that had not been used in ontology construction: a genome-wide screen for genes required for growth across 418 experimental conditions[36] and a database of direct gene targets for each of 183 transcription factors[37]. NeXO identified significantly enriched (false discovery rate < 5%) terms in the sets of required genes for 244 of the experimental conditions (58%)—an improvement over all three GO hierarchies (**Fig. 3e**). NeXO also yielded enriched terms for 126 transcription factor target sets (69%), matching the best result for GO but with a smaller number of terms (**Fig. 3f**). Thus, the data-driven ontology provides functionally relevant terms covering a wide spectrum of yeast biology to an extent comparable with manually curated efforts.

Using NeXO to identify ontology terms and relations

Many NeXO terms aligned well with GO terms, but perhaps even more interesting are the terms and relations in NeXO that were not cataloged in GO. Of these 449 terms and 123 relations had particularly strong network support (Online Methods and **Supplementary Tables 2** and **3**). To further explore previously uncataloged terms and

Figure 4 Evaluation of protein trafficking terms using genetic interaction profiling. (**a**) Correlation of genetic interaction profiles among genes associated with Golgi apparatus in NeXO, for gene pairs annotated to subterms of different sizes. For each gene pair, the smallest subterm common to both genes is considered. (**b**) NeXO branch for term NeXO:9763 containing *NNF2*, YEL043W, *MTC1* and *SFT2*. Green lines indicate high genetic interaction profile correlations ($R > 0.4$). Node colors indicate similarity of genetic interaction profiles within each nested module (mean profile correlation is provided in parenthesis). NeXO IDs are provided for terms that are discussed in more detail in the text. (**c**) Comparison of a subset of the genetic interaction scores from the profiles generated from *nnf2*Δ, *yel043w*Δ, *mtc1*Δ, *sft2*Δ and *gmh1*Δ. Yellow and blue correspond to positive and negative interactions, respectively. (**d**) Distribution of correlation coefficients of the *nnf2*Δ genetic interaction profile versus a selection of ~750 other genes, highlighting the *yel043w*Δ profile as an extremely high correlation. (**e**) Nested modules or protein complexes in close proximity to NeXO:9763. Border colors indicate similarity of genetic interaction profiles within each nested module. Nodes in gray are discussed in more detail in the text.



relations, we generated quantitative genetic interaction profiles (Online Methods and **Supplementary Table 4**) for 73 genes annotated beneath term NeXO:9965. This term aligned strongly to the 'Golgi apparatus' component of GO and contained known Golgi subcomponents as well as 32 uncataloged descendent terms. Notably, the genetic interaction profiles of genes annotated to Golgi were much more highly correlated than those of a background set of genes (**Fig. 4a**). We observed higher correlations for gene pairs annotated to a common small term, and lower correlations for gene pairs annotated to larger terms, lending support for the hierarchical organization recovered by NeXO.

For example, the uncharacterized term NeXO:9763 represents several layers of substructure corroborated by the new genetic interaction profiles (**Fig. 4b**). These profiles are, on average, more highly correlated to each other ($R = 0.27$) than are random gene pairs ($R = 0.02$) or gene pairs assigned generally to Golgi ($R = 0.20$). Within this new term NeXO identifies a highly specific term joining *NNF2* and the gene encoding YEL043W (NeXO:8060, **Fig. 4b**). Comparison of the genetic interaction profiles generated from *nnf2*Δ and *yel043w*Δ revealed a strikingly high correlation ($R = 0.84$) (**Fig. 4b,c**), suggesting a close functional relationship. Indeed, of all the genetic interaction profiles measured, the *nnf2*Δ profile is most correlated with that of *yel043w*Δ (**Fig. 4d**). High genetic interaction correlations are also observed with deletions of *MTC1* and *SFT2*—genes that, together with *GMH1*, form another new term (NeXO:9270) that is adjacent to *NNF2* and YEL043W in NeXO (**Fig. 4b,d**). Thus, the genetic profiling data are highly consistent with the NeXO structure, that is, genes assigned to the same specific terms have higher genetic profile correlations than genes assigned to the same general terms (**Fig. 4b,e**).

These new terms are positioned next to the retromer, a complex that regulates recycling transmembrane receptors from endosome to the trans-Golgi network[38], and the HOPS and Corvet complexes, which serve as tethering complexes by capturing endosomal vesicles[39,40] (**Fig. 4e**), strongly suggesting that the new terms represent integral components in endosomal and Golgi regulation. Consistent with this notion, we observed strong negative genetic interactions of *nnf2*Δ and *yel043w*Δ with deletions of components of the (i) retromer (*VPS5* and *VPS17*); (ii) HOPS (*VPS41* and *VAM6*); (iii) the COG complex (*COG6* and *COG8*), a regulator of the Golgi-glycosylation machinery; (iv) RIC1, RGP1 and YPT6, a pathway required for fusion of endosome-derived vesicles with the Golgi, in which Ric1 and Ypt6 are co-complexed and act as a nucleotide exchange factor for the GTPase *YPT6*; and (v) *VPS30* and *VPS38*, which encode components of the PI3-kinase Complex II (**Fig. 4e**). A number of other uncharacterized terms are supported by the new genetic interaction profiles, including the term NeXO:8891, which is composed of *VPS8*, *VPS21* and *VPS9* and is placed directly next to retromer subcomponents *VPS5* and *VPS17*, with which it shares very strong genetic profile correlations (**Fig. 4e** and **Supplementary Fig. 4a**). Although additional work will be required to understand the exact function of these factors, the genetic analysis provides good evidence for NeXO's ability to identify new components and functions and to pinpoint hierarchical relationships with known components.

In addition to terms supported by genetic interaction profiling, we also found 115 uncharacterized terms that were enriched for genes required for growth under specific environmental conditions, providing

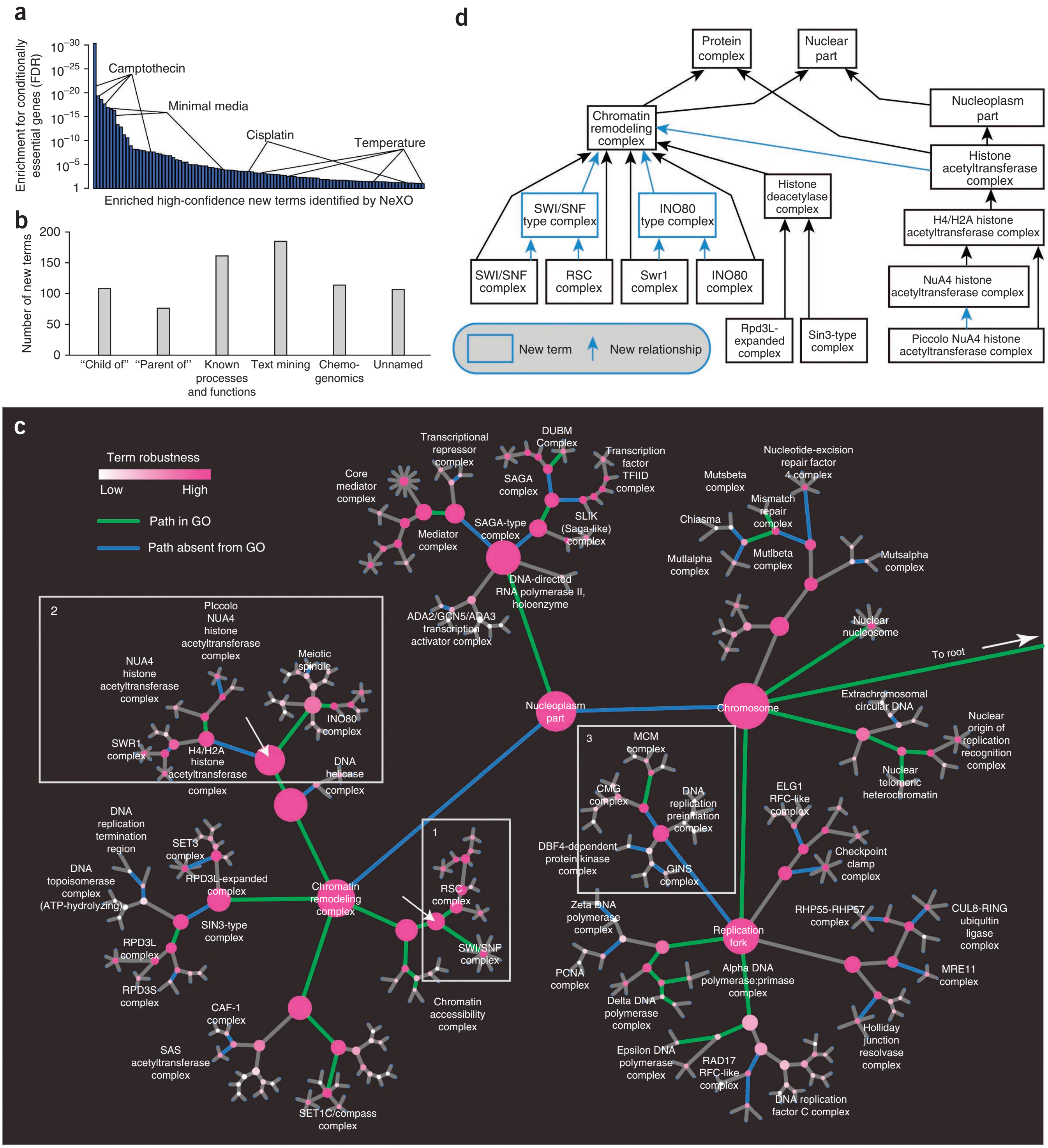


Figure 5 Updating GO with additional terms and term relations. (**a**) Robust terms enriched for genes necessary for growth under an experimental condition (false discovery rate < 10%). Top-scoring conditions for selected terms are indicated. (**b**) The number of terms derived using NeXO that can be named based on parent-child relationships with aligned terms, alignment to other GO ontologies, text mining or chemogenomic data. (**c**) The branch of NeXO containing the chromatin remodeling, DNA replication and DNA repair machinery. Node colors represent term robustness. Node labels indicate aligned GO Cellular Component terms where applicable. Green paths indicate consistent term-term relations between NeXO and GO, that is, ancestor-descendant pairs in NeXO aligned to GO terms that are also in an ancestor-descendant relation. A blue edge indicates a relation present in NeXO but absent from GO, such that the two GO terms aligned to the child and the nearest aligned NeXO ancestor, respectively, are not in a descendant-ancestor relation. Insets indicate regions of NeXO discussed in the main text and arrows indicate selected terms identified using NeXO. (**d**) Examples of NeXO-derived terms (blue rectangles) and term-term relations (blue arrows) that were submitted to the GO Consortium based on this work.

useful insight into the functions of these terms (**Fig. 5a** and **Supplementary Table 5**). For example, term NeXO:6375 was composed of six poorly characterized genes: *MTC2*, *MTC4*, *MTC6*, *DLT1*, YBR197C and YPR153W, of which the first three had been associated with maintenance of telomere capping[41] (**Supplementary Fig. 5**). This term was placed under the mitochondrial component by NeXO next to genes annotated with oxoreductase activity and antimonite transport. Interestingly, deletion of any of the six genes increases yeast sensitivity to mercury chloride ($HgCl_2$) (**Supplementary Fig. 5**), an agent that promotes oxidative stress and has been found to induce apoptosis through a mitochondrial-dependent pathway[42]. Although the association of these genes with mitochondrial function has not yet been established, some telomere maintenance genes, including telomerase and *MTC3*, are known to localize to the mitochondria[43,44] (telomerase, in particular, under oxidative stress), supporting the localization of the new term in NeXO.

We pursued several other bioinformatic means of gaining biological insight into the terms identified by NeXO (**Fig. 5b** and **Supplementary Table 2**). In particular, we identified many cases in which an unknown term could be assigned a temporary name based on (i) its relationship with a known parent or child term, (ii) alignment to a term in the GO Biological Process or Molecular Function hierarchy or (iii) text mining the 'Description' field of the *Saccharomyces* Genome Database[45] for text phrases that are common among the genes assigned to the term (**Fig. 5b** and **Supplementary Table 2**). We also found that the network data can be further mined to suggest the type of relationship between terms, such as the 'part_of' and 'is_a' relations used in the GO Cellular Component ontology (**Supplementary Fig. 6**). 'Part_of' relations indicate the child is an actual portion of the parent (such as subunits of a protein complex), whereas 'is_a' relations define the child as a particular kind or subtype of the parent (such as a family of complexes related by function). Further work will be required to determine the best strategy to automatically specify the types for all relations in NeXO.

Using NeXO to systematically update and expand GO

We were also able to recognize many NeXO-derived terms and term relations as absent from GO but having strong support in the literature. For instance, term NeXO:6164 groups together *BLS1*, *SNN1*, *CNL1*—encoding subunits of the BLOC complex, which was recently defined in yeast[46] but not yet incorporated in the yeast GO. In chromatin remodeling (**Fig. 5c**, insets 1,2), the SWI/SNF and RSC complexes are grouped under the same robust new parent term, as are the INO80 and SWR1 complexes. Although neither of these parent terms was documented in GO (**Fig. 5d**), both are well documented in the literature[47]. An additional NeXO-derived term also joins SWR1 and NUA4, two chromatin-modifying complexes with overlapping functions and components[48], which have been proposed to function as the single Tip60 complex in humans[49,50] including homologs of the INO80 complex[50]. NeXO correctly places both histone acetyltransferases (HATs) and histone deacetylase complexes (HDACs) under the parent term 'Chromatin Remodeling Complex'[51] whereas, in GO, HATs are descendants of this term but HDACs are not (**Fig. 5d**). NeXO also correctly identifies the Piccolo NUA4 complex as part of the parent NUA4 complex (**Fig. 5c**, inset 2, and **Fig. 5d**) and the CMG complex as part of the DNA replication preinitiation complex[52] (**Fig. 5c**, inset 3), relations that were missing from the current GO.

All of these NeXO-derived terms and relations were submitted to the GO Consortium for inclusion in the ontology and the following changes were incorporated by the ontology editor: GO:0043189: H4/H2A histone acetyltransferase complex was made a child of GO:0016585:chromatin remodeling complex, GO:0035267:NuA4 histone acetyltransferase complex was made a parent of GO:0032777: Piccolo NuA4 histone acetyltransferase complex, and GO:0031261: DNA replication preinitiation complex was made a parent of GO:0071162:CMG complex. A new term named 'INO80 type complex' was made a parent of GO:0000812:SWR1 and GO:0031011: INO80, and GO:0070603:SWI/SNF type complex is now a parent of GO:0016586:RSC and GO:0016514:SWI/SNF complex. In addition the products of yeast genes *BLS1*, *SNN1* and *CNL1* were annotated to the BLOC-1 complex. Thus, NeXO provides a systematic means of directing literature curation efforts to biological mechanisms that are known but have not yet been considered by curators. A cursory inspection of the complete list of additional terms and relations derived using NeXO indicates that many more of these may already have some literature support and are good candidates for further GO curation.

DISCUSSION

A key challenge in biology is to capture knowledge about the cell in a way that is accurate, unbiased and scalable. Toward this goal, we have described our efforts to systematically construct an entire gene ontology directly from large-scale network data and compare it to the manually constructed GO. Our approach involves (i) probabilistic clustering of networks to yield a hierarchy of putative terms and relations, (ii) transformation of the hierarchy to match the structure of an ontology and (iii) alignment of the resulting NeXO ontology with GO in order to name the terms and term relations of known biology and to identify those that are newly identified by NeXO.

Although NeXO and GO have similar functionalities—browsing the hierarchy of terms, searching for gene-to-term associations and performing functional enrichment—a key difference is that NeXO does not assign common English language names and definitions to all terms. In NeXO, terms that do not align with previous knowledge (by alignment to GO, text mining or functional enrichment in omics data) have only a systematic ID assigned. Lack of a common name does not imply lesser importance, however, but only that the term represents a biological entity not previously named by a human investigator. This process of systematically identifying and naming entities in a cell is not unlike the process of gene finding in a sequenced genome. Whereas genes are defined by their nucleotide sequences, cellular components and functions are defined by the intrinsic patterns of interaction shared by their subunits. Sequence analysis of genomes routinely identifies unknown genes that are initially assigned systematic IDs, and it is precisely these uncharacterized entities that present some of the most interesting opportunities for future study.

GO and NeXO complement each other in mutually beneficial ways. One of GO's main values lies in providing a uniform gold-standard vocabulary for referencing well-characterized cellular components and functions, but we also see room for a parallel ontology that is tied directly to data and is not limited by prior knowledge and curation. In this respect, ontology alignment provides the means to map between these two types of ontologies to enable the bidirectional transfer of both well-established and new information. As we have shown, this process can be used to objectively identify additional terms and relations that are missing from GO, some of which have now been added by the GO Consortium. An intriguing question is whether the most robust terms and relations in a data-driven ontology should be automatically created in GO, perhaps with special evidence codes akin to the codes already in use for gene-to-term associations[53].

Networks have long been instrumental in representing and visualizing biological relationships. The challenge now is to transform these networks into representations that capture the multiscale modularity inherent in all biological systems. Although such representations are present in manually curated gene ontologies, we have shown that gene ontologies can also be assembled and curated automatically from high-throughput network data. The research reported in this manuscript raises the possibility that, given the appropriate tools, ontologies might evolve over time with the addition of each new network map or high-throughput experiment that is published. More importantly, it enables a philosophical shift in bioinformatic analysis, from a regime in which the ontology is viewed as gold standard to one in which it is the major result.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to J. Bader and Y. Park for assistance with the HAC-ML software and to I. Lee and E. Marcotte for advice on YeastNet. We would also like to thank D. Botstein, C. Boone, R. Sharan and R. Shamir for constructive advice on the manuscript. This work was generously supported by US National Institutes of Health grants P41-GM103504 and P50-GM085764 (TI) and R01-GM084448 and P50-GM081879 (N.J.K.). N.J.K. is a Searle Scholar and Keck Young Investigator.

AUTHOR CONTRIBUTIONS

J.D. and T.I. conceived and designed the analysis. J.D. performed initial data analysis, constructed the NeXO ontology and performed all computational experiments. M.K. designed and implemented the ontology alignment procedure with guidance from J.D. M.S. and N.J.K. performed the quantitative genetic interaction profiling and interpreted the data. R.B. and J.M.C. investigated and curated the new ontology terms and relations. J.D. and T.I. wrote the manuscript. All authors contributed to the manuscript and approved its final version.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nbt.2463.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
2. Musen, M.A. *et al.* The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* **19**, 190–195 (2012).
3. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
4. Fraser, A.G. & Marcotte, E.M. A probabilistic view of gene function. *Nat. Genet.* **36**, 559–564 (2004).
5. Leonelli, S., Diehl, A.D., Christie, K.R., Harris, M.A. & Lomax, J. How the gene ontology evolves. *BMC Bioinformatics* **12**, 325 (2011).
6. Krallinger, M., Leitner, F. & Valencia, A. Analysis of biological processes and diseases using text mining approaches. *Methods Mol. Biol.* **593**, 341–382 (2010).
7. Raychaudhuri, S., Chang, J.T., Sutphin, P.D. & Altman, R.B. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.* **12**, 203–214 (2002).
8. Pena-Castillo, L. *et al.* A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol.* **9** (suppl.1), S2 (2008).
9. Buitelaar, P. & Cimiano, P. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, Vol. 167 (IOS Press, Amsterdam, 2008).
10. Coulet, A., Shah, N.H., Garten, Y., Musen, M. & Altman, R.B. Using text to build semantic networks for pharmacogenomics. *J. Biomed. Inform.* **43**, 1009–1019 (2010).
11. Collins, S.R. *et al.* Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6**, 439–450 (2007).
12. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
13. Tarassov, K. *et al.* An *in vivo* map of the yeast protein interactome. *Science* **320**, 1465–1470 (2008).
14. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
15. Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. *Science* **333**, 601–607 (2011).
16. Gasch, A.P. *et al.* Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell* **12**, 2987–3003 (2001).
17. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
18. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
19. Myers, C.L. *et al.* Discovery of biological networks from diverse functional genomic data. *Genome Biol.* **6**, R114 (2005).
20. Lee, I., Li, Z. & Marcotte, E.M. An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS ONE* **2**, e988 (2007).
21. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
22. Girvan, M. & Newman, M.E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821–7826 (2002).
23. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
24. Khatri, P. & Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21**, 3587–3595 (2005).
25. D'haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **23**, 1499–1501 (2005).
26. Gibbons, F.D. & Roth, F.P. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* **12**, 1574–1581 (2002).
27. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabasi, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
28. Dotan-Cohen, D., Letovsky, S., Melkman, A.A. & Kasif, S. Biological process linkage networks. *PLoS ONE* **4**, e5313 (2009).
29. Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. USA* **101**, 2981–2986 (2004).
30. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566 (2005).
31. Jaimovich, A., Rinott, R., Schuldiner, M., Margalit, H. & Friedman, N. Modularity and directionality in genetic interaction maps. *Bioinformatics* **26**, i228–i236 (2010).
32. Park, Y. & Bader, J.S. Resolving the structure of interactomes with hierarchical agglomerative clustering. *BMC Bioinformatics* **12** (suppl.1), S44 (2011).
33. Clauset, A., Moore, C. & Newman, M.E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).
34. Jean-Mary, Y.R., Shironoshita, E.P. & Kabuka, M.R. Ontology Matching with Semantic Verification. *Web Semant.* **7**, 235–251 (2009).
35. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
36. Hillenmeyer, M.E. *et al.* Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome Biol.* **11**, R30 (2010).
37. Abdulrehman, D. *et al.* YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface. *Nucleic Acids Res.* **39**, D136–D140 (2011).
38. Seaman, M.N. Recycle your receptors with retromer. *Trends Cell Biol.* **15**, 68–75 (2005).
39. Nickerson, D.P., Brett, C.L. & Merz, A.J. Vps-C complexes: gatekeepers of endolysosomal traffic. *Curr. Opin. Cell Biol.* **21**, 543–551 (2009).
40. Peplowska, K., Markgraf, D.F., Ostrowicz, C.W., Bange, G. & Ungermann, C. The CORVET tethering complex interacts with the yeast Rab5 homolog Vps21 and is involved in endo-lysosomal biogenesis. *Dev. Cell* **12**, 739–750 (2007).
41. 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).
42. Araragi, S. *et al.* Mercuric chloride induces apoptosis via a mitochondrial-dependent pathway in human leukemia cells. *Toxicology* **184**, 1–9 (2003).
43. Saretzki, G. Telomerase, mitochondria and oxidative stress. *Exp. Gerontol.* **44**, 485–492 (2009).
44. Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
45. Cherry, J.M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).
46. Hayes, M.J., Bryon, K., Satkurunathan, J. & Levine, T.P. Yeast homologues of three BLOC-1 subunits highlight KxDL proteins as conserved interactors of BLOC-1. *Traffic* **12**, 260–268 (2011).
47. Clapier, C.R. & Cairns, B.R. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).
48. Lu, P.Y., Levesque, N. & Kobor, M.S. NuA4 and SWR1-C: two chromatin-modifying complexes with overlapping functions and components. *Biochem. Cell Biol.* **87**, 799–815 (2009).
49. Auger, A. *et al.* Eaf1 is the platform for NuA4 molecular assembly that evolutionarily links chromatin acetylation to ATP-dependent exchange of histone H2A variants. *Mol. Cell. Biol.* **28**, 2257–2270 (2008).
50. van Attikum, H. & Gasser, S.M. The histone code at DNA breaks: a guide to repair? *Nat. Rev. Mol. Cell Biol.* **6**, 757–765 (2005).
51. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
52. Evrin, C. *et al.* A double-hexameric MCM2–7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc. Natl. Acad. Sci. USA* **106**, 20240–20245 (2009).
53. Hong, E.L. *et al.* Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.* **36**, D577–D581 (2008).

ONLINE METHODS

Input networks. We obtained four yeast networks from public databases, corresponding to physical protein-protein interactions (BioGRID[54]), synthetic-lethal and epistatic genetic interactions (DRYGIN database[14]; http://drygin.ccbr.utoronto.ca/), co-expression relationships (Stanford Microarray Database[55]) and an integrated functional network (YeastNet[20]; http://www.yeastnet.org/) (**Supplementary Table 1**). Genetic interaction data contained precomputed Pearson correlations between all pairs of genetic interaction profiles of 4,417 genes. Expression data (5,053 genes, 1,683 arrays) were filtered to remove genes not present in the majority of arrays and arrays not covering at least 80% of genes, retaining 1,113 arrays and 4,660 genes for which pairwise Pearson correlation coefficient was computed.

To create **Figure 1b**,**c**, we selected the same number of top-scoring interactions from each data set to match the 62,885 physical protein-protein interactions in BioGRID for yeast (**Supplementary Table 1**). All remaining analyses were conducted using a single high-confidence interaction data set that included all physical protein-protein interactions from BioGRID that were observed in at least two independent studies, genetic interaction profiles with Pearson correlation coefficient ≥ 0.2 (cutoff as previously determined[14]) and gene co-expression profiles with Pearson correlation coefficient ≥ 0.93 (determined to provide the same enrichment for biological process co-membership as the genetic interactions with correlation ≥ 0.2), and high-scoring interactions from YeastNet (log-likelihood score (LLS) ≥ 3) that had not yet been covered by any of the previous networks (**Supplementary Tables 1** and **6**). We also attempted to use a single integrated network and corresponding LLS threshold based on YeastNet alone. However, YeastNet was released in 2007 and updating it with more recently published interaction data (e.g., in refs. 12,14) requires multiple nontrivial optimizations involving the original raw data sets (Insuk Lee, Yonsei University, Seoul, Korea, personal communication), which we elected not to attempt here.

Determining network support for specific and general GO terms. We defined the network density of a term t in the GO as $Density(t) = \frac{NE(t)}{NP(t)}$, where $NE(t)$ is the number of edges between genes assigned to t, and $NP(t)$ is the number of gene pairs in the term, that is, $\binom{n}{2}$, where n is the number of genes in t. To rule out the possibility that small terms in the ontology could account for the significant density of larger terms (**Fig. 1b**), we devised a permutation scheme that preserved the interactions within small terms (of size $k < k_0$), and permuted all other interactions while preserving node degrees (**Fig. 1c**).

Assembly of the NeXO. The goal is to construct a gene ontology $O = (G, M)$, where G is a directed acyclic graph in which nodes are terms and directed edges are parent→child term relations, and M: $t \rightarrow x$ is a function that maps each term t to a set of genes x. To construct an ontology from networks, we first identify a hierarchy of network communities using a probabilistic algorithm for community detection[32]. This method constructs a binary tree T in which the leaves are genes and each internal non-leaf node represents the join of two child nodes ($c1$, $c2$). The probability of the network data D given T is expressed as:

$$P(D \mid T) = \prod_{c1,c2 \in joins(T)} P_{c1,c2}$$

where the probability $P_{c1,c2}$ takes the form:

$$P_{c1,c2} = B(e_{c1,c2} + 1, h_{c1,c2} + 1)$$

where B is the Beta function, and $e_{c1,c2}$ and $h_{c1,c2}$ refer to the edges and non-edges (holes) crossing between genes assigned to the left subtree $c1$ and genes assigned to the right subtree. The Beta function scores highly if the gene pairs of interest are primarily edges or primarily holes, that is, have coherent behavior. A maximum-likelihood optimization of T is performed. We found that this method performs better in the ontology construction pipeline than several other standard hierarchical clustering algorithms[21–23] that could also be used to produce a binary tree T (**Supplementary Figs. 7** and **8**).

T serves as a computationally tractable approximation of the ontology graph G, but it artificially imposes that every term (except those at the root and leaves) connects to exactly two specialized terms and a single more general term (the root has two children terms and no parents, whereas leaves have a single parent term and no children). Hence, we modify T as follows. First, we test the degree to which each non-leaf node contributes to the overall score. Nodes that do not contribute to the score are removed such that the node's parent is connected directly to each of its children. The procedure is based on a previously proposed post-processing step that uses Bayesian model selection to remove nodes in the top and bottom layers of the tree[32]. In the present work this procedure is extended and applied to all nodes in the tree (first to the nodes joining genes, as proposed previously, and then to all other internal nodes). The criteria for removing an internal parent node p is based on evaluating the probability of the data under the original tree and under an updated version in which p is replaced by its children ($c1$, $c2$,..,cn) as expressed by the following ratio:

$$\lambda_{c1,c2,..,cn} = \prod_{s=1}^{K} \frac{P_{p,s}}{P_{c1,s}P_{c2,s} \cdots P_{cn,s}}$$

where s ($1 \le s \le K$) is one of K siblings of the node p. We chose to remove p and substitute its position by the children nodes when $\lambda_{c1,\ c2,..,cn} < 1$, additionally requiring that the interaction density between the children $c1$, $c2$,..,cn is not greater than that between p and its siblings. By replacing p with its children we create a multiway (>2) join associated with the parent of p.

Second, we provide a simple heuristic for supplementing T with new connections from child nodes c to second parents p when such relations are supported by the network data (such that T is no longer a tree but remains a directed acyclic graph). Starting from the leaves we iteratively consider all node pairs (c, p) such that the number of genes assigned to c is less than the number assigned to p. Node p is identified as an additional parent of c if:

1. Nodes p and c are not already on the same path or children of the same node.
2. There is a dense pattern of interactions connecting genes assigned to c and genes assigned to p ($Density \ge 0.3$; hypergeometric P-value < 0.05).
3. The sets of genes associated with p and c together form a dense cluster

$$(Density(p \cup C) \ge \frac{1}{2}(Density(p)).$$

The final result is a DAG T where non-leaf nodes correspond to terms and leaves correspond to genes. To return the NeXO ontology, G is defined as T minus its leaves. M is defined by mapping each term to the set of genes below it in T.

Preparation of GO. The GO OBO and annotation files were downloaded from http://www.geneontology.org/GO.downloads.ontology.shtml (files used to generate the results were current as of Dec. 19, 2011). We considered four basic types of GO relations: "is_a", "part_of", "regulates" and "has_part" and excluded relations annotated with a "NOT" clause. To prepare GO for ontology alignment against NeXO, we removed GO terms that did not have any (direct or indirect) gene annotations in *S. cerevisiae*. We also identified redundant GO terms that had the same gene content as their direct descendants. Because these terms may obscure information about the actual functional hierarchy in yeast (by imposing artificial layers in the hierarchy), we iteratively replaced each mutually redundant parent-child pair by the more specific term (the child in the parent-child relationship). In total, 709, 1,960 and 2,849 were retained in the GO Cellular Component, Molecular Function and Biological Process ontologies, respectively.

Ontology alignment. Given two ontologies, O_1 with n_1 terms and O_2 with n_2 terms, an ontology alignment A is a mapping of terms between ontologies such that each term in O_1 maps to at most one term in O_2, and vice versa. To provide a method for aligning gene ontologies in which the terms refer to sets of genes (technically, the set of genes assigned to a term defines the 'label' of that term), we developed an algorithm motivated by a previously proposed method called ASMOV[34] for aligning semantic ontologies. Term mapping in our alignment procedure is evaluated using a score function that considers the similarity of the sets of genes assigned to the terms (the so-called intrinsic

doi:10.1038/nbt.2463

term similarity) and the relative position of the terms in the hierarchy (relational similarity). The alignment process is iterative, with iteration k producing an ontology alignment A_k. First, an $n_1 \times n_2$ term similarity matrix T_k is calculated, where $0 \le T_k(i,j) \le 1$ represents the similarity of term $i \in O_1$ to term $j \in O_2$. $T_k(i,j)$ is composed of the intrinsic similarity $I(i,j)$ and the relational similarity $R_k(i,j)$:

$$T_k(i,j) = \begin{cases} I(i,j), & k = 0 \\ 0.75\, I(i,j) + 0.25 R_k(i,j), & k > 0 \end{cases}$$

$I(i,j)$ is taken as the Jaccard Index of the sets of genes assigned to i and j: $I(i,j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|}$. The matrix I is precomputed and does not change throughout the alignment process. The relational similarity $R_k(i,j)$ is calculated by determining the similarity between the sets of terms (P_i, P_j) that are the parents of i and j, and the similarity between the sets of terms (C_i, C_j) that are the children of i and j:

$$R_k(i,j) = \begin{cases} \frac{S(P_i,P_j) + S(C_i,C_j)}{2}, & \text{(internal nodes)} \\ S(C_i,C_j), & \text{(root)} \end{cases}$$

The set similarity S is calculated using the term similarity matrix T_{k-1} from the previous iteration:

$$S(X,Y) = \frac{SOS}{|X| + |Y| - SOS}$$

$$SOS = \sum_{(x,y) \in L} T_{k-1}(x,y)$$

where L is a local alignment of X to Y, determined by greedily choosing pairs (x,y) with the highest T_{k-1} while ensuring that each element of X or Y participates in no more than one pair. Based in $R_k(i,j)$ we can now calculate $T_k(i,j)$ and determine the new alignment A_k according to the following greedy algorithm:

0. Initializations: Initialize A_k as the empty mapping. Initialize L as a sorted list of term pairs (i,j) in decreasing order of $T_k(i,j)$.

1. Select the top pair (i,j) from L.

2. Check if (i,j) conflicts with any pair already contained in A_k. Two mappings (e_1, e_2) and (e_1', e_2') conflict if:

a. **Non-uniqueness:** $e_1 = e_1'$ or $e_2 = e_2'$.

b. **Parent-child crisscross:** Either e_1 is a descendant of e_1' in O_1 and e_2 is an ancestor of e_2' in O_2, or e_1' is a descendant of e_1 in O_1 and e_2' is an ancestor of e_2 in O_2.

3. If there is no conflict with any of the mappings already in A_k, add (i,j) to A_k.

4. Remove (i,j) from L.

5. If all of the elements of O_1 or O_2 are mapped, or L contains only mappings below a threshold similarity value (set to 0.01), then alignment A_k is complete. Otherwise, go to step 1.

If A_k matches a previous A_i ($i < k$), this mapping is returned as the final alignment and the final alignment score S_k for each term t is determined as:

$$S_k(t) = \begin{cases} T_k(t, A_k(t)), & \text{if } t \text{ is mapped by the alignment } A_k \\ O, & \text{otherwise} \end{cases}$$

Otherwise, the above algorithm is restarted at iteration k+1.

Calculating the false-discovery rate for ontology alignment. The false-discovery rate (FDR) of term alignment was calculated as:

$$FDR(t) = \frac{\frac{1}{n}\sum_{i=1}^{n} N_{R_i}(t)}{N(t)}$$

where $N_{R_i}(t)$ is the number of terms in the random permutation i that have an alignment score $\ge t$, and $N(t)$ is the number of terms in the actual computed ontology that have an alignment score $\ge t$. We set a minimum score threshold value $t \ge 0.1$ for large terms and higher threshold values for small terms so as to maintain an FDR < 10% within each size group (**Supplementary Fig. 9**).

Scoring robustness of NeXO terms. To determine the most confident terms in NeXO, we devised a measure of term quality that considers the network support of the term and its robustness to random perturbations of the input data. The network support $NS(t)$ for a term t is defined as the enrichment for interactions connecting genes assigned to the term (−log(P-value) estimated based on the hypergeometric distribution). The bootstrap score $B(t)$ for term t is calculated by randomly removing 5% of the edges in the input network and reconstructing a new bootstrapped ontology and aligning it to the original NeXO:

$$B(t) = \frac{1}{n\sum_{i=1}^{n} S_i(t)}$$

where S_i is the alignment score for term t when NeXO is aligned to the i-th bootstrapped ontology. The final robustness score for each term is calculated as a geometric mean of network support for the term and its bootstrap value:

$$R(t) = \sqrt{NS(t)B(t)}$$

This robustness score significantly enriches for terms in NeXO that align to GO terms (**Supplementary Fig. 10**) and thus is used to prioritize novel term candidates that are highly robust but are not yet present in GO.

Genetic interaction profiling. Genetic interaction profiling was done as previously described[56,57] based on a 73 × 741 (query × array) design. The 73 query genes were chosen arbitrarily from the entire set of 162 genes annotated at or beneath NeXO term 9965 (Golgi Apparatus). The 741 array genes were chosen to cover a representative sample of genes more broadly related to protein trafficking and lipid metabolism, including the 73 genes used as queries. Measured colony sizes for each double mutant (combining a deletion of a query gene with a deletion of an array gene) were processed to yield quantitative 'S-scores', indicating the degree to which the observed growth was greater than (positive S-score) or less than (negative S-score) expected[56]. The complete set of genetic interaction profiles for all 73 queries is given in **Supplementary Table 4**. The background set of genetic interaction profiles considered in **Fig. 4a,d** was drawn from unpublished data covering the entire space of 741 × 741 interactions.

54. Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* **39**, D698–D704 (2011).

55. Hubble, J. *et al.* Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Res.* **37**, D898–D901 (2009).

56. Collins, S.R., Roguev, A. & Krogan, N.J. Quantitative genetic interaction mapping using the E-MAP approach. *Methods Enzymol.* **470**, 205–231 (2010).

57. Schuldiner, M. *et al.* Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507–519 (2005).