

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Transformation of High-Throughput Data into Hierarchical Cellular Models Enables Biological Prediction and Discovery

Permalink

<https://escholarship.org/uc/item/6nm9r56v>

Author

Kramer, Michael Harris

Publication Date

2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

**Transformation of High-Throughput Data into Hierarchical
Cellular Models Enables Biological Prediction and Discovery**

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Biomedical Sciences

by

Michael Harris Kramer

Committee in charge:

Professor Trey Ideker, Chair
Professor Vineet Bafna
Professor Steve Briggs
Professor Kun-Liang Guan
Professor Gene Yeo

2016

Copyright

Michael Harris Kramer, 2016

All rights reserved

This Dissertation of Michael Kramer is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2016

DEDICATION

I dedicate this thesis to my family, whose support has been instrumental for me to reach this milestone. To my parents, for all the years of encouragement, patience and confidence. To my siblings, Matt and Lauren, for trying to keep me grounded and humble. To my grandparents, who always make it clear that no goal is too lofty. And to my wife, Stephanie, who does too much for me to list here but, simply put, makes every day a great day. I love you all.

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Table of Contents	v
List of Figures	vi
List of Tables	ix
Acknowledgements	x
Vita.....	xii
Abstract of the Dissertation	xiii
Chapter 1 Introduction	1
Chapter 2 A gene ontology inferred from molecular networks	8
Chapter 3 Inferring gene ontologies from pairwise similarity data	34
Chapter 4 Active Interaction Mapping reveals the hiearachical organization of autophagy	44
Chapter 5 Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems	89

LIST OF FIGURES

Chapter 1

Figure 1. Structure of the Gene Ontology	2
--	---

Chapter 2

Figure 1. Automated assembly and alignment of gene ontologies	9
---	---

Figure 2. The NeXO Ontology	10
-----------------------------------	----

Figure 3. Validation	11
----------------------------	----

Figure 4. Evaluation of protein trafficking terms using genetic interaction profiling	12
---	----

Figure 5. Updating GO with additional terms and relations	13
---	----

Supplementary Figure 1. Alignment results for NeXO computed without the YeastNet and co-expression networks	22
---	----

Supplementary Figure 2. Alignment results for NeXO using unfiltered networks	23
--	----

Supplementary Figure 3. Alignment results using GO with excluded high-throughput annotations	24
--	----

Supplementary Figure 4. New NeXO term associated with the retromer complex.....	25
---	----

Supplementary Figure 5. New term enriched for mutants sensitive to mercury chloride (HgCl ₂).....	26
---	----

Supplementary Figure 6. Distinguishing “is_a” and “part_of” relations in NeXO	27
---	----

Supplementary Figure 7. Representation of the proteasome when using other clustering methods in the NeXO approach.....	28
--	----

Supplementary Figure 8. Alignment results for NeXO using other hierarchical clustering methods.....	29
---	----

Supplementary Figure 9. Alignment score threshold at 10% FDR	30
--	----

Supplementary Figure 10. Distribution of robustness for aligned and unaligned terms ...	31
---	----

Chapter 3

Figure 1. CliXO Method	36
------------------------------	----

Figure 2. Inferring an ontology from semantic similarity.....	39
Figure 3. CliXO with Noise	39
Figure 4. Inferring an Ontology from –omics data	40
Figure 5. Stability of NeXO and CliXO with respect to parameters	41
Figure 6. CliXO Ontology from YeastNet v3.....	41
Chapter 4	
Figure 1. Program for Active Interaction Mapping	51
Figure 2. Assembly of hierarchical model from integrated data	53
Figure 3. First-generation autophagy ontology.....	54
Figure 4. Active analysis of data types leads to conditional genetic interaction maps.....	57
Figure 5. Second-generation ontology	61
Figure 6. Characterization of AtgO 2.0	63
Figure 7. Atg26 is a component of the Cvt pathway	69
Supplementary Figure 1. Performance degradation as interactions are removed.....	84
Chapter 5	
Figure 1. Patterns of Genetic Interaction Reflect the Hierarchical Structure of the GO ...	91
Figure 2. The Ontotype Method of Translating Genotype to Phenotype	92
Figure 3. Genome-wide Prediction of Pairwise Genetic Interactions in Yeast	93
Figure 4. The Functionalized GO	94
Figure 5. Elucidating the Genetic Logic of DNA Repair and the Nuclear Lumen.....	96
Figure 6. Prediction of Triple Mutants	97
Supplementary Figure 1. Prediction of pairwise genetic interactions under a stringent cross-validation setup	102
Supplementary Figure 2. Fuller characterization of genes in an ontology increases prediction accuracy	103

Supplementary Figure 3. Segregation of positive and negative interactions across F_{GO} .	104
Supplementary Figure 4. Genetic logic of the nuclear lumen	105
Supplementary Figure 5. Prediction performance versus the depth of decision trees	106
Supplementary Figure 6. Prediction of synthetic lethal interactions curated in the <i>Saccharomyces</i> Genome Database	107

LIST OF TABLES

Chapter 1

Supplementary Table 1. Input networks investigated in this study	32
--	----

Chapter 5

Table 1. Top New Functional Relationships in F_{GO}	95
---	----

ACKNOWLEDGEMENTS

I would like to acknowledge my thesis advisor, Trey Ideker, for countless hours of excellent advice and support during my PhD. I am continually amazed by his ability to pinpoint interesting questions in science and to quickly grasp where the kernel of a highly technical idea can be taken to make the most important contribution. I am a much better scientist because of him.

I would like to acknowledge the many members of the Ideker lab who have been friends and excellent colleagues during my PhD. There are many of you whose pointed questions and suggestions greatly influenced all stages of the work presented here. I would especially like to acknowledge Janusz Dutkowski, who led the initial charge into the world of data-driven ontologies, and Mike Yu, who has been involved in almost every crucial discussion and brainstorming session which has led to the work presented here.

I would like to acknowledge Jean-Claude Farré, whose guidance and patient teaching in the has been particularly appreciated and who is personally responsible for a large majority of my technical skill in the wet lab.

I would like to acknowledge my thesis committee, Vineet Bafna, Steve Briggs, Kun-Liang Guan and Gene Yeo for the helpful advice and thoughtful questions. I would like to acknowledge Steve Briggs for encouraging me from the beginning of my graduate work to be comfortable stepping out from behind the computer and getting my hands “wet”. I would like to acknowledge Vineet Bafna for spending a particularly significant amount of time providing incredibly insightful and invaluable algorithmic advice which was crucial for the development of the algorithms presented here.

Chapter 2, in full, is a reprint of the material as it appears in: Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T. “A gene ontology inferred from molecular networks.” *Nature Biotechnology*. 2013;31(1):38-45. The dissertation author was the second author of this work.

Chapter 3, in full, is a reprint of the material as it appears in: Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T. “Inferring gene ontologies from pairwise similarity data.” *Bioinformatics*. 2014;30(12):i34-42.. The dissertation author was the primary investigator and first author of this work.

Chapter 4, in full, is being revised and prepared for resubmission for publication. Kramer M*, Farre JC*, Mitra K, Yu MK, Licon K, Balakrishnan R, Cherry JM, Subramani S, Ideker T. “Active Interaction Mapping reveals the hierarchical organization of autophagy.” 2016. The dissertation author was the primary investigator and co-first author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in: Yu MK, Kramer M, Dutkowski J, Srivas R, Licon K, Kreisberg J, Ng CT, Krogan N, Sharan R, Ideker T. “Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems.” *Cell Systems*. 2016;2(2):77-88. The dissertation author was the second author of this work.

VITA

Education

- 2010 Bachelor of Arts, Engineering Sciences
Harvard College, Cambridge, MA
- 2016 Doctor of Philosophy, Biomedical Sciences
University of California, San Diego

PUBLICATIONS

Uthgenannt BA, **Kramer MH**, Hwu JA, Wopenka B, Silva MJ. Skeletal self-repair: stress fracture healing by rapid formation and densification of woven bone. *Journal of Bone and Mineral Research*. 2007;22(10):1548-56. doi: 10.1359/jbmr.0070614.

Kramer MH, Breydo E, Shubina M, Babcock K, Einbinder JS, Turchin A. Prevalence and factors affecting home blood pressure documentation in routine clinical care: a retrospective study. *BMC Health Services Research*. 2010;10:139. doi: 10.1186/1472-6963-10-139.

Dutkowski J, **Kramer M**, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T. A gene ontology inferred from molecular networks. *Nature Biotechnology*. 2013;31(1):38-45. doi: 10.1038/nbt.2463.

Haynes BC, Maier EJ, **Kramer MH**, Wang PI, Brown H, Brent MR. Mapping functional transcription factor networks from gene expression data. *Genome Research*. 2013;23(8):1319-28. doi: 10.1101/gr.150904.112.

Dutkowski J, Ono K, **Kramer M**, Yu M, Pratt D, Demchak B, Ideker T. NeXO Web: the NeXO ontology database and visualization platform. *Nucleic Acids Research*. 2014;42(Database issue):D1269-74. doi: 10.1093/nar/gkt1192.

Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T. Inferring gene ontologies from pairwise similarity data. *Bioinformatics*. 2014;30(12):i34-42. doi: 10.1093/bioinformatics/btu282.

Yu MK, **Kramer M**, Dutkowski J, Srivas R, Licon K, Kreisberg J, Ng CT, Krogan N, Sharan R, Ideker T. Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems. *Cell Systems*. 2016;2(2):77-88. doi: 10.1016/j.cels.2016.02.003.

ABSTRACT OF THE DISSERTATION

Transformation of High-Throughput Data into Hierarchical
Cellular Models Enables Biological Prediction and Discovery

by

Michael Harris Kramer

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2016

Professor Trey Ideker, Chair

A holy grail of bioinformatics is the creation of whole-cell models with the ability to enhance human understanding and facilitate discovery. To this end, a successful and widely-used effort is the Gene Ontology (GO), a massive project to manually annotate genes into terms describing molecular functions, biological processes and cellular components and provide relations between terms, e.g. capturing that “small ribosomal

subunit” and “large ribosomal subunit” come together to make “ribosome”. GO is widely used to understand the function of a gene or group of genes. Unfortunately, GO is limited by the effort required to create and update it by hand. It exists only for well-studied organisms and even then in one, generic form per organism with limited overall genome coverage and bias towards well-studied genes and functions. It is not possible to learn about an uncharacterized gene or discover a new function using GO, and one cannot quickly assemble an ontology model for a new organism, cell-type or disease-state.

Here we change this state of affairs by developing and utilizing the concept of purely data-driven gene ontologies. In chapter two, we show that large networks of gene and protein interactions in *Saccharomyces cerevisiae* can be used to computationally infer a data-driven ontology whose coverage and power are equivalent to those of the manually-curated GO. In chapter three we further develop the algorithmic foundations for data-driven ontologies, laying the groundwork for machine learning to intelligently integrate many types of experimental data into ontology models. In chapter four, we focus on a cellular process (autophagy in *Saccharomyces cerevisiae*) and develop a framework (Active Interaction Mapping) which guides experimental selection, systematically improves an existing process-specific ontology model and uncovers new autophagy biology. Finally, in chapter five, we illustrate the power of hierarchical whole-cell ontology models for biological modeling by demonstrating an ontology-based framework for translation of genotype to phenotype.

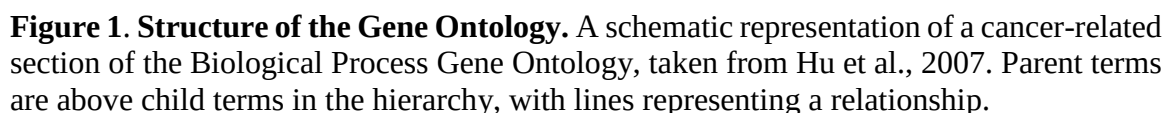
Overall, this work provides a roadmap to construct data-driven, hierarchical models of gene function for the whole cell or a specific cellular process and illustrates the power of these models for both discovery of new biology and biological modeling.

CHAPTER 1

INTRODUCTION

Whole cell models with the ability to illuminate useful biology have long been a primary goal of the field of bioinformatics (Tomita, 2001; Tomita et al., 1999). One of the most successful and widely-used efforts to this end has been the Gene Ontology (GO) project, an exhaustive effort to represent the functions of genes in cellular level processes across organisms (Ashburner et al., 2000; Gene Ontology, 2001). GO is "the default source of functional annotations for virtually every experimental system and the gold standard for measuring the success of bioinformatic methods" (Dolinski and Botstein, 2013). It is extensively utilized by researchers in a wide variety of situations, from those studying specific questions of basic biology to those trying to interpret the possible effect of a gene mutation discovered in a GWAS study (Holmans et al., 2009; Wang et al., 2010). For those who have performed genome-wide experiments, overlap of a set of experimentally discovered genes with genes annotated to a given GO term is the gold standard method for determining the function of such a gene set (Huang da et al., 2009; Khatri and Draghici, 2005; Maere et al., 2005).

The GO project actually consists of two distinct but complementary efforts - one to construct a set of terms which represent the inner workings of the generic eukaryotic cell, and another to assign genes to those terms for specific organisms. For the first effort, GO curators have constructed and currently maintain and update three general sets of terms which form three distinct ontologies designed to capture different aspects of gene function - Molecular Function, Cellular Component, and Biological Process. Each of these ontologies is a hierarchical set of terms, such that more general terms are "parents" to less



Several issues, however, have arisen as GO has grown to its current, expansive state (it now encompasses >38,000 unique terms and >64,000 term-term relations which are

used to annotate >80 different species) (Fraser and Marcotte, 2004; Leonelli et al., 2011). First, despite stringent curation standards, it has been difficult to maintain consistency in how terms are created and genes are annotated across domains of expertise, as different expert curators are responsible for different sections and species of GO. Second, there is a strong bias in the GO in favor of coverage and annotation of well-studied functions and processes and away from the annotation and coverage of terms representing newly discovered or less well-studied functions and processes. Third, it simply takes a massive amount of manual effort to create, update and maintain both the ontologies and the species-specific annotations. This has led to a system where the GO annotations for various species are kept up-to-date to varying degrees. Even in a best case scenario, researchers are able to support only a single version of GO annotations for each organism – there is no such thing as a cell-type specific, condition specific or disease-state specific GO. All of these issues lead to problems when trying to use GO ranging from the insertion of bias into the training of a newly developed bioinformatic method with GO as a gold standard, to a researcher missing the true functional effect of a treatment condition under examination because the area of GO representing that function has not been updated, to a researcher working on a less well-studied organism being required to infer function solely based on orthology with genes in another organism's GO.

It has previously been suggested that many of these problems with GO could be addressed by an alternative to manually curated ontology construction - direct construction of ontologies from data (Fraser and Marcotte, 2004). This goal, however, has proven elusive despite recognition by those involved in the original founding of GO that "a scalable, automated approach would eventually be required because of the size and

complexity of the ontology" (Dolinski and Botstein, 2013). Still, there has been some progress in related tasks or tasks which form a subset of those required to create a de novo ontology directly from data. There have been relatively successful attempts to infer functional relationships between individual genes (Costello et al., 2009; Kim et al., 2008; Lee et al., 2010; Lee et al., 2004; Lee et al., 2008; Lee et al., 2007). There have been efforts to infer the correct placement of unannotated genes into the existing Gene Ontology hierarchy (Pena-Castillo et al., 2008; Yan et al., 2010). There have been numerous algorithms created to "cluster" genes hierarchically into functional groups based on network data (Eisen et al., 1998; Girvan and Newman, 2002; Park and Bader, 2011; Zhang and Horvath, 2005). This hierarchy, however, is generally flattened with a cut at a given level and isolated gene sets are then examined individually for enriched overlap with GO terms (D'Haeseleer, 2005; Gibbons and Roth, 2002; Khatri and Draghici, 2005). There have even been some attempts to show that smaller clusters of genes can sometimes be grouped into larger, biologically meaningful clusters (Dotan-Cohen et al., 2009; Jaimovich et al., 2010; Kelley and Ideker, 2005; Ravasz et al., 2002; Tanay et al., 2004). However, there has not previously been a comprehensive method to create a purely data-based, hierarchical, whole cell ontology model including both the hierarchical relationships between terms and assignments of genes to terms. Here we will examine the potential to do just this. We will also explore several potential uses of these hierarchical cell models, including 1) use of a data-driven hierarchical model to guide interaction mapping of yeast autophagy and extract high-value biologically relevant information from a new high-throughput experimental dataset, and 2) the usage of the hierarchical cellular models created to translate genotype to phenotype.

REFERENCES

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25, 25-29.
- Costello, J.C., Dalkilic, M.M., Beason, S.M., Gehlhausen, J.R., Patwardhan, R., Middha, S., Eads, B.D., and Andrews, J.R. (2009). Gene networks in *Drosophila melanogaster*: integrating experimental data to predict gene function. *Genome biology* 10, R97.
- D'Haeseleer, P. (2005). How does gene expression clustering work? *Nature biotechnology* 23, 1499-1501.
- Dolinski, K., and Botstein, D. (2013). Automating the construction of gene ontologies. *Nature biotechnology* 31, 34-35.
- Dotan-Cohen, D., Letovsky, S., Melkman, A.A., and Kasif, S. (2009). Biological process linkage networks. *PloS one* 4, e5313.
- Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95, 14863-14868.
- Fraser, A.G., and Marcotte, E.M. (2004). A probabilistic view of gene function. *Nature genetics* 36, 559-564.
- Gene Ontology, C. (2001). Creating the gene ontology resource: design and implementation. *Genome research* 11, 1425-1433.
- Gibbons, F.D., and Roth, F.P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome research* 12, 1574-1581.
- Girvan, M., and Newman, M.E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* 99, 7821-7826.
- Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Wellcome Trust Case-Control, C., Owen, M.J., O'Donovan, M.C., and Craddock, N. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American journal of human genetics* 85, 13-24.
- Hu, P., Bader, G., Wigle, D.A., and Emili, A. (2007). Computational prediction of cancer-gene function. *Nature reviews Cancer* 7, 23-34.

- Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 37, 1-13.
- Jaimovich, A., Rinott, R., Schuldiner, M., Margalit, H., and Friedman, N. (2010). Modularity and directionality in genetic interaction maps. *Bioinformatics* 26, i228-236.
- Kelley, R., and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nature biotechnology* 23, 561-566.
- Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587-3595.
- Kim, W.K., Krumpelman, C., and Marcotte, E.M. (2008). Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome biology* 9 Suppl 1, S5.
- Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M., and Rhee, S.Y. (2010). Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nature biotechnology* 28, 149-156.
- Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. *Science* 306, 1555-1558.
- Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., and Marcotte, E.M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nature genetics* 40, 181-188.
- Lee, I., Li, Z., and Marcotte, E.M. (2007). An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PloS one* 2, e988.
- Leonelli, S., Diehl, A.D., Christie, K.R., Harris, M.A., and Lomax, J. (2011). How the gene ontology evolves. *BMC bioinformatics* 12, 325.
- Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448-3449.
- Park, Y., and Bader, J.S. (2011). Resolving the structure of interactomes with hierarchical agglomerative clustering. *BMC bioinformatics* 12 Suppl 1, S44.
- Pena-Castillo, L., Tasan, M., Myers, C.L., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W.K., Krumpelman, C., Tian, W., Obozinski, G., Qi, Y., Mostafavi, S., Lin, G.N., Berriz, G.F., Gibbons, F.D., Lanckriet, G., Qiu, J., Grant, C., Barutcuoglu, Z., Hill, D.P., Warde-Farley, D., Grouios, C., Ray, D., Blake, J.A., Deng, M., Jordan, M.I., Noble, W.S., Morris, Q., Klein-Seetharaman, J., Bar-Joseph, Z., Chen, T., Sun, F., Troyanskaya, O.G., Marcotte, E.M., Xu, D., Hughes, T.R., and Roth, F.P.

(2008). A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome biology* 9 Suppl 1, S2.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabasi, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551-1555.

Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the United States of America* 101, 2981-2986.

Tomita, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends in biotechnology* 19, 205-210.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T.S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., and Hutchison, C.A., 3rd (1999). E-CELL: software environment for whole-cell simulation. *Bioinformatics* 15, 72-84.

Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature reviews Genetics* 11, 843-854.

Yan, H., Venkatesan, K., Beaver, J.E., Klitgord, N., Yildirim, M.A., Hao, T., Hill, D.E., Cusick, M.E., Perrimon, N., Roth, F.P., and Vidal, M. (2010). A genome-wide gene function prediction resource for *Drosophila melanogaster*. *PloS one* 5, e12139.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* 4, Article17.

CHAPTER 2

Reprint of “A gene ontology inferred from molecular networks.” Published January 2013 in *Nature Biotechnology*.

A gene ontology inferred from molecular networks

Janusz Dutkowski¹, Michael Kramer¹, Michal A Surma^{2,3}, Rama Balakrishnan⁴, J Michael Cherry⁴, 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 genes³. 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 terms⁸ 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 suggested⁴). High-throughput measurements of genetic and protein interactions^{11–15} and mRNA expression profiles¹⁶, 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 YeastNet²⁰ (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

¹Departments of Medicine and Bioengineering, University of California San Diego, La Jolla, California, USA. ²Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. ³Department of Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, California, USA. ⁴Department of Genetics, Stanford University, Stanford, California, USA. ⁵J. David Gladstone Institutes, San Francisco, California, USA. ⁶Institute 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

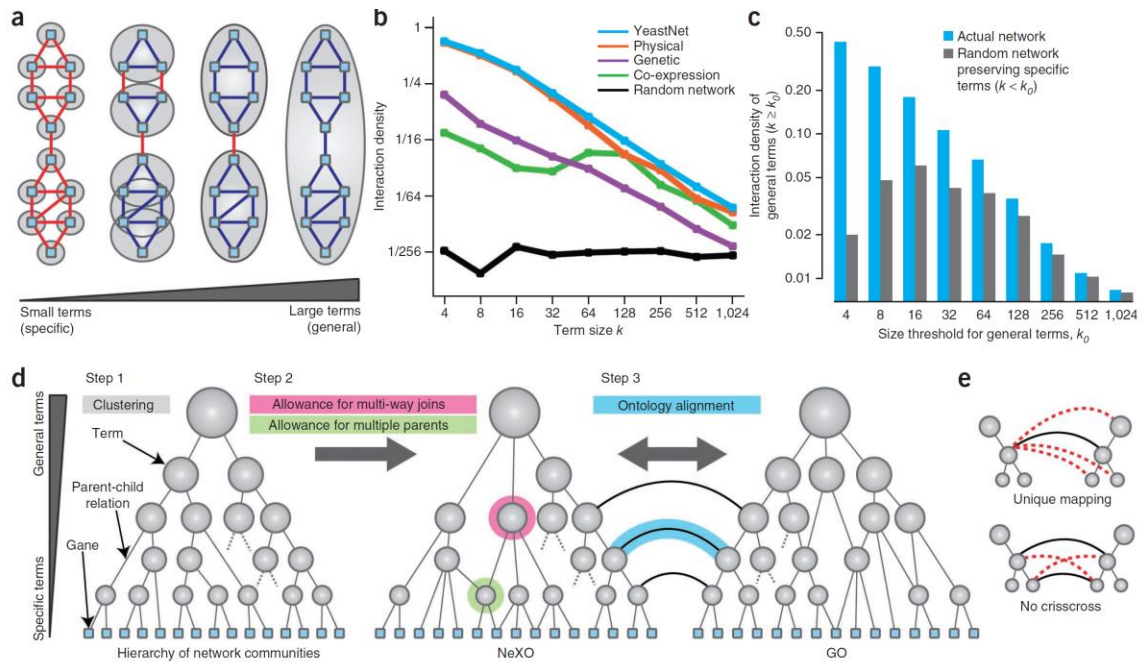


Figure 1 Automated assembly and alignment of gene ontologies. (a) Specific GO terms correspond to small gene communities in an interaction network, which are nested within larger communities corresponding to more general terms. Gray ovals represent terms, blue squares represent genes and links represent 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 sciences³⁴, and matches terms between ontologies based on similar gene assignments and similar positions in

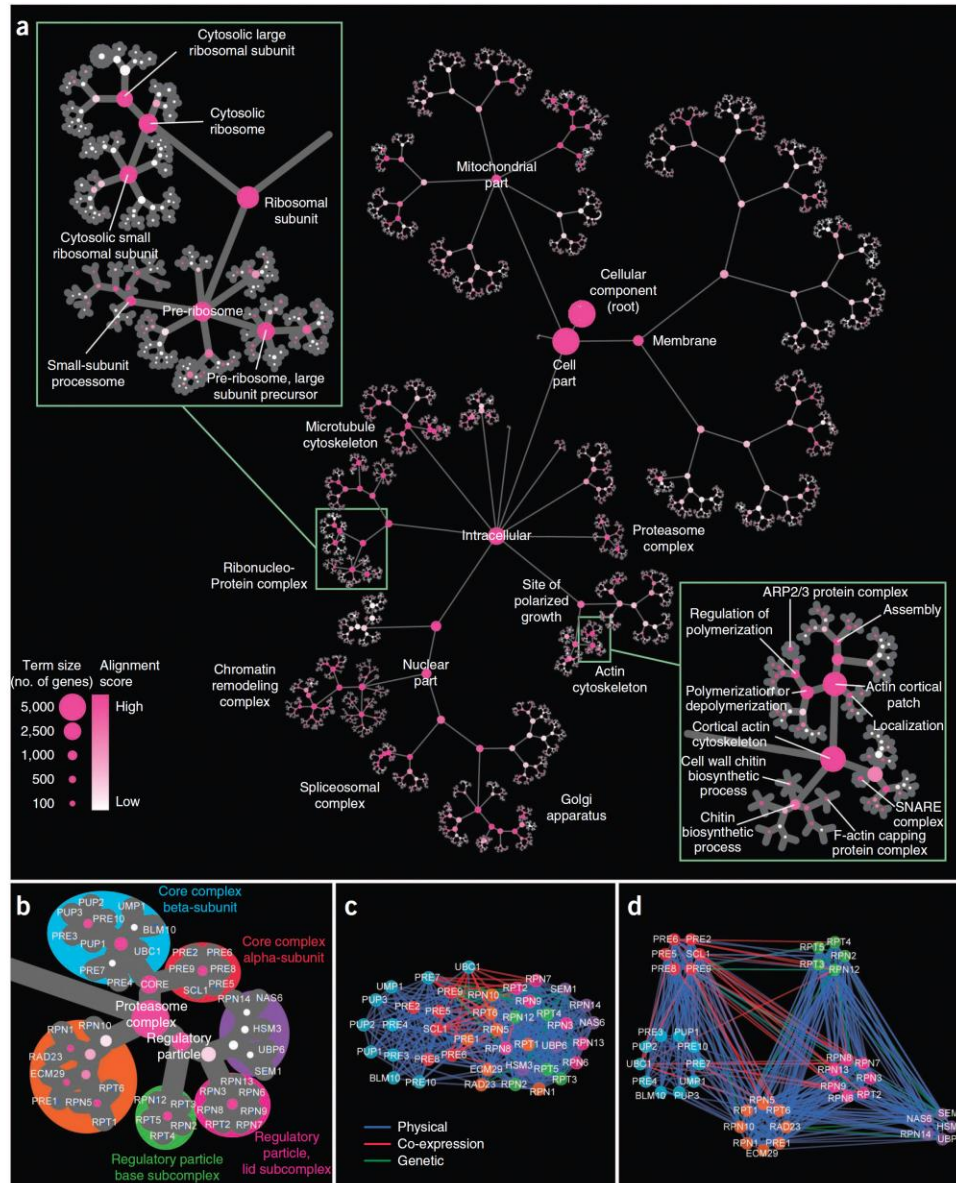


Figure 2 The NeXO ontology. (a) The ontology is shown as a tree, with nodes indicating terms and edges indicating hierarchical relations between terms. Node sizes indicate the number of genes assigned to a term. Node colors represent the degree of correspondence to a term in GO as determined by ontology 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 enrichment³⁵, 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

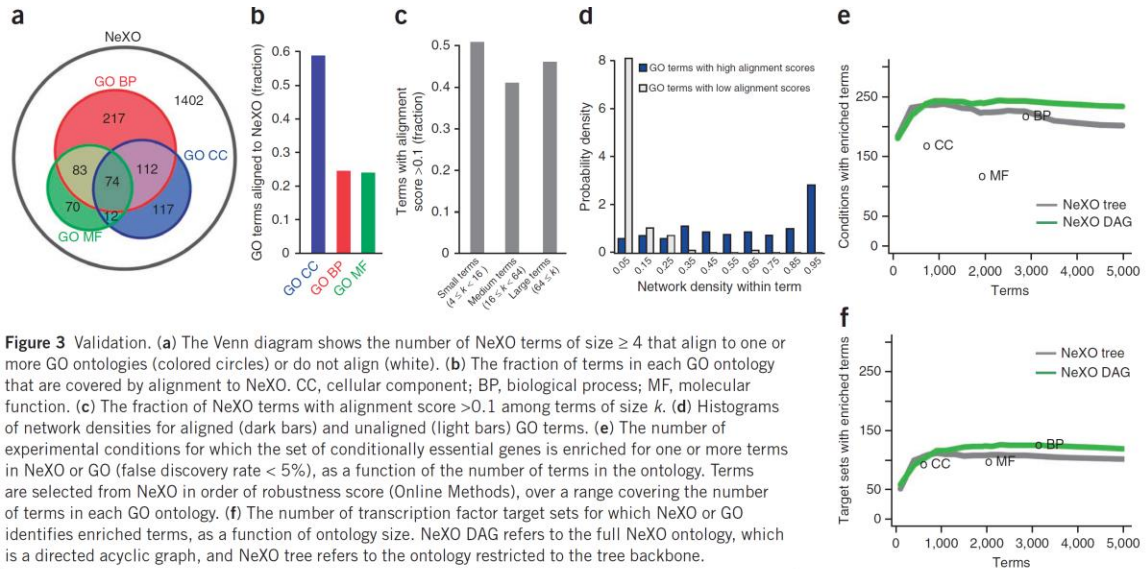


Figure 3 Validation. (a) The Venn diagram shows the number of NeXO terms of size ≥ 4 that align to one or more GO ontologies (colored circles) or do not align (white). (b) The fraction of terms in each GO ontology that are covered by alignment to NeXO. CC, cellular component; BP, biological process; 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 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

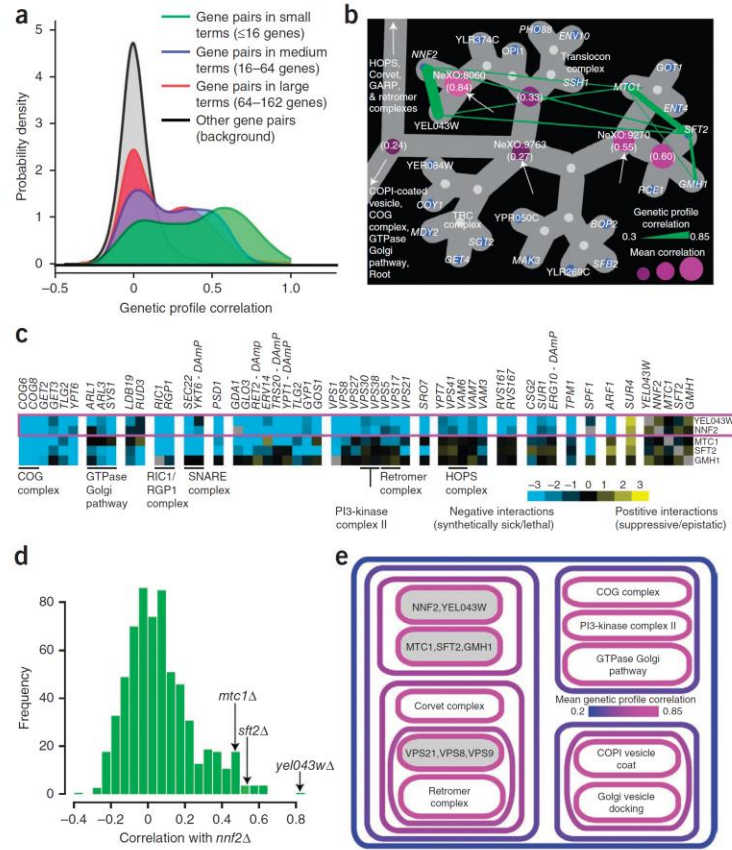
good network support. We considered that some of this observed correspondence between NeXO and GO might occur by construction, since cut-off thresholds for two of the input networks (YeastNet and co-expression) had been optimized to connect genes with similar GO Biological Process annotations. However, 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 conditions³⁶ and a database of direct gene targets for each of 183 transcription factors³⁷. 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 *nrf2Δ*, *yel043wΔ*, *mtc1Δ*, *sft2Δ* and *gmh1Δ*. Yellow and blue correspond to positive and negative interactions, respectively. (d) Distribution of correlation coefficients of the *nrf2Δ* genetic interaction profile versus a selection of ~750 other genes, highlighting the *yel043wΔ* 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 descendant 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 *nrf2Δ* and *yel043wΔ* 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 *nrf2Δ* profile is most correlated with that of *yel043wΔ* (**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 network³⁸, 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 *nrf2Δ* and *yel043wΔ* 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

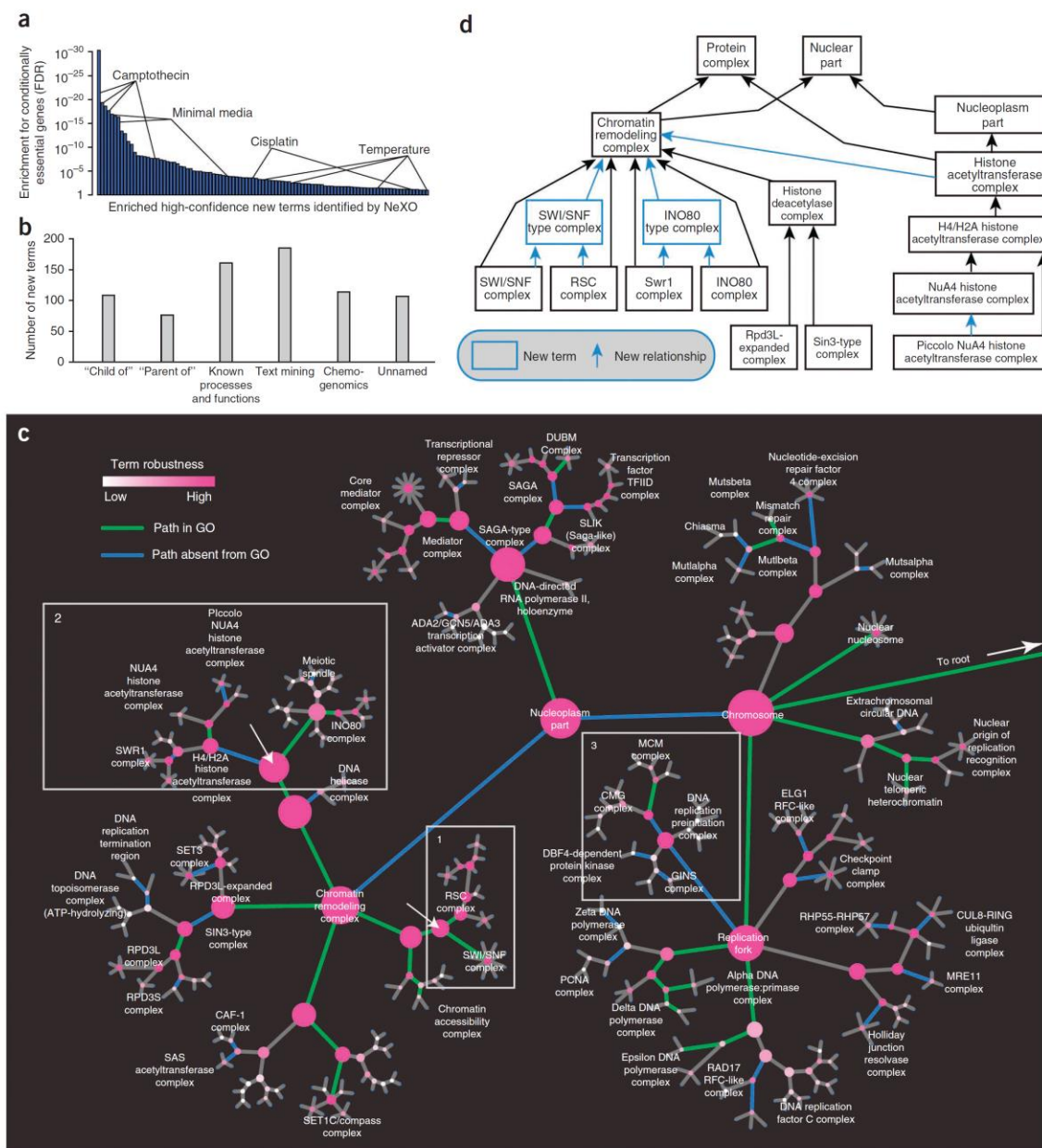


Figure 5 Updating GO with additional terms and term relations. (a) Robust terms enriched for genes necessary for growth under an experimental condition (false discovery rate < 10%). Top-scoring conditions for selected terms are indicated. (b) The number of terms derived using NeXO that can be named based 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 capping⁴¹ (**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 (HgCl₂) (**Supplementary Fig. 5**), an agent that promotes oxidative stress and has been found to induce apoptosis through a mitochondrial-dependent pathway⁴². 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 Database⁴⁵ 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 yeast⁴⁶ 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 literature⁴⁷. An additional NeXO-derived term also joins SWR1 and NUA4, two chromatin-modifying complexes with overlapping functions and components⁴⁸, which have been proposed to function as the single Tip60 complex in humans^{49,50} including homologs of the INO80 complex⁵⁰. 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 complex⁵² (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 associations⁵³.

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/dofinder/10.1038/nbt.2463>.

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

- Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
- Musen, M.A. *et al.* The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* **19**, 190–195 (2012).
- Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
- Fraser, A.G. & Marcotte, E.M. A probabilistic view of gene function. *Nat. Genet.* **36**, 559–564 (2004).
- Leonelli, S., Diehl, A.D., Christie, K.R., Harris, M.A. & Lomax, J. How the gene ontology evolves. *BMC Bioinformatics* **12**, 325 (2011).
- Krallinger, M., Leitner, F. & Valencia, A. Analysis of biological processes and diseases using text mining approaches. *Methods Mol. Biol.* **593**, 341–382 (2010).
- 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).
- 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).
- Buitelaar, P. & Cimiano, P. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, Vol. 167 (IOS Press, Amsterdam, 2008).
- 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).
- Collins, S.R. *et al.* Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6**, 439–450 (2007).
- Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
- Tarassov, K. *et al.* An *in vivo* map of the yeast protein interactome. *Science* **320**, 1465–1470 (2008).
- Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
- Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. *Science* **333**, 601–607 (2011).
- 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).
- Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
- Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
- Myers, C.L. *et al.* Discovery of biological networks from diverse functional genomic data. *Genome Biol.* **6**, R114 (2005).
- 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).
- 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).
- Girvan, M. & Newman, M.E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821–7826 (2002).
- Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
- Khatiri, P. & Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21**, 3587–3595 (2005).
- D'haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **23**, 1499–1501 (2005).
- Gibbons, F.D. & Roth, F.P. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* **12**, 1574–1581 (2002).
- Ravasz, E., Somera, A.L., Mongru, D.A., Olvai, Z.N. & Barabasi, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
- Dotan-Cohen, D., Letovsky, S., Melkman, A.A. & Kasif, S. Biological process linkage networks. *PLoS ONE* **4**, e5313 (2009).
- Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomics data. *Proc. Natl. Acad. Sci. USA* **101**, 2981–2986 (2004).
- Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566 (2005).
- Jaimovich, A., Rinott, R., Schuldiner, M., Margalit, H. & Friedman, N. Modularity and directionality in genetic interaction maps. *Bioinformatics* **26**, i228–i236 (2010).
- Park, Y. & Bader, J.S. Resolving the structure of interactomes with hierarchical agglomerative clustering. *BMC Bioinformatics* **12** (suppl.1), S44 (2011).
- Clauset, A., Moore, C. & Newman, M.E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).
- Jean-Mary, Y.R., Shironoshita, E.P. & Kabuka, M.R. Ontology Matching with Semantic Verification. *Web Semant.* **7**, 235–251 (2009).
- 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).
- 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).
- Abdelrehman, 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).
- Seaman, M.N. Recycle your receptors with retromer. *Trends Cell Biol.* **15**, 68–75 (2005).
- Nickerson, D.P., Brett, C.L. & Merz, A.J. Vps-C complexes: gatekeepers of endolysosomal traffic. *Curr. Opin. Cell Biol.* **21**, 543–551 (2009).
- 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).
- Addinall, S.G. *et al.* A genome-wide suppressor and enhancer analysis of *cdc13-1* reveals varied cellular processes influencing telomere capping in *Saccharomyces cerevisiae*. *Genetics* **180**, 2251–2266 (2008).
- Araragi, S. *et al.* Mercuric chloride induces apoptosis via a mitochondrial-dependent pathway in human leukemia cells. *Toxicology* **184**, 1–9 (2003).
- Saretzki, G. Telomerase, mitochondria and oxidative stress. *Exp. Gerontol.* **44**, 485–492 (2009).
- Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
- Cherry, J.M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).
- Hayes, M.J., Bryon, K., Satkuranathan, 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).
- Clapier, C.R. & Cairns, B.R. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).
- 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).
- 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).
- 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).
- Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
- Evin, 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).
- 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 (BioGRID⁵⁴), synthetic-lethal and epistatic genetic interactions (DRYGIN database¹⁴; <http://drygin.ccb.utoronto.ca/>), co-expression relationships (Stanford Microarray Database⁵⁵) and an integrated functional network (YeastNet²⁰; <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 determined¹⁴) 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 $\rightarrow$ 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 detection³². 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 (c_1, c_2). The probability of the network data D given T is expressed as:

$$P(D|T) = \prod_{c_1, c_2 \in \text{joins}(T)} P_{c_1, c_2}$$

where the probability P_{c_1, c_2} takes the form:

$$P_{c_1, c_2} = B(e_{c_1, c_2} + 1, h_{c_1, c_2} + 1)$$

where B is the Beta function, and e_{c_1, c_2} and h_{c_1, c_2} refer to the edges and non-edges (holes) crossing between genes assigned to the left subtree c_1 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 algorithms²¹⁻²³ 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 tree³². 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 ($c_1, c_2, \dots, c_n$) as expressed by the following ratio:

$$\lambda_{c_1, c_2, \dots, c_n} = \frac{\prod_{s=1}^K P_{p,s}}{P_{c_1,s} P_{c_2,s} \dots P_{c_n,s}}$$

where s ($1 \leq s \leq K$) is one of K siblings of the node p . We chose to remove p and substitute its position by the children nodes when $\lambda_{c_1, c_2, \dots, c_n} < 1$, additionally requiring that the interaction density between the children $c_1, c_2, \dots, c_n$ 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 ($Density \geq 0.3$; hypergeometric P -value < 0.05).
3. The sets of genes associated with p and c together form a dense cluster

$$(Density(p \cup c) \geq \frac{1}{2}(Density(p))).$$

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 ASMOV³⁴ 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

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 $0 \leq T_k(i, j) \leq 1$ 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 :

$$I(i, j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|}. \text{ The matrix } I \text{ is precomputed and does not change through-}$$

out 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 :

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

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:

$$S_k(t) = \begin{cases} T_k(t, A_k(t)), & \text{if } t \text{ is mapped by the alignment } A_k \\ 0, & \text{otherwise} \end{cases}$$

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 $\geq t$, and $N(t)$ is the number of terms in the actual computed ontology that have an alignment score $\geq t$. We set a minimum score threshold value $t \geq 0.1$ 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 ($-\log(P\text{-value})$ 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 $\times$ 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) expected⁵⁶. 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).

Supplementary Materials for

A gene ontology inferred from molecular networks

Janusz Dutkowski, Michael Kramer, Michal A Surma, Rama Balakrishnan, J Michael Cherry, Nevan J Krogan & Trey Ideker

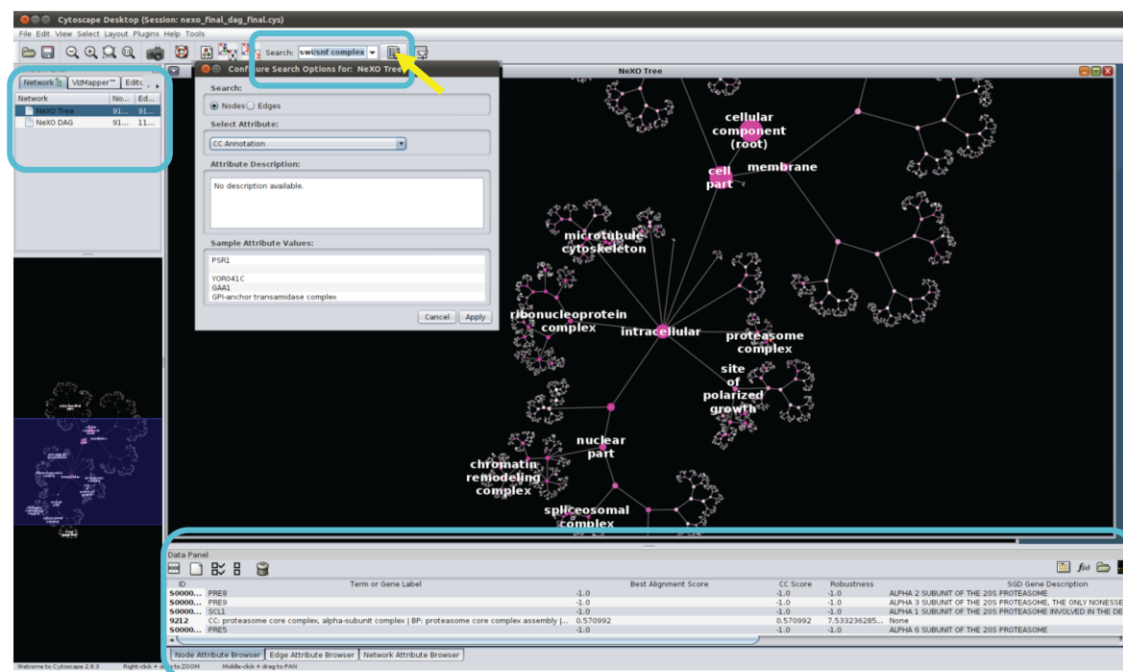
1. Supplementary Note 1. NeXO User Manual: Navigation with Cytoscape and OBO-Edit
2. Supplementary Figure 1. Alignment results for NeXO without YeastNet and expression networks
3. Supplementary Figure 2. Alignment results for NeXO using unfiltered networks
4. Supplementary Figure 3. Alignment results using GO with excluded high-throughput annotations
5. Supplementary Figure 4. New NeXO term associated with the retromer complex
6. Supplementary Figure 5. New term enriched for mutants sensitive to mercury chloride (HgCl_2)
7. Supplementary Figure 6. Distinguishing “is_a” and “part_of” relations in NeXO
8. Supplementary Figure 7. Representation of the proteasome with other clustering methods
9. Supplementary Figure 8. Alignment results for NeXO using other hierarchical clustering methods
10. Supplementary Figure 9. Alignment score threshold at 10% FDR
11. Supplementary Figure 10. Distribution of robustness for aligned and unaligned terms
12. Supplementary Table 1. Input networks investigated in this study
13. Supplementary Tables 2-6 (Excel spreadsheet files provided separately)
14. Supplementary File 1. NeXO in Cytoscape Format (File NeXO.cys provided separately)
15. Supplementary File 2. NeXO in Open Biomedical Ontology (OBO) Format (File NeXO.obo provided separately)

NeXO USER MANUAL

Navigation with Cytoscape

The NeXO.cys (Supplementary File 1) is a Cytoscape session file that can be loaded directly into Cytoscape version 2.8.3 (available from <http://www.cytoscape.org/>) to allow for visualization and interactive analysis of the entire ontology. The session provides three versions of NeXO: the tree, displaying the backbone of the ontology, the DAG where supplementary term-term relations are added, and also the DAG with additional gene-term assignments. The following steps illustrate an exemplary analysis using NeXO.cys and Cytoscape.

1. **Launch Cytoscape** and use the **File→Open** menu to locate and select the file NeXO.cys.
2. Select the **NeXO Tree** from the Network tab of the Control Panel at left (below screenshot).



3. Select the menu item **View→Show Graphics Details** (if the menu reads **Hide Graphics Details**, this command is already active). NeXO is visualized with high-level terms labeled via alignment to the GO Cellular Component ontology. Node size scales with the number of genes assigned at or below that term. Node color intensity scales with the Best Alignment Score to GO. See above screenshot and Fig. 2 in the main text.
4. **Select parts of the tree using the mouse** (e.g. the proteasome subtree). Information on selected nodes (terms or genes) are shown in the table in the Data Panel. The main attributes (columns) are as follows. Further details about the derivation of each attribute are provided in Supplementary Methods.

ID

The NeXO identifier (terms) or SGD identifier (genes) of the node.

Term or Gene Label

For terms, names are transferred from GO alignment (scores > 0.1, CC = Cellular Component, BP = Biological Process, MF = Molecular Function). For genes, common names are shown or, if unavailable, the ORF is shown.

CC Score

The alignment score vs. GO Cellular Component (defaults to -1.0 for genes).

Best Alignment Score

Maximum alignment score against the CC, BP, and MF ontologies of GO.

Robustness

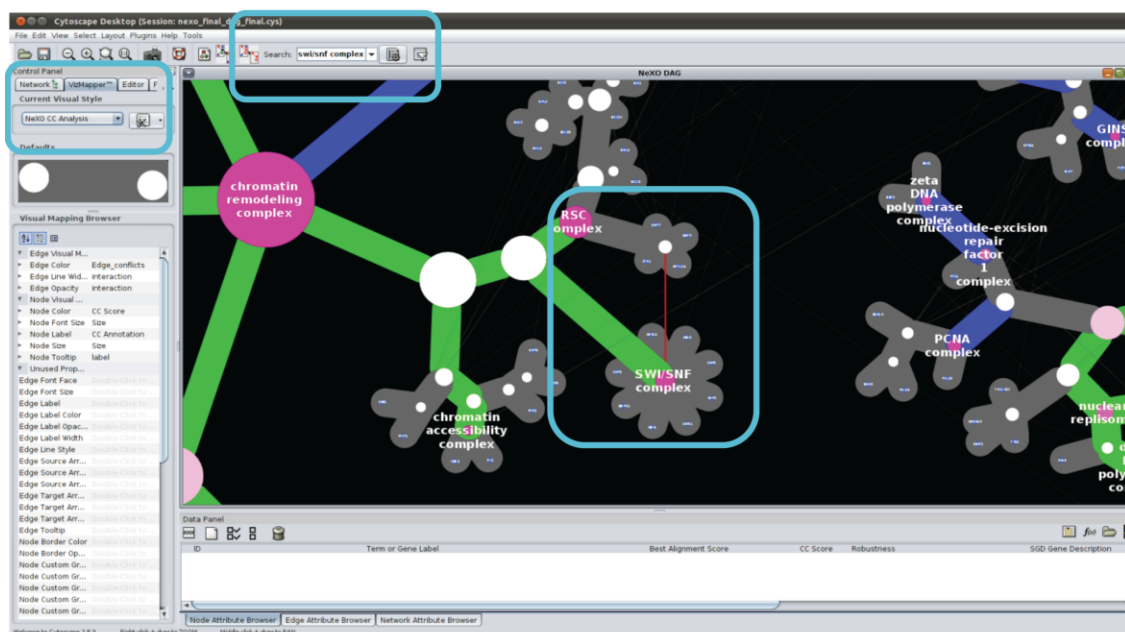
The term robustness score (Supplementary Methods, defaults to -1.0 for genes).

SGD Gene Description

Provides the free-form Gene Description field from SGD (genes only).

Additional attributes such as term definitions (CC, BP, MF), term size (number of assigned genes) and interaction density can also be loaded using the **Data Panel**.

5. Select the **VizMapper** tab of the Control Panel (below screenshot) and switch from the **NeXO Overview** to the **NeXO CC Analysis** visual style. This display will color the nodes according to their alignment score against the GO Cellular Component ontology. Green paths indicate consistent term-term relations between the two ontologies, i.e. ancestor-descendant pairs that are aligned to GO terms that are also in an ancestor-descendant relation. In contrast, a blue edge (parent to child) indicates a relation present in NeXO but absent from GO, such that the GO terms aligned to the child and the nearest aligned NeXO ancestor are not in a descendant-ancestor relation in GO. Only NeXO terms with high alignment score (CC score ≥ 0.2) are considered in this relation analysis – other terms are neutral for determining agreements and conflicts (e.g., a green path may pass through them).
6. Search NeXO for cellular components or genes of interest. Use the mouse to push the button to the right of the **Search field** (above screenshot) to configure search options. Select the **CC Annotation** attribute and push **Apply**. Type “swi/snf” into the search field and press **Enter**. Gene names are also searchable using this attribute. The SWI/SNF complex will be shown.
7. Select the **NeXO DAG** or **NeXO DAG (filtered)** from the Network tab of the Control Panel and repeat steps 3-6. In the full DAG version of the ontology, edges are added which connect child terms to additional parents (see below example, in which a subunit of RSC is connected to the SWI/SNF complex). Similar analysis can also be performed using NeXO DAG with additional gene-term assignments.



Navigation with OBO-Edit

The NeXO ontology can also be explored using designated ontology editors and browsers such as OBO-Edit (available from <http://www.oboedit.org/>) as shown in the below screenshot. For this purpose we have provided the file NeXO.obo (Supplementary File 2) which conforms to the Open Biological and Biomedical Ontology (OBO) format. Using OBO-Edit the NeXO ontology can be explored in three basic ways:

- Using the Ontology Tree Editor
- Using the Graph Editor
- By searching for keywords appearing in NeXO term annotations.

Terms in the NeXO.obo file are annotated with text labels transferred through alignment to GO. The label indicates matchings with score ≥ 0.1 to any one of the three GO ontologies: Cellular Component (CC), Biological Process (BP), and Molecular Function (MF). Additional information about the term such as its robustness score, interaction density, and alignment scores are provided in the Comment field.

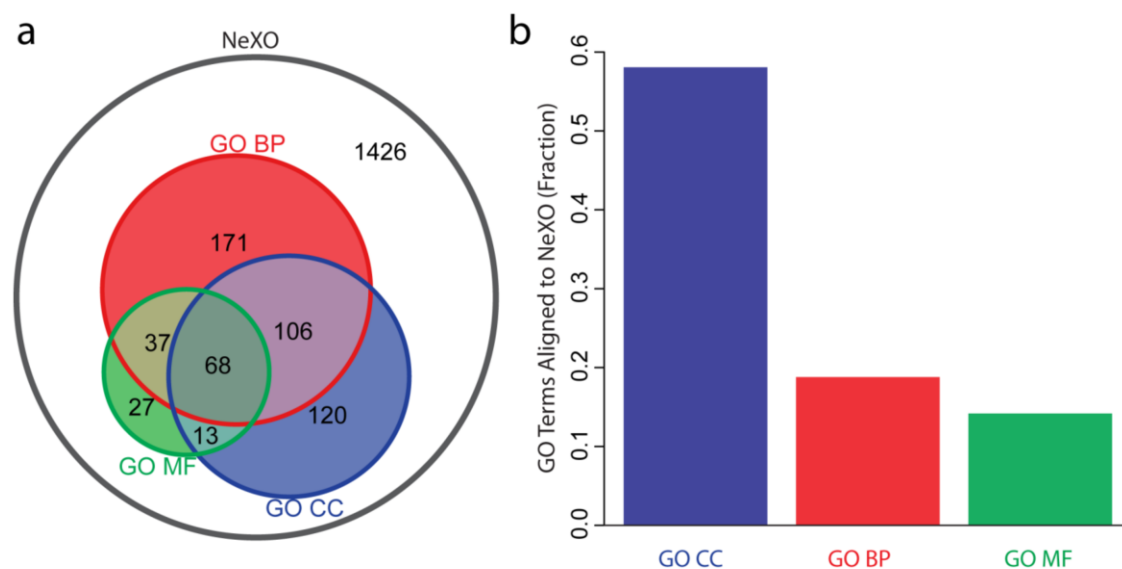
Each term relationship is annotated with one of two types: “child_of” or “linked_to”. The “X child_of Y” relationship indicates that X is a child of Y in the tree forming the backbone of the NeXO ontology. “X linked to Y” indicates that X was linked to Y by an additional edge in the NeXO DAG.

The screenshot displays the OBO-Edit application window, which is divided into several panes. The top pane shows the 'Ontology Tree Editor' with a hierarchical tree of ontology terms. The middle pane shows the 'Graph Editor' with a network diagram of terms and their relationships. The right pane shows the 'Text Editor' with a search results window and a detailed view of a specific term.

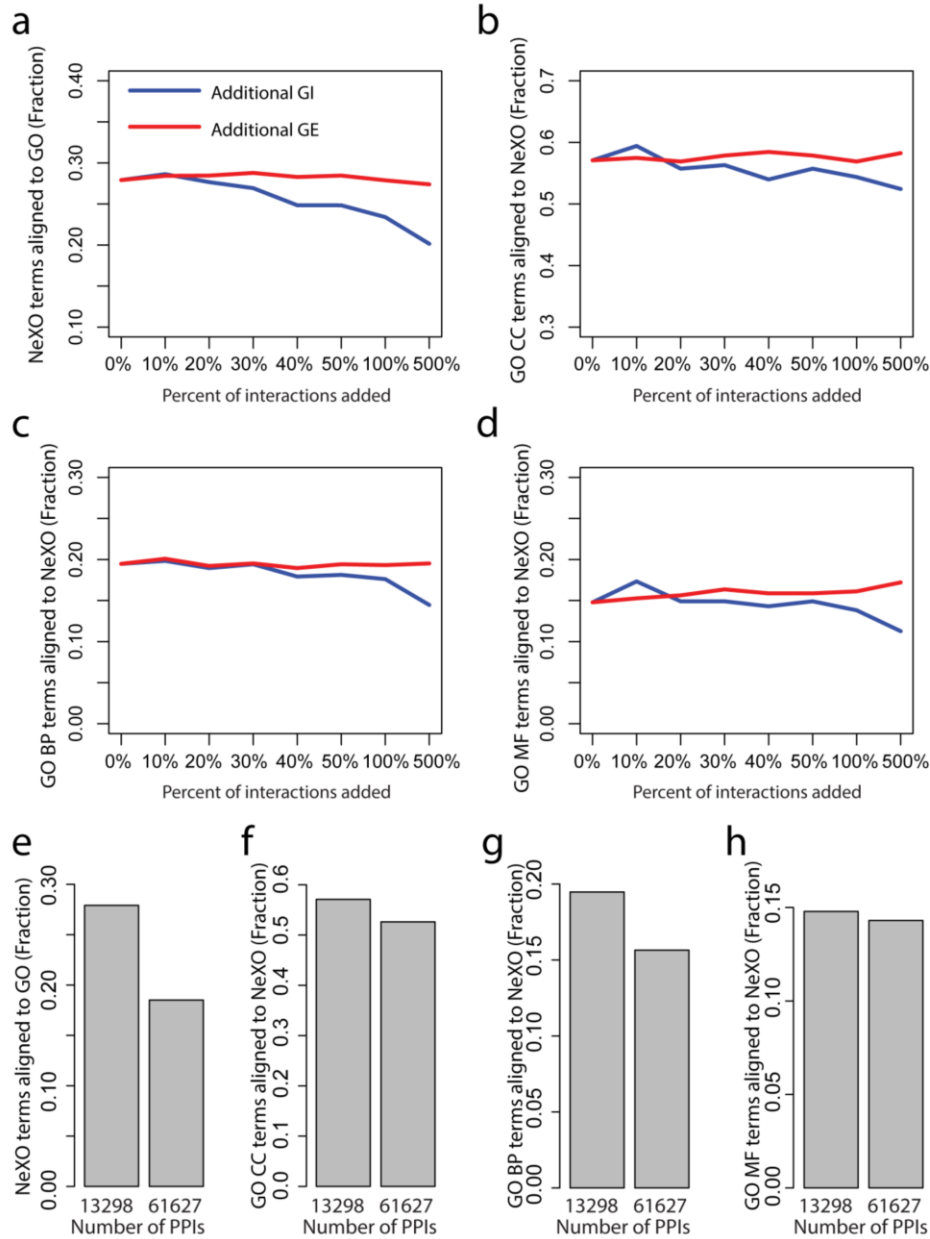
Ontology Tree Editor: The tree shows a hierarchy of terms. The root is 'CC: chromatin remodeling complex | BP: nucleosome organization | MF: ...'. Below it are terms like 'CC: RSC complex | BP: nucleosome disassembly | MF: DNA-dependent ATPase activity' and 'CC: SWI/SNF complex | BP: nucleosome mobilization | MF: positive regulation of transcription, DNA-dependent'.

Graph Editor: The graph shows terms as nodes connected by edges. The nodes are labeled with their full names, including the CC, BP, and MF annotations. The edges represent the 'child_of' or 'linked_to' relationships.

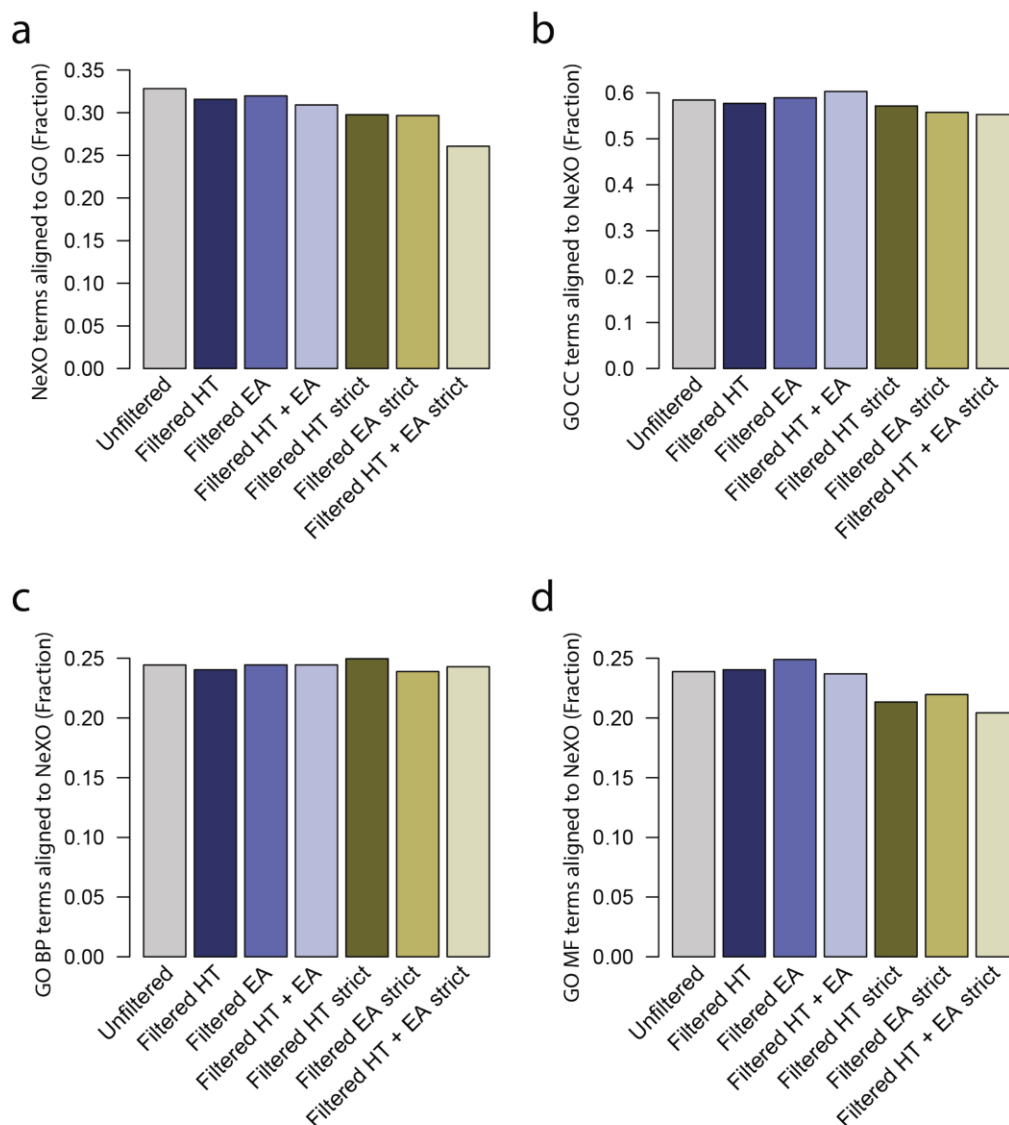
Text Editor: The search results window shows the results of a search for 'swi/snf'. The results list the terms and their annotations. The detailed view of a term shows its name, namespace, and comment. The comment includes the robustness score, interaction density, and alignment scores.



Supplementary Figure 1. Alignment results for NeXO computed without the YeastNet and co-expression networks. (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. The overlap with GO (in particular GO Cellular Components) closely matches the results obtained with the original NeXO that included information from

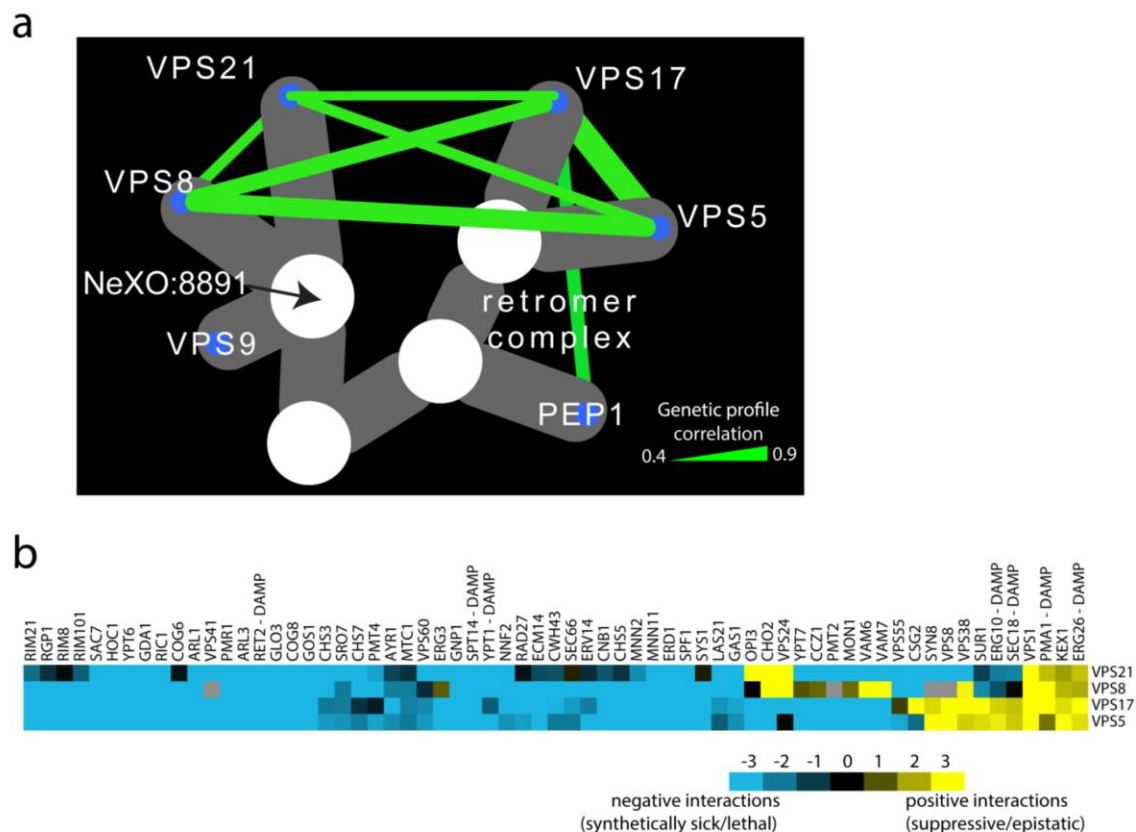


Supplementary Figure 2. Alignment results for NeXO using unfiltered networks. The percent of NeXO terms that aligned to GO (a), and GO terms that aligned to NeXO (b-d) when NeXO is built using more permissive interaction networks with 0-500% more genetic (GI) or co-expression (GE) interactions. The percent of NeXO terms that aligned to GO (e), and GO terms that aligned to NeXO (f-h) when NeXO is built using the original network and a more permissive interaction network which includes all 61627 physical protein-protein interactions (PPIs) in BioGRID. We observe that alignment results are robust across a wide range of more permissive input networks. Note that in all cases the input networks contain PPI, GI, and GE data, but exclude YeastNet to make sure that GO is not used for network training.

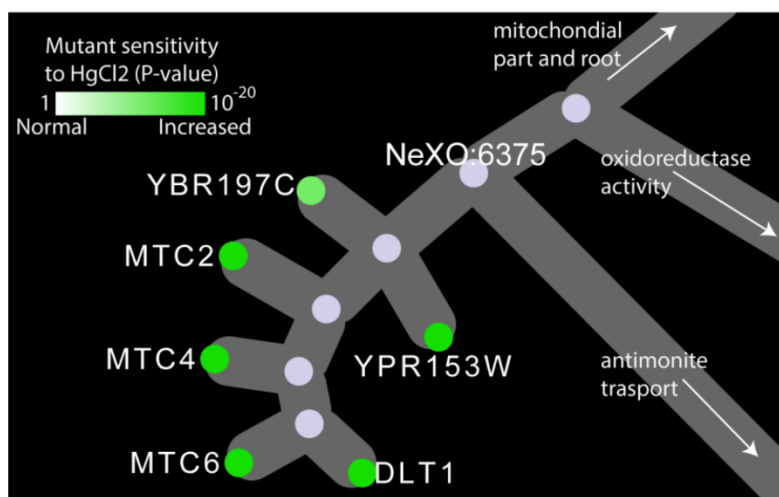


Supplementary Figure 3. Alignment results using GO with excluded high-throughput annotations.

The percent of NeXO terms that aligned to GO (a), and GO terms that aligned to NeXO (b-d) when gene-to-term association based on high-throughput interaction data (HT) and/or those inferred from electronic annotations (EA) are removed from GO. The specific evidence codes related to high-throughput interaction data (HT) are: IPI - Inferred from Physical Interaction; IEP - Inferred from Expression Pattern; and IGI - Inferred from Genetic Interaction. In the ‘strict’ version we removed all associations of gene X to term Y if at least one of these associations met the condition for removal. In the non-strict version gene X could still be associated with term Y if it had at least one evidence code that did not meet the condition for removal. NeXO alignment results appear robust to removal of HT and EA annotations from GO.

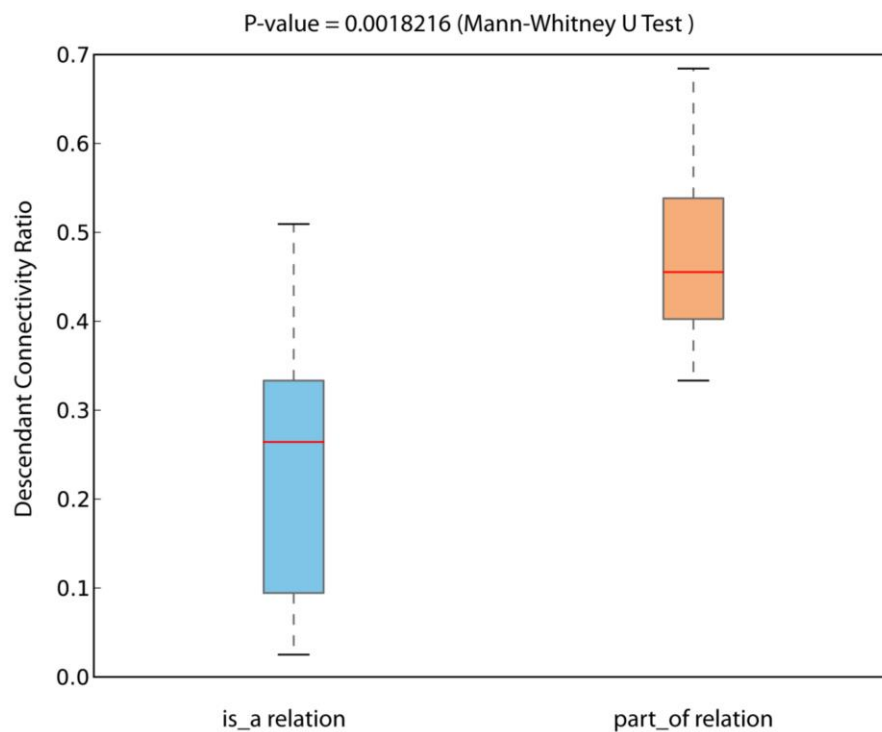


Supplementary Figure 4. New NeXO term associated with the retromer complex. New term NeXO:8891, composed of *VPS8*, *VPS21* and *VPS9*, and placed directly next to the retromer complex in NeXO (a). The new genetic interaction screen identified very high profile correlation between genes in the new term (*VPS8*, *VPS21*) and the retromer subunits *VPS17* and *VPS5* (a,b), providing further support for the position of the new term next to the retromer complex in the NeXO ontology.

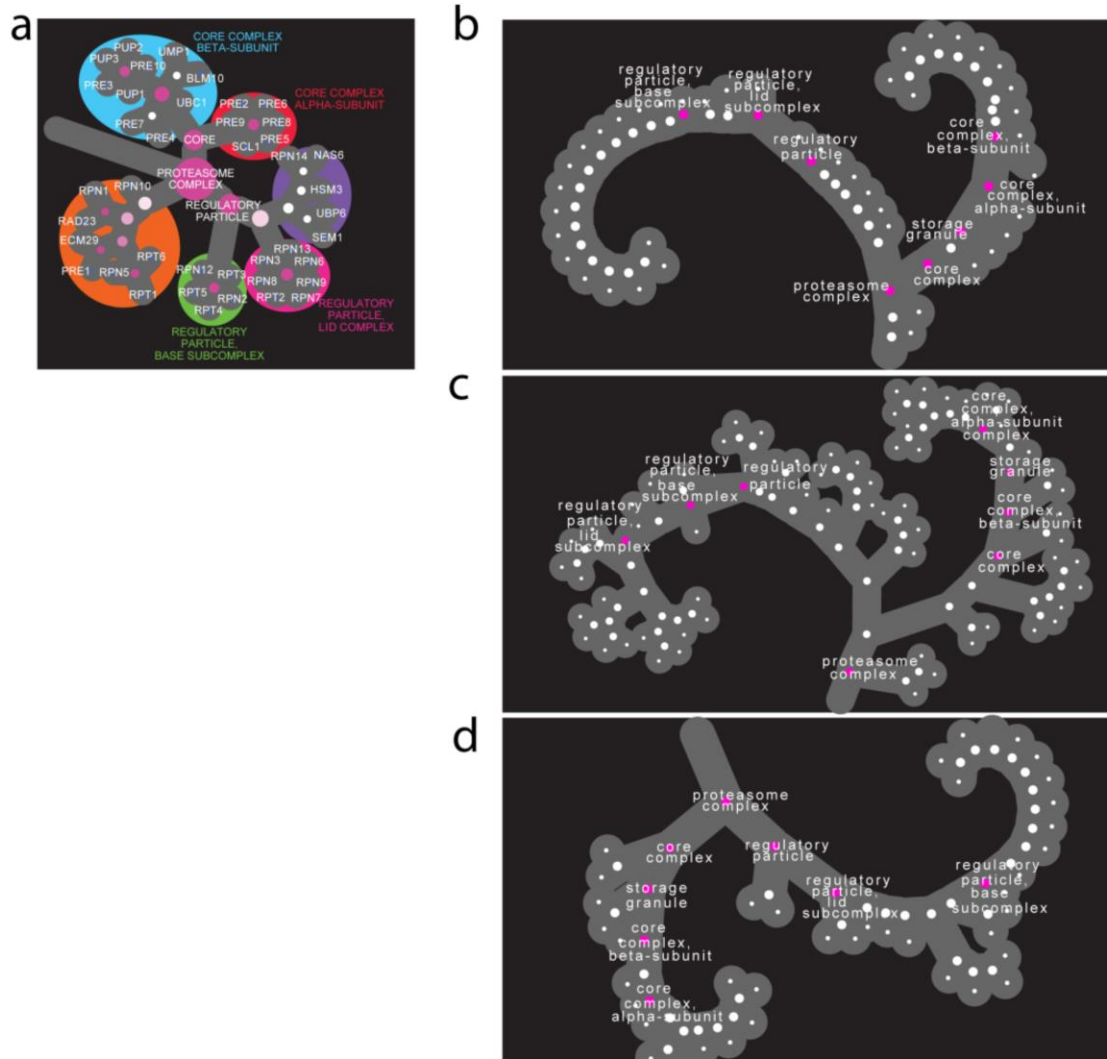


Supplementary Figure 5. New term enriched for mutants sensitive to mercury chloride (HgCl₂).

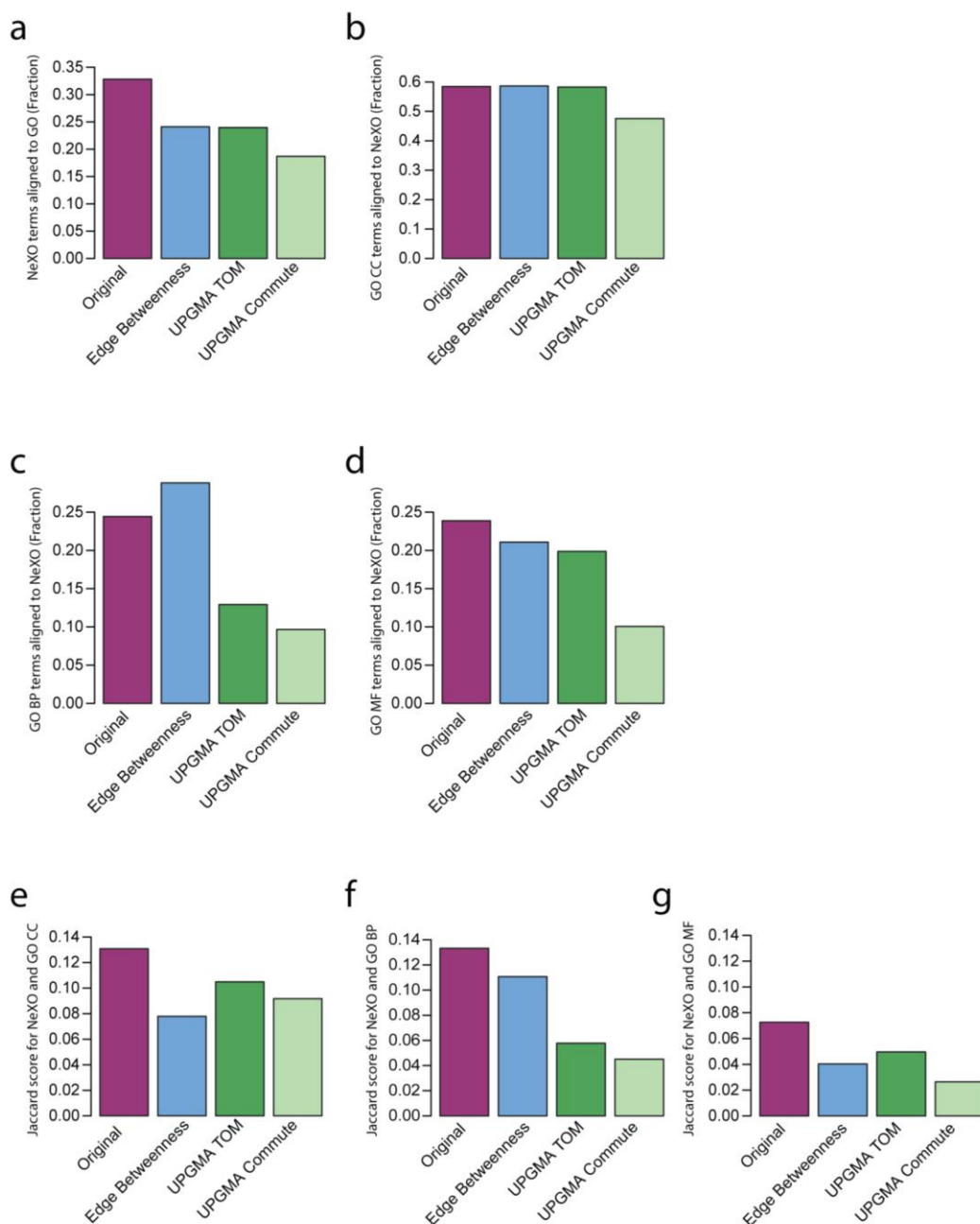
The term NeXO:6375 is composed of six poorly uncharacterized ORFs: *MTC2*, *MTC4*, *MTC6*, *DLT1*, *YBR197C* and *YPR153W*. NeXO places this terms under the mitochondrial component next to genes annotated with oxidoreductase activity and antimonite transport. Deletions of each of 6 genes in this term make yeast sensitive to mercury chloride (HgCl₂).



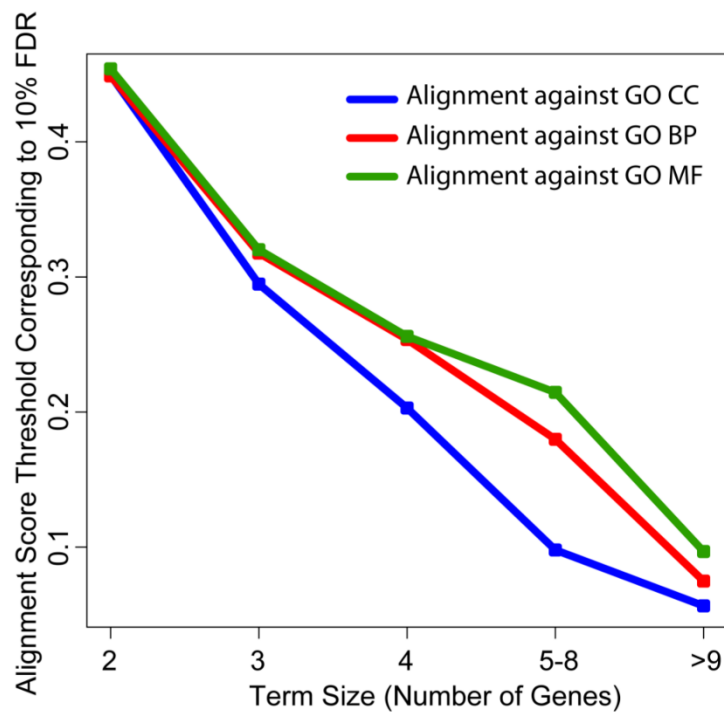
Supplementary Figure 6. Distinguishing “is_a” and “part_of” relations in NeXO. For each parent-child relation in NeXO that aligned to a parent-child relation in GO, the parent was examined to compute its Child Connectivity Ratio (CCR). This was the ratio of the number of molecular interactions spanning different children of the parent to the total number of interactions falling within the parent. CCR values for parent-child relations in NeXO mapping to “is_a” relations in GO are significantly lower than for NeXO relations mapping to “part_of” relations.



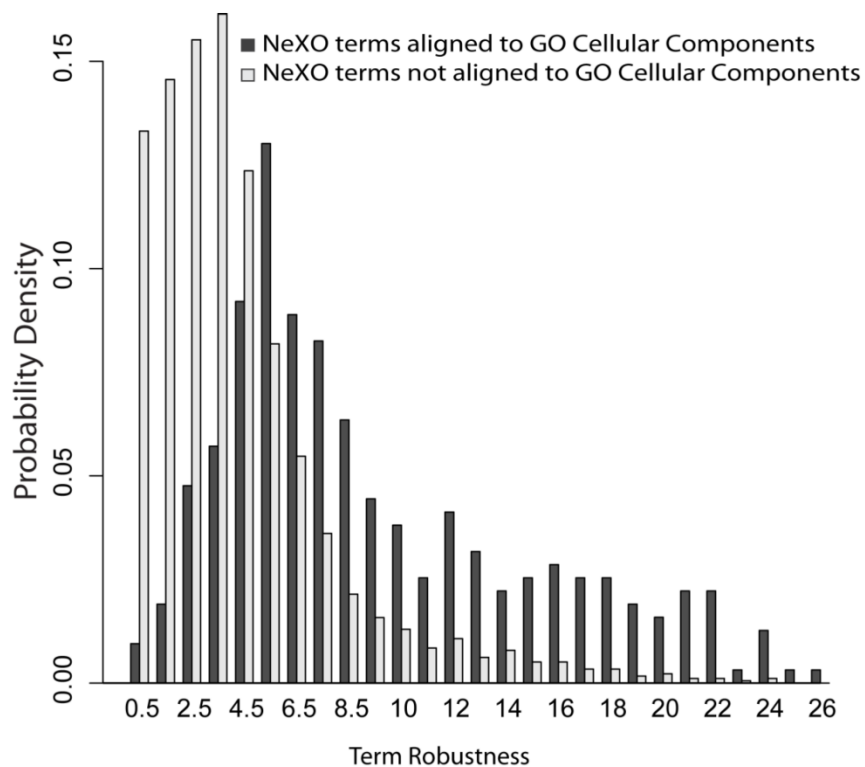
Supplementary Figure 7. Representation of the proteasome when using other clustering methods in the NeXO approach: original (a), edge betweenness (b), UPGMA based on topological overlap distance (c), and UPGMA based on commute distance (d). NeXO alignment is performed to map the resulting trees to the GO Cellular Component ontology. The alternative clustering methods (b-d) do not recover the hierarchy to the same extent as the original method used by NeXO (a). While core and regulatory particles are correctly identified, organization of more specific components is incorrect (b-d).



Supplementary Figure 8. Alignment results for NeXO using other hierarchical clustering methods. Percent of NeXO terms that aligned significantly ($FDR = 10\%$) to any of the three GO ontologies (a), percent of terms in each of the GO ontologies that aligned to NeXO (b-d). Jaccard index comparing the set of terms in NeXO to the set of GO Cellular Components (e), GO Biological Processes (f), and GO Molecular Functions (g). The original method achieves better agreement with GO than other standard methods.



Supplementary Figure 9. Alignment score threshold at 10% FDR. In the ontology alignment of NeXO to GO, different alignment score thresholds were used for different NeXO term sizes in order to maintain a false discovery rate of 10%. CC = Cellular Component, BP = Biological Process, MF = Molecular Function.



Supplementary Figure 10. Distribution of robustness for aligned and unaligned terms. The distributions are well separated for robustness > 5.

Supplementary Table 1. Input networks investigated in this study

Network Type	Used in Figs. 1B, C		HQ used in all other analyses	
	Number of Genes	Number of Interactions	Number of Genes	Number of Interactions
Physical Protein-Protein	5,658	61,627	3,401	13,298
Co-Expression	3,915	61,627	228	1,705
Genetic	4,396	61,627	3,090	11,240
YeastNet	5,268	61,627	3,351	10,573
Integrated High-Quality (HQ)	--	--	5,051	29,789

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of the material as it appears in: Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T. “A gene ontology inferred from molecular networks.” *Nature Biotechnology*. 2013;31(1):38-45. The dissertation author was the second author of this work.

CHAPTER 3

Reprint of “Inferring gene ontologies from pairwise similarity data.” Published June 15, 2014 in the *ISMB* issue of *Bioinformatics*.

Inferring gene ontologies from pairwise similarity data

Michael Kramer¹, Janusz Dutkowski¹, Michael Yu¹, Vineet Bafna² and Trey Ideker^{1,*}

¹Department of Medicine and ²Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA

ABSTRACT

Motivation: While the manually curated Gene Ontology (GO) is widely used, inferring a GO directly from -omics data is a compelling new problem. Recognizing that ontologies are a directed acyclic graph (DAG) of terms and hierarchical relations, algorithms are needed that:

- (1) analyze a full matrix of gene–gene pairwise similarities from -omics data;
- (2) infer true hierarchical structure in these data rather than enforcing hierarchy as a computational artifact; and
- (3) respect biological pleiotropy, by which a term in the hierarchy can relate to multiple higher level terms.

Methods addressing these requirements are just beginning to emerge—none has been evaluated for GO inference.

Methods: We consider two algorithms [Clique Extracted Ontology (CliXO), LocalFitness] that uniquely satisfy these requirements, compared with methods including standard clustering. CliXO is a new approach that finds maximal cliques in a network induced by progressive thresholding of a similarity matrix. We evaluate each method’s ability to reconstruct the GO biological process ontology from a similarity matrix based on (a) semantic similarities for GO itself or (b) three -omics datasets for yeast.

Results: For task (a) using semantic similarity, CliXO accurately reconstructs GO (>99% precision, recall) and outperforms other approaches (<20% precision, <20% recall). For task (b) using -omics data, CliXO outperforms other methods using two -omics datasets and achieves ~30% precision and recall using YeastNet v3, similar to an earlier approach (Network Extracted Ontology) and better than LocalFitness or standard clustering (20–25% precision, recall).

Conclusion: This study provides algorithmic foundation for building gene ontologies by capturing hierarchical and pleiotropic structure embedded in biomolecular data.

Contact: tideker@ucsd.edu

1 INTRODUCTION

Ontologies have proven very useful for capturing and organizing knowledge as a hierarchical set of terms and their interrelationships. In biology, one of the most successful and widely used ontologies is from the Gene Ontology (GO) project, a major effort to represent gene functions in cellular level processes across organisms (Ashburner *et al.*, 2000; Gene Ontology Consortium, 2001). GO is ‘the default source of functional annotations for virtually every experimental system and the gold standard for measuring the success of bioinformatic methods’ (Dolinski and Botstein, 2013). It is extensively used by researchers in a wide variety of situations, such as understanding

the function of genes discovered in a Genome Wide Association Study (Holmans *et al.*, 2009; Wang *et al.*, 2010) or computationally predicting functions for uncharacterized genes (Pena-Castillo *et al.*, 2008; Yan *et al.*, 2010).

An important feature of GO is that the ontology structure is constructed by a diverse team of scientists according to their best abilities to curate the published scientific literature. As the amount of cell biological literature increases, however, curating the ontology structure has become a painstaking effort that is proving difficult to scale up and systematize (Alterovitz *et al.*, 2010). Moreover, human curation necessarily favors biological entities that have been well studied and misses the large proportion of cell biology that is not yet known or has not yet been curated. For these reasons, it is not possible to directly learn about an uncharacterized gene or discover a new function using GO, and one cannot quickly assemble an ontology model for a new organism, let alone a specific cell type or disease state.

Recently, it has been shown by some of us that a GO can be inferred directly from molecular data as a complement to further curation efforts (www.nexontology.org) (Dutkowski *et al.*, 2013). For ontology curators, this approach ‘is extremely valuable in three ways. First...it finds connections missed by curators. Second, it will save huge amounts of curation time by pointing curators to the data that matter. Third, it provides a quality-control check on the GO that is unbiased by the vagaries of publication policies, as it is based only on the data themselves’ (Dolinski and Botstein, 2013). Furthermore, the ability to rapidly generate ontologies from data opens up new possibilities for the use of ontologies in general. ‘For example, data-driven ontologies generated from diseased and normal samples could be compared. This would be a novel way to look at what goes awry in particular disease states, providing the context and perspective of complex, interrelated biological processes’ (Dolinski and Botstein, 2013). Such an ontology model may also serve as the basis for an intelligent, predictive agent, as one of us has described elsewhere (Carvunis and Ideker, 2014). Despite these possibilities inspired by an initial attempt (Dutkowski *et al.*, 2013) it remains an open question as to how best algorithmically to infer an ontology from molecular data.

To understand the challenges involved in inferring an ontology from data, we first must recognize that ontologies contain both syntactic information (terms and their structural relations) as well as semantic information (relations between terms have defined meanings—in GO these include ‘is a’, ‘part of’ and ‘regulates’ relations). Both our previous work and this work will focus on inferring the syntactic information—the ontology terms, their relations and the annotations of genes to terms. This syntactic information is the most commonly used information by biologists using GO as a gold standard.

*To whom correspondence should be addressed.

GO is structured as a rooted, directed acyclic graph (DAG), where gene annotations propagate up the hierarchy from child terms to parent terms through ‘is a’ and ‘part of’ relations. For example, in the biological process (BP) branch, ‘M phase of mitotic cell cycle’ and ‘interphase of mitotic cell cycle’ are child terms beneath ‘mitotic cell cycle’. Since GO is a DAG, any term can have multiple parents and/or multiple children. Allowing for multiple children per term is necessary to capture the biological truth that many molecular machines have more than two subunits, or that a BP such as the interphase of cell cycle can be split into more than two phases (G1, S, G2). Multiple parents are necessary to capture the biological principle of pleiotropy—the reuse of genes and/or subunits, or the classification of a BP into multiple higher categories of process (e.g. ‘Sulfur Amino Acid Metabolic Process’ is a child of both ‘Cellular Amino Acid Metabolic Process’ and ‘Sulfur Compound Metabolic Process’). In order to faithfully infer an ontology of genes like GO, we need an algorithm able to infer a DAG.

In previous work, we developed a multistep algorithm for constructing a Network Extracted Ontology (NeXO), which infers a DAG structure from data summarized in the form of a molecular interaction or gene similarity network (Dutkowski *et al.*, 2013). This algorithm first infers a binary tree using the HAC-ML algorithm (Park and Bader, 2011). Then, splits without network support are collapsed, creating nodes with more than 2 children. Finally, a post-processing step searches for additional parent-child relationships between existing nodes, allowing for a single node to have multiple parents.

The NeXO algorithm does not use quantitative information about gene interaction or similarity, i.e. it assumes an input network for which the edges are unweighted. However, typical genome-scale data such as gene expression correlation populate a full matrix of similarity scores between pairs of genes where all of the information is contained in the weights. While thresholding these data is one way to interpret them as unweighted networks, this results in information loss not only below but also above the threshold, as hierarchical structure embedded in the weights is collapsed. Furthermore, when combining multiple types of data (as will be undoubtedly required for construction of a complete ontology of gene function), it is useful to give greater weight to gene pairs which have evidence of similarity in multiple datasets (Kim *et al.*, 2014). Since the NeXO algorithm is unable to use valuable information contained in the weights of edges in an input network, it is instead forced to rely on setting a proper threshold below which all information is ignored and above which relative weights are lost.

The essence of the DAG inference problem is to detect communities of genes that span a wide range of sizes/scales and can nest hierarchically as well as share arbitrary subsets of members. Furthermore, the algorithm should be able to construct a DAG using the information contained in weighted networks, and it should be capable of operating on a nearly complete, weighted graph with a genome-scale number of nodes (thousands) and edges (millions). A number of existing clustering algorithms address parts of this challenge. For example, there are many commonly used hierarchical clustering algorithms which use similarity or distance scores as input and infer a nested hierarchy of clusters (Flores *et al.*, 1951; Sneath and Sokal, 1973; Sokal

and Michener, 1958; Sørensen, 1948; Ward, 1963). These methods, however, rely on iterative joining of pairs of terms, resulting in forced construction of a binary tree. Clusters cannot overlap (i.e. have multiple parents for a single node) or have >2 children, and the number of clusters inferred is fixed at $n-1$ where n is the number of terminal nodes.

There have recently been a handful of algorithms which construct hierarchies with overlapping clusters, by creating a first level of overlapping clusters with terminal nodes and then combining these base clusters into higher level clusters (Becker *et al.*, 2012; Kovacs *et al.*, 2010; Kumpula *et al.*, 2008; Sales-Pardo *et al.*, 2007). These algorithms all present solutions that allow multiple parents and multiple children at the initial level of clusters, and some operate on weighted networks. However, these methods restrict each higher level cluster to having a single parent. This aspect inherently limits the types of relations that can be discovered by these methods.

Another relevant method has been proposed that hierarchically clusters links in a graph rather than edges (Ahn *et al.*, 2010). This method is quite flexible and allows for each node in the graph to appear in multiple clusters via participation in multiple edges. Such a method still restricts the types of clusters it creates, however. First, an edge can only participate in a single cluster at each level of the hierarchy (although a pair of nodes may incidentally participate in multiple clusters via other edges). Secondly, the hierarchical clustering of the edges allows only for binary joins between edges and therefore each lower level cluster will be joined with exactly one other cluster at each step. This process creates $e-1$ clusters, where e is the number of edges in the input graph. While the original paper proposes a method for determining a single optimal cut in the hierarchy for determining clusters, how to determine all levels of the hierarchy that are meaningful and not an artifact of construction remains an open question.

The LocalFitness algorithm, which has recently been proposed in the physics community, is to our knowledge the only previous approach that constructs a DAG from an unweighted or weighted network and has a principled way of determining which clusters are robust (Lancichinetti *et al.*, 2009). This method constructs clusters at a given level in the hierarchy by optimizing a fitness function for each of many potentially overlapping clusters built out from multiple seed nodes. The fitness function includes a parameter that is tuned to find clusters at multiple levels of the hierarchy. Graph partitionings that are stable across comparatively wide ranges of this parameter are used.

Here we present a basic formulation of the ontology inference problem followed by a description of a new method, called Clique Extracted Ontology (CliXO), for ontology inference based on progressive identification of maximal cliques. We evaluate CliXO in comparison with LocalFitness as well as several other methods such as the NeXO algorithm and standard clustering. Methods are evaluated based on their ability to reconstruct the GO from two starting datasets: (a) pairwise semantic similarities, which are derived directly from GO itself, and (b) three different -omics datasets (genetic interaction profile correlation, gene expression correlation and an integrated -omics dataset for yeast).

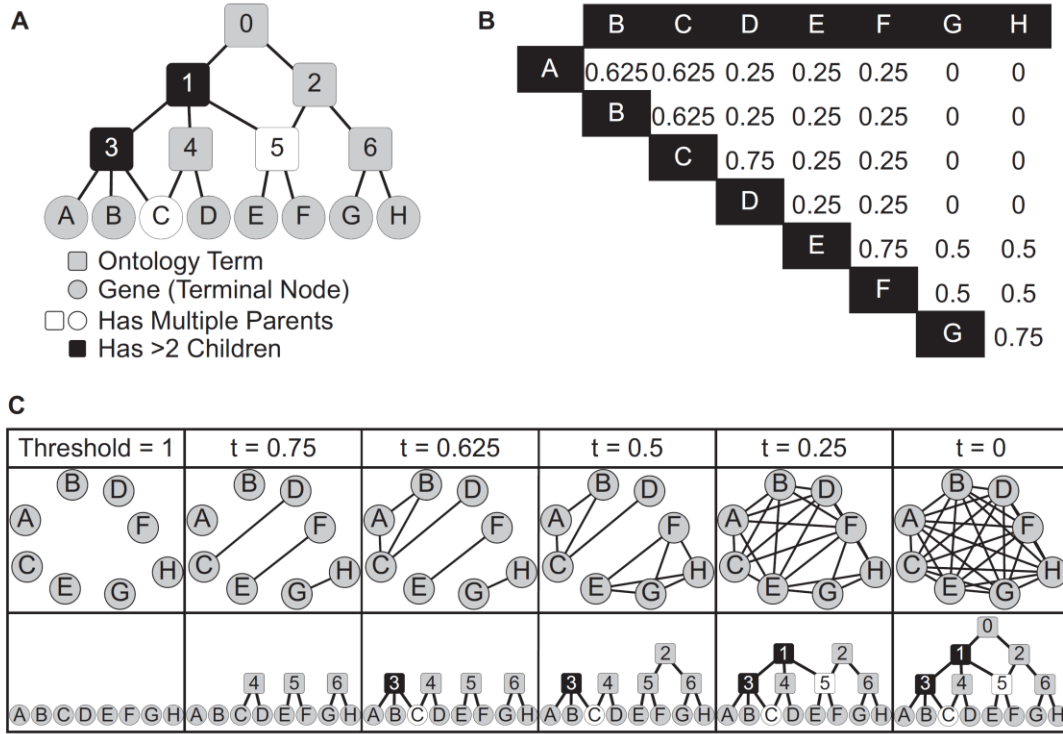


Fig. 1. CliXO method. (A) An example ontology with genes A-H and terms 0-6. (B) Semantic similarity scores calculated from the ontology in (A). (C) Example showing reconstruction of the ontology in (A) from the similarity scores in (B). As the threshold is decreased, edges that equal or exceed the threshold are added to the graph. At each new threshold, maximal cliques in the graph, corresponding to terms are found and added to the inferred ontology

2 METHODS

2.1 Basics

Define an *ontology-graph* as a weighted DAG $G = [T, N, E, w, r]$ with the following properties:

- The nodes in G are terminal (set T ; no outgoing edges), or non-terminal (set N). Non-terminal nodes in the ontology are also called *terms*. G has a single root $r \in N$ from which all nodes can be reached.
- Ultrametric property: a non-terminal node u in G has constant distance, denoted $w(u)$, to all its terminal descendants, denoted by $L(u)$.
- Witness property: for non-terminal node u , and all terminal nodes $a \notin L(u)$, there exists $b \in L(u)$ such that $w(a, b) > 2w(u)$.
- $|G|$ is the total number of non-terminal nodes in G .
- For any pair of terminal nodes a, b , let $w(a, b)$ denote the shortest path between a, b in G .
- Define a *Least Common Ancestor* $\text{lca}(a, b)$ such that $w(a, b) = 2w(\text{lca}(a, b))$.

Figure 1A shows an example of an ontology graph, with non-terminal nodes 0, ..., 6 and terminal nodes A, ..., H. This model of an ontology allows for the grouping of elements (terminal nodes) into coherent clusters of elements (i.e. non-terminal nodes) that are closer to each other than to other nodes based on pairwise distances between elements.

2.1.1 The Ontology DAG reconstruction problem (perfect case)

INPUT: A set of terminal nodes (i.e. genes) T , and a distance matrix M between all pairs in M (as shown in Figure 1B).

OUTPUT: An ontology-graph G with T as the set of terminal vertices and for all $a, b \in T$, $w(a, b) = M(a, b)$.

The input distances M rarely satisfy the ontology distances perfectly and therefore, in the *imperfect case*, we must compute the ontology DAG that best represents M .

2.1.2 The Ontology DAG reconstruction problem (imperfect case)

INPUT: A set of terminal nodes (i.e. genes) T , a distance matrix M between all pairs and a user-provided noise parameter α .

OUTPUT: An ontology-graph G with T as the set of terminal vertices, and which maximizes $|G|$ while satisfying the following:

- (1) For all $a, b \in T$, $w(a, b) \geq M(a, b)$.
- (2) For non-terminal node u , and all terminal nodes $a \notin L(u)$, there exists $b \in L(u)$ such that $M(a, b) > 2w(u) + \alpha$.
- (3) For each non-terminal node u , there must be at least one pair of terminal nodes $a, b \in L(u)$ for which $2w(u) + \alpha < 2w(v)$ for all v with $a, b \in L(v)$, $v \neq u$.
- (4) For all $u \in N$, $w(u) = 2 \max_{a, b \in L(u)} M(a, b)$

2.1.3 Distance versus similarity scores We note that in many cases, the relationship between terms is presented in the form of a similarity matrix, rather than a distance matrix. The two are interchangeable for the algorithms proposed below. For example, subtract the similarity scores from a large constant to get a valid distance function between two terms.

2.2 CliXO algorithm

We consider a simple heuristic as described in Figure 1C. Consider an undirected graph U with nodes $\in T$ and no edges. Let S be a stack of all pairs (a, b) sorted by distances $M(a, b)$, with the smallest value at the top.

Algorithm CLIXO(S)

Input: Stack S of sorted distances
Output: Non-terminal nodes in the ontology-graph
1. $C_G = \{\}$
2. **while** ($S \neq \{\}$)
3. $(a, b) = \text{top}(S)$; $t = M(a, b)$
4. **while** ($M(a, b) == t$)
5. $(a, b) = \text{Pop}(S)$
6. Add edge (a, b) to U
7. $C_{\text{cur}} = \text{Set of maximal cliques in } U$
8. $C_G = C_G \cup C_{\text{cur}}$
9. (C_G forms the set of non-terminal nodes in G)

Note that each maximal clique in U corresponds to a node of the ontology graph, with an obvious hierarchy. We use an algorithm proposed by Chiba and Nishizeki (1985) to compute all maximal cliques in time $O(m\mu a(U))$ where m is the number of edges in U , μ is the number of maximal cliques in U and $a(U)$ is the arboricity of U ($a(U) < n$ where $n = |T|$). While output-efficient, the number of cliques can be large. Moreover, the algorithm may output the same cliques repeatedly.

We modify the algorithm above to maintain only *informative* maximal cliques defined as follows. At any time in the algorithm, a pair of terminal nodes a, b is *explained* if a and b are both elements of a single clique which is already in C_G . A clique is *necessary* if it contains a pair of terminal nodes c, d which is not contained in any other single clique in C_{cur} . A clique $C \in C_{\text{cur}}$ is *informative* if it is both necessary and contains at least one pair of non-explained terms. Only informative cliques are added to C_G .

Next, we check cliques dynamically for maximality and informativeness. Define $N(a)$ as the set of vertices adjacent to a in U . Let C_a represent the set of maximal cliques in C_{cur} containing a . Consider Steps 5 and 6, where (a, b) is popped off S and edge (a, b) is added to U . For each clique $C' \in C_a$, we create a new clique $C' \cap N(b) \cup \{b\}$. Similarly, for each clique $C' \in C_b$, we will create a new clique $C' \cap N(a) \cup \{a\}$. These are the only new cliques. Each new clique C_{new} is checked for maximality. If maximal, add each C_{new} to C_{cur} , and remove from C_{cur} all cliques in C_a or in C_b that are contained in (i.e. a subset of) a C_{new} . Finally, when we are ready to increase the threshold [all (a, b) matching the current threshold have been popped], we check all cliques in C_{cur} and retain only the informative ones. CliXO with dynamic checking for maximality and informativeness can compute all informative maximal cliques in $O(Mn\gamma)$ where M is the number of non-zero edges in M , $n = |T|$ and γ is the number of informative cliques in the final output G . Practically, $\gamma \ll \mu$, where μ is the total number of all maximal cliques at each unique t in U , and this results in significant performance increase.

2.2.1 Imperfect case In the *imperfect case*, the only change is in Step 8 of the CliXO algorithm. Instead of adding all informative cliques in C_{cur} to C_G each time a new threshold is reached, we add to C_G only those informative cliques C in C_{cur} for which $\max_{a,b \in C} M(a, b) < t - \alpha$.

This algorithm can perfectly distinguish between signal and noise when α is smaller than s , the smallest distance between any child-parent node pair in the ontology and larger than a noise value $n(G)$, determined by the following procedure. For each term $u \in G$, order by value all $M(a, b)$ where $a, b \in u$, $n(u)$ for a non-terminal node u is the maximum difference between adjacent values in this ordered list and $n(G) = \max_{u \in N} n(u)$. It is not possible to directly determine $n(G)$ without knowing the true ontology structure, but it is a useful concept for understanding α . On a conceptual and practical level, $n(G)$ can be estimated as roughly $2 \times$ the standard error of measured pairwise distances (e.g. when the experiments used to measure distance are replicated both technically and biologically). α can be set to this estimated $n(G)$.

It may not always be possible to set α so that $n(G) < \alpha < s$ because sometimes $n(G) > s$. In this case, if $\alpha < n(G)$, it will result in the creation of extraneous terms that are a subset of a 'real' term. If $\alpha > s$ then any 'real' child node c that is close to its parent node p , i.e. $2(w(p) - w(c)) < \alpha$, will not be included in the ontology. One practical strategy for dealing with this case is for the user to focus on a particular small, well-known section of the ontology (e.g. a molecular machine like the proteasome) and observe it over various levels of α . If many extraneous terms seem to be being created, then α is likely too low. If known levels of the hierarchy are being collapsed, then α is likely too high.

2.2.2 Missing edges (false negatives) We expect that in real molecular data, there will be a number of missing edges—i.e. measurements for $a, b \in T$ where $M(a, b) \gg w(a, b)$ in the 'true' ontology. These missing edges will cause splitting of maximal cliques into highly overlapping smaller cliques. For example, a 'true' term of size k with just one missing edge in the measured distances M will result in two smaller terms of size $k + 1$ in the DAG inferred from M .

2.2.3 Algorithm: Missing Edges We maintain the properties of the noisy measurements case. However, we now add a user-defined parameter β and the ability to edit M , the input distance matrix, when we infer that an edge is likely missing (i.e. is a false negative). β is a parameter where $0 < \beta \leq 1$. Two cliques C_i and C_j are considered highly overlapping by the algorithm if for all $a \in C_i \cup C_j$,

$$\frac{|N(a) \cap (C_i \cup C_j)|}{|C_i \cup C_j| - 1} \geq \beta.$$

We modify the CliXO algorithm, again in Step 8. Before an informative, maximal clique $C_i \in C_{\text{cur}}$ is added to C_G , we check all other $C_j \in C_{\text{cur}}$ for high overlap with C_i . For any C_j found to be highly overlapping with C_i , we also check for any $C_k \in C_{\text{cur}}$ that are highly overlapping with C_j . Recall $w(C) = \max_{a,b \in C} M(a, b)$. For all such highly overlapping pairs of cliques C_i and C_j , we then set $M(a, b) = \max(w(C_i), w(C_j))$ for all $a, b \in (C_i \cup C_j)$ with $M(a, b) > \max(w(C_i), w(C_j))$. We add all edges adjusted in M to U and update C_{cur} .

2.2.4 Other algorithms considered We evaluate ontologies inferred using the NeXO algorithm (Dutkowski *et al.*, 2013), the Local Fitness Algorithm (Lancichinetti *et al.*, 2009), and several standard hierarchical clustering algorithms including the Unweighted Pairwise Group Method (UPGMA; Sokal and Michener, 1958), the Weighted Pairwise Group Method (WPGMA; Sneath and Sokal, 1973), Ward's minimum variance method (Ward, 1963), Complete Linkage (Sorensen 1948), and Single Linkage (Flores *et al.*, 1951).

2.3 Evaluation of reconstructed ontologies

2.3.1 Ontology alignment Ontologies were aligned as in Dutkowski *et al.* (2013). Briefly, 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. Term mapping in our alignment procedure is evaluated using a score

function that considers the similarity of the sets of genes assigned to terms (intrinsic similarity) and the similarity of the ontology hierarchy surrounding each term (relational similarity). Each pair of terms which is aligned receives an alignment score ranging from 0 to 1 where 1 represents identity in both intrinsic and relational similarity.

To calculate false discovery rate (FDR) of term alignment, we first create n ontologies with the same structure as the inferred ontology O_i but with all gene labels randomly permuted. The FDR of a term alignment at a given score t is then:

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

Where $N_{R_i}(t)$ is the number of terms in the random permutation i that have an alignment score $\geq t$, and $N(t)$ is the number of terms in the ontology O_i that have an alignment score $\geq t$. Terms are binned by size for the FDR computation, and we set a minimum score threshold value $t \geq 0.1$ for large terms and higher threshold values for small terms so as to maintain an FDR $< 5\%$ within each size group.

The alignment procedure also respects the preservation of structural relationships across ontologies being aligned. It does this by ensuring that there are no *conflicts* between individual mappings in the final alignment. There are two versions of the alignment—in the *permissive* version, a conflict exists between mappings (e_1, e_2) and (e_1', e_2') where $e_1, e_1' \in O_1$ and $e_2, e_2' \in O_2$ if there is a parent-child criss-cross—that is, 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 . In the *strict* version, there is a conflict if e_1 is a descendant of e_1' in O_1 and e_2 is not a descendant of e_2' in O_2 . The *permissive* version places most value on the alignment between individual nodes, whereas the *strict* version also requires that the relations between nodes are preserved.

2.3.2 Precision and recall of inferred ontologies For evaluating inferred ontologies, we first align the inferred ontology I to the gold standard ontology G , in our case GO. This gives us an alignment A . We then determine precision and recall as

$$\text{precision} = \frac{|A|}{|I|}, \text{ recall} = \frac{|A|}{|G|}$$

where $|A|$ is the number of terms aligned with $t \geq 0.1$ and FDR $< 5\%$.

3 RESULTS AND DISCUSSION

3.1 Ontology inference from semantic similarity

We began by testing performance using input similarity data that are perfectly consistent with the target ontology. For this task, we calculated the Resnik semantic similarity (Resnik, 1995) of all gene pairs in the BP branch of GO to form a complete, weighted graph. Using these data, CliXO was able to nearly identically infer the original ontology, with $>98\%$ of BP terms identically reconstructed (Fig. 2A) and 100% of inferred terms aligned to a term in BP (Fig. 2B). The other algorithms tested, including several common hierarchical clustering methods which build trees and a method for directly creating a DAG (Local Fitness), struggled with this task by comparison. The Local Fitness algorithm, while capable of operating on a complete, weighted graph, struggles to distinguish the borders of terms, instead joining true terms into larger groups which do not recapitulate the true ontology (recall $< 5\%$, precision $< 30\%$ by alignment). The hierarchical clustering methods behave similarly to each other (with the exception of single linkage which performs

worse). While these methods often reconstruct BP terms (recall of $\sim 20\%$ terms identical, $\sim 70\%$ terms aligned), these methods suffer from the creation of numerous extra terms. Since they create a binary tree, they are forced by construction to create terms even if they are not supported by the data. This greatly damages the precision of these methods. The ontologies created by these methods contain roughly $3\times$ as many terms as the true ontology. CliXO, on the other hand, finds the borders of each term from the data—it is not forced to create terms where they do not exist.

The NeXO algorithm, while not able to operate on a complete graph, can be run on these data by applying a threshold at a semantic similarity value and keeping only edges above this threshold. While this results in information loss, especially about larger ontology terms and substructure of smaller terms, NeXO is capable of at least approximating many terms ($\sim 40\%$ recall, $\sim 70\%$ precision by alignment) when using a network with edges consisting of the top 10000 semantic similarity scores (these top scores actually perfectly specify the terms of size 10 genes or fewer, as the vast majority of edges in the semantic similarity network are due to larger terms). Still, the terms produced by the NeXO algorithm are only an approximation of the true terms (precision, recall $< 10\%$ by identity). Furthermore, the precision slips dramatically with almost no gain in recall as the threshold is lowered and NeXO uses more edges (recall 42%, precision 62% by alignment using the top 20000 edges; recall 43%, precision 49% by alignment using the top 100000 edges).

The 83 GO BP terms not identically reconstructed by CliXO are those where information loss occurs in the conversion to semantic similarities. There are two such possible cases. First, two highly overlapping terms are indistinguishable from one larger term if the genes in the overlap already belong to a term with higher similarity. For example, if $\{a, b, c, d\} = T_1$, $\{a, b, c, e\} = T_2$, $\{d, e\} \in T_3$, then CliXO cannot distinguish T_1 from T_2 . Therefore, one term with $\{a, b, c, d, e\}$ will be created. Second, a term is ‘invisible’ in the semantic similarity data if all gene pairs in that term also belong to higher similarity terms. CliXO cannot discover these ‘invisible’ terms.

3.2 Accounting for noise

Of course, we cannot expect to have experimental data that exactly model a semantic similarity measure. We expect three types of noise in experimental data—false positives, false negatives and Gaussian noise.

The simplest of these cases is false positives (cases where a gene pair is falsely assigned a high similarity). Assuming that false positives are randomly scattered throughout the network, we can expect that these false positives are unlikely to form and/or complete a clique. Rather, these will lead to the creation of small maximal cliques, usually containing only the two genes with the false positive measurement between them. In some cases, a particularly well-placed false positive may create an extraneous clique of size 3. For this reason, we choose to ignore terms of size 2 and 3 when our algorithm is run on experimental data.

Gaussian noise in measurement will cause gene pairs that all belong to the same ontology term to have slightly different measured similarities. To simulate this, we added Gaussian noise centered at 0 with varying SDs to the calculated Resnik semantic

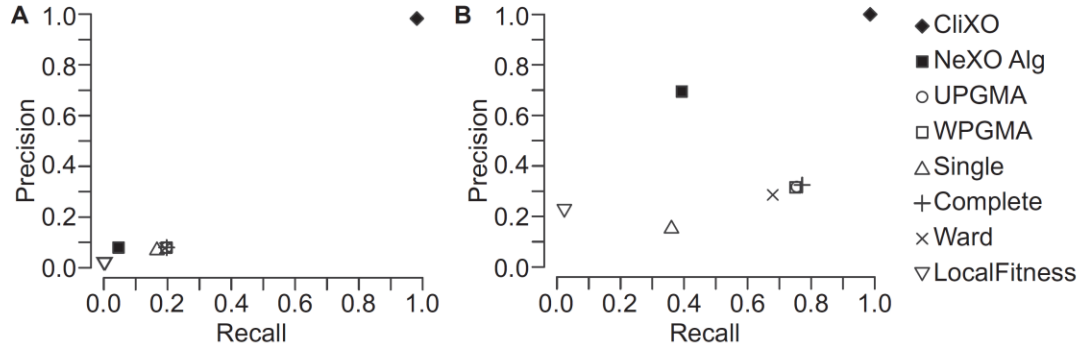


Fig. 2. Inferring an ontology from semantic similarity. Precision–recall plots for ontologies inferred from GO BP Resnik semantic similarities. True positive inferred terms are identical (A) or aligned (B) to a GO BP term using permissive alignment

similarities and then inferred the ontology. Without the α parameter ($\alpha = 0$), CliXO incorrectly infers many intermediate terms (leading to low precision, Fig. 3A) in addition to the correct terms (leading to high recall, Fig. 3B). When the α parameter is included and optimized (Section 2.2.1), then the CliXO algorithm is able to distinguish signal from noise. This allows for stable performance of CliXO even in the presence of noise. It should be noted that the median signal in this ontology is actually two orders of magnitude above the minimum signal, so even at a relative noise of 0.1, some terms in the ontology are already completely washed out due to noise. For all values where noise is less than minimum signal, CliXO with the α parameter is able to perfectly reconstruct the input ontology.

False negatives are cases where two genes belong to the same term but the edge between them either does not exist or has an extremely small similarity. The effect of a missing edge from a clique of size k genes is to split that clique into two smaller cliques of size $k-1$ genes that share $k-2$ genes. To account for these missing edges, our method incorporates the parameter β which causes the algorithm to find and merge highly overlapping cliques (Section 2.2.3). Without this parameter (equivalent to $\beta = 1$), CliXO immediately infers far too many terms, resulting in very low precision even with only 1–10% of edges missing (Fig. 3C). With $\beta = 0.5$, CliXO is able to maintain very high precision and recall even with large percentages of edges removed (precision and recall remain $>80\%$ with half of all edges removed and $>50\%$ with 80% of all edges removed, Fig. 3C, D). It should be noted that there is a slight hit in recall by using $\beta = 0.5$ when no edges are removed—this is because there are a small number of real terms in the ontology which qualify as highly overlapping, and these terms are erroneously merged. Still, recall remains $\sim 90\%$ with precision near 100%.

A thought experiment may reveal why the algorithm is robust with a fixed β over a wide range of missing edges. The operation to infer missing edges is based on finding pairs of highly overlapping cliques and filling in missing edges between them. If a single edge is missing from an otherwise complete clique C , this will create two highly overlapping cliques $C_1, C_2 \subset C$ which will each contain one unique element. C_1 and C_2 will be merged by the algorithm to create C . If a second edge is removed which does

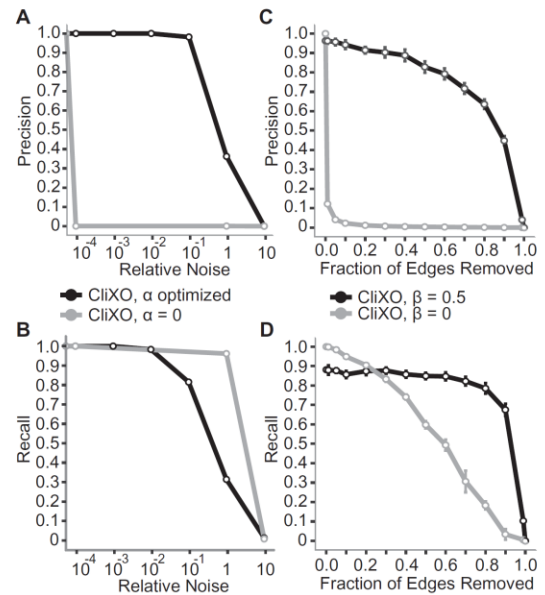


Fig. 3. CliXO with Noise. Precision (A, C) and recall (B, D) for ontology inferred from Resnik semantic similarities beneath and including CC Biogenesis BP term (GO:0044085). True positives by strict alignment. Varying levels of Gaussian noise added to (A, B) or edges removed from (C, D) semantic similarity measure. Relative noise = $1/\text{Median Signal to Noise ratio}$. Error bars represent standard error over 10 tested networks with random noise or edges removed per point

not share a vertex with the first, then these two cliques will now each split into two cliques (for a total of four cliques— $C_{11}, C_{12} \subset C_1$ and $C_{21}, C_{22} \subset C_2$). Even if β is set so that C_{11} is not considered highly overlapping with C_{21} or C_{22} , still C_{11} will first be joined with C_{12} to form C_1 while C_{21} and C_{22} are joined to form C_2 . C_1 and C_2 will then be found to be highly overlapping and

are therefore joined to make C . As more edges are removed from the original clique C this will simply lead to extra rounds of clique merging before C is recovered (this breaks down a bit as missing edges share vertices, but with a β set appropriately low this can be overcome). Because of this thought experiment and the experimental results in Figure 3C and D where the algorithm displays robustness to a wide range of missing edges with $\beta = 0.5$, we have set $\beta = 0.5$ for all experiments using -omics data. We recommend this setting for all use cases where missing edges are expected.

3.3 Ontology inference from -omics data for yeast

With CliXO performing well even in the presence of noise and significant levels of false negatives, we turned to the problem of inferring an ontology directly from experimental data. For this purpose, we used three -omics datasets—genetic interaction (GI) profile Pearson correlations as provided by Costanzo *et al.* (2010), gene expression (GE) profile correlations across all arrays in the Stanford Microarray Database (Hubble *et al.*, 2009) and YeastNet v3 (Kim *et al.*, 2014), a recently released functional gene network which combines evidence from multiple experimental sources to provide a weighted functional relationship between genes. All three of these datasets provide quantitative weights that correlate with the semantic similarities calculated from GO BP (Fig. 4A, C and E).

We noted that in all three of these datasets, there is a tight and strong correlation with the GO BP semantic similarity values at the very highest end. However, this tight correlation loosens as the edge weights in the -omics data decrease, as seen by wider ranges of scores and also several consecutive levels of scores indistinguishable from the lowest scores. This indicates increasing levels of false positives as the edge weights decrease. At this point we note that the CliXO algorithm outputs terms in order of decreasing weight, where a term weight is equal to the lowest similarity edge between any two genes in that term. This, combined with the decreasing reliability of the -omics data as the edge weights decrease, means that a single run of CliXO with a fixed α outputs terms in order of reliability, with the most reliable first. As such, a precision–recall curve can be calculated for a single CliXO ontology where a single point represents the alignment to GO of a CliXO ontology containing all terms above a given term weight. These precision–recall curves for CliXO at a fixed α are shown in Figure 4B, D and F, with the top left points representing a term weight equivalent to the top 10000 edges and the bottom right point representing all terms for GI and YeastNet v3 and a term weight equivalent to the top 100000 edges in the GE network.

We applied CliXO and several other algorithms to the problem of inferring the ontology from each of these three -omics datasets and aligned the resulting ontologies to GO BP. We report precision–recall curves for a range of α values, where lower α creates more terms and results in higher recall but at the cost of lower precision. For genetic interaction profile correlations, CliXO was the clear winner, with all α values resulting in precision–recall curves with higher precision–recall (50–10%, 3–8% by strict alignment) than competing methods including LocalFitness (15% precision, 5% recall), UPGMA (8%, 6.4%) and NeXO (10%, 5% at threshold of top 10000 edges).

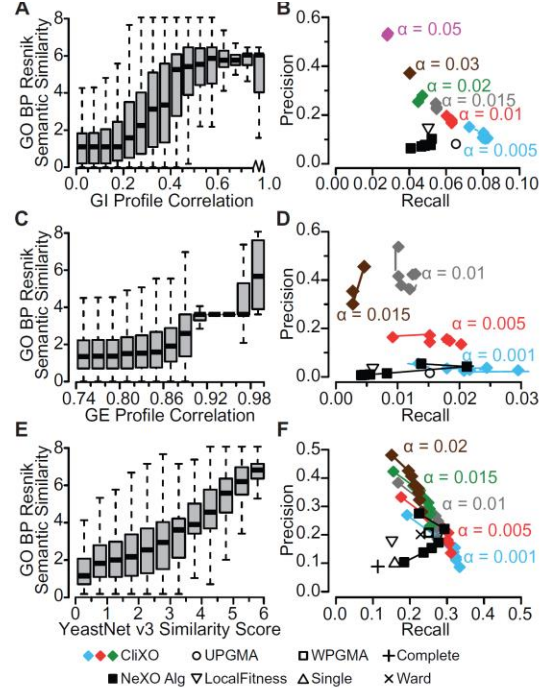


Fig. 4. Inferring an Ontology from -omics data. (A, C, E) Pairwise similarity scores from data versus BP Resnik Semantic Similarity. A = Genetic Interaction (GI) profile Pearson correlation as provided by Costanzo *et al.* (2010); C = Gene expression (GE) Pearson correlation from Stanford Microarray Database (SMD); E = YeastNet v3. (B, D, F) Precision–recall plot for ontologies reconstructed using various methods and evaluated by strict alignment to GO BP. Data from GI profile correlation (B), GE profile correlation (D) or YeastNet v3 similarity (F). CliXO with varying α is shown by color change. At a given α parameter, the precision–recall curve for the CliXO ontology (terms ordered by weight at which they are inferred) is shown. NeXO results are shown with varying threshold edge weights for generating an unweighted input network

Next, several algorithms were run on the Pearson correlations of gene expression profiles across all arrays in the Stanford Microarray Database. All methods scored relatively poorly for the recall of GO BP terms (<3%). This is likely because these profiles were calculated across all arrays, causing information from individual experiments (which each represent relatively few arrays of the total) to be lost. The CliXO algorithm, however, was the only algorithm capable of separating the information present from noise, producing precision as high as >50% (Fig. 4D) for some parameter settings, greatly outperforming other algorithms (<10% precision).

We next applied all algorithms to YeastNet v3. In this test, all versions of CliXO (precision 17–41%, recall 31–21% by strict alignment) and the NeXO algorithm with an optimized threshold of the top 20000 edges (precision ~24%, recall ~28% by strict alignment) outperformed other methods tested (Fig. 4F).

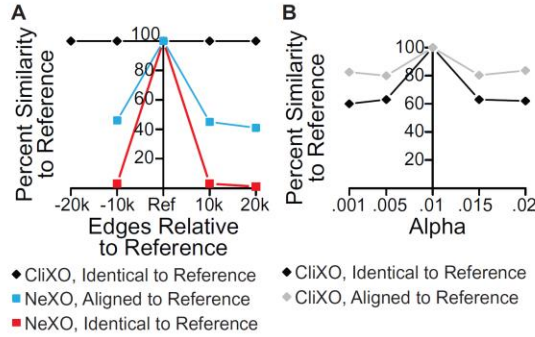


Fig. 5. Stability of NeXO and CliXO with respect to parameters. (A) Both NeXO and CliXO algorithms run with varying numbers of edges from YeastNet v3. Reference is best performing result (20k for NeXO, 30k, $\alpha = 0.01$ for CliXO). Percent similarity defined as number of terms either identical or aligned strictly divided by the number of terms in the smaller of the two ontologies produced. (B) CliXO run with varying levels of α . Reference is $\alpha = 0.01$

We note that while NeXO with an optimized threshold parameter is capable of matching CliXO in precision and recall, the performance of the NeXO algorithm is much more sensitive to this parameter choice than CliXO is to the choice of either α or a term weight threshold. In fact, the performance of the NeXO algorithm degrades quite rapidly as its threshold is loosened (Fig. 4F). Furthermore, the NeXO algorithm is quite inconsistent in its output as the threshold is varied. When compared with the best NeXO using the top 20,000 edges in YeastNet v3, no other versions of NeXO reproduced >4% of the terms exactly, and less than half of the terms aligned strictly (Fig. 5A). Without a reliable gold standard, it would be quite difficult to set this threshold for NeXO reliably.

CliXO, on the other hand, simply offers the user a tradeoff between precision and recall with its parameters. As α increases, so does precision at the cost of decreased recall. Furthermore, the CliXO algorithm is quite consistent as its parameters are varied, in contrast to the NeXO algorithm. As more terms are considered (lower edge weight threshold), all previously created terms are unaffected (Fig. 5A). Varying α also preserves the majority of terms, with >60% identical and >80% aligned strictly to CliXO with $\alpha = 0.01$ (Fig. 5B). This behavior allows the user to have confidence that CliXO is producing reliable output over a range of parameter settings. Furthermore, the effect of tuning these parameters is predictable.

The CliXO inferred ontology with $\alpha = 0.01$ infers 1833 terms from the top 30,000 edges in YeastNet v3 with a precision of 30% and a recall of 25% by strict alignment. The resulting structure is in fact a DAG and not a tree, as many nodes have multiple parents (Fig. 6A) and multiple children (Fig. 6B). Genes are more likely to be annotated directly to multiple parents (range 1–12, median 2) than internal nodes (range 1–3, 32 have multiple parents). Created terms range in size, roughly following a power law distribution as expected (Fig. 6C). The inferred ontology aligns not only to terms in the BP branch of GO, but also to the cellular component (CC) and Molecular Function (MF)

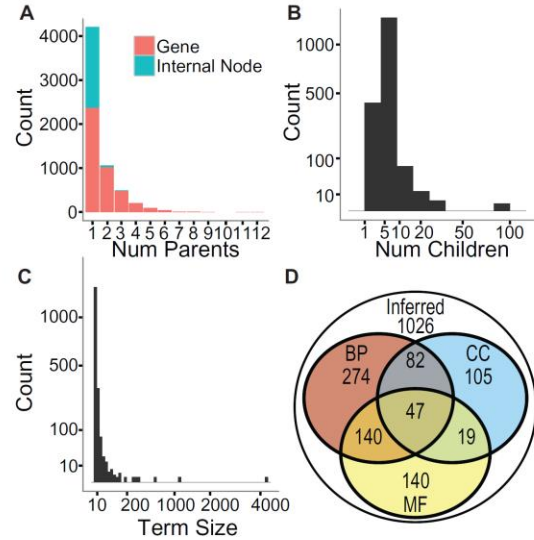


Fig. 6. CliXO Ontology from YeastNet v3. (A–C) Statistics about CliXO inferred ontology with $\alpha = 0.01$, including number of parents (A) and number of children per non-terminal node (B) and distribution of term sizes inferred (C). (D) Number of terms in CliXO inferred ontology with $\alpha = 0.01$ aligned strictly with FDR <5% against all three branches of GO (CC, MF, BP). Unaligned terms in outer circle

branches. Overall, 44% of the 1833 terms in this inferred ontology align to at least one of the branches in GO at an FDR of <5% using strict alignment (Fig. 6D).

It should be noted that YeastNet v3 learns functional relationships from data in a supervised fashion, with GO BP term comembership serving as the gold standard for training. Therefore, there is some circularity in reconstructing GO BP from this network. However, comparisons between reconstruction methods are still valid. Furthermore, the training takes place only at the level of the weight for each experiment type, so there is ample opportunity for a gene pair to score highly despite no known relationship in GO.

3.4 Further applications of CliXO algorithm

Here we have explored the *de novo* construction of ontologies from pairwise similarity data and compared several potential algorithms to do so. We have also introduced a new algorithm, CliXO, which performs well at this task and also may have further applications. First, CliXO may be useful as a way to combine manually curated ontologies with -omics data to create an updated ontology. For example, semantic similarities calculated from a manually curated ontology could be adjusted based on support in -omics data, followed by reconstruction of the ontology using CliXO. The updated ontology could then be aligned to and compared with the original ontology, revealing areas where -omics data supports changes to the original ontology.

Furthermore, CliXO can be viewed as a generic algorithm to convert pairwise similarity measurements between individual

entities into a hierarchical DAG where terms represent groups of entities which are similar to each other and a single entity may belong to more than one term. The exact meaning of the resulting structure will change depending on the similarity measurement used, but in general there will be an intuitive understanding of the result. For example, one could take a similarity measurement between cancer patients' genomes and then use the CliXO algorithm to construct a hierarchical DAG where each term represents a group of patients that are genetically similar to each other—i.e. genetic subtypes of cancer. As another example, one could take measurements of connection similarity in a social network and use CliXO to discover social groups in this network. We leave these applications for future work.

4 CONCLUSION

Here we have explored the requirements for inferring an ontology from a similarity matrix and available methods for doing so, as well as proposed a new method, CliXO. We found that CliXO outperforms other methods when a reliable similarity matrix is available. When using -omics data, we found that CliXO outperforms other methods on two of three datasets tested and ties the NeXO algorithm on the third, where both were able to successfully infer an ontology similar to GO BP. Furthermore, CliXO proved significantly more stable than NeXO to changes in parameters. This study provides the algorithmic foundation for building gene ontologies by capturing hierarchical and pleiotropic structure embedded in biomolecular data.

Funding: National Resource for Network Biology and the Cytoscape Project (National Institutes of Health grants P41GM103504 and R01GM070743); Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program (to M.K.). VB was supported in parts by grants from the National Science Foundation (CCF-1115206 and IIS-1318386), and from the NIH (U54 HL108460, 1P01HD070494).

Conflict of Interest: none declared.

REFERENCES

- Ahn,Y.Y. *et al.* (2010) Link communities reveal multiscale complexity in networks. *Nature*, **466**, 761–764.
- Alterovitz,G. *et al.* (2010) Ontology engineering. *Nat. Biotechnol.*, **28**, 128–130.
- Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
- Becker,E. *et al.* (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, **28**, 84–90.
- Carvunis,A. and Ideker,T. (2014) Siri of the cell: what biology could learn from the iPhone. *Cell*, **157**, 534–538.
- Chiba,N. and Nishizeki,T. (1985) Arboricity and subgraph listing algorithms. *SIAM J. Comput.*, **14**, 210–223.
- Costanzo,M. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
- Dolinski,K. and Botstein,D. (2013) Automating the construction of gene ontologies. *Nat. Biotechnol.*, **31**, 34–35.
- Dutkowski,J. *et al.* (2013) A gene ontology inferred from molecular networks. *Nat. Biotechnol.*, **31**, 38–45.
- Florek,K. *et al.* (1951) Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.*, **2**, 282–285.
- Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
- Holmans,P. *et al.* (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.*, **85**, 13–24.
- Hubble,J. *et al.* (2009) Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Res.*, **37**, D898–D901.
- Kim,H. *et al.* (2014) YeastNet v3: a public database of data-specific and integrated functional gene networks for *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **42**, D731–D736.
- Kovacs,I.A. *et al.* (2010) Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS One*, **5**, e12528+.
- Kumpula,J.M. *et al.* (2008) Sequential algorithm for fast clique percolation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **78**, 026109.
- Lancichinetti,A. *et al.* (2009) Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.*, **11**, 033015.
- Park,Y. and Bader,J.S. (2011) Resolving the structure of interactomes with hierarchical agglomerative clustering. *BMC Bioinformatics*, **12** (Suppl. 1), S44.
- Pena-Castillo,L. *et al.* (2008) A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol.*, **9** (Suppl. 1), S2.
- Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. *Int. Joint Conf. Artif.*, **1**, 448–453.
- Sales-Pardo,M. *et al.* (2007) Extracting the hierarchical organization of complex systems. *Proc. Natl Acad. Sci. USA*, **104**, 15224–15229.
- Sneath,P.H. and Sokal,R.R. (1973) *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. W. H. Freeman, San Francisco, USA.
- Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin*, **38**, 1409–1438.
- Sørensen,T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, **5**, 1–34.
- Wang,K. *et al.* (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 843–854.
- Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
- Yan,H. *et al.* (2010) A genome-wide gene function prediction resource for *Drosophila melanogaster*. *PLoS One*, **5**, e12139.

ACKNOWLEDGEMENTS

Chapter 3, in full, is a reprint of the material as it appears in: Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T. “Inferring gene ontologies from pairwise similarity data.” *Bioinformatics*. 2014;30(12):i34-42.. The dissertation author was the primary investigator and first author of this work.

CHAPTER 4

Active Interaction Mapping reveals the hierarchical organization of autophagy

Michael H. Kramer^{1*}, Jean-Claude Farré^{2*}, Koyel Mitra¹, Michael Ku Yu¹, Katherine Licon¹, Rama Balakrishnan³, J. Michael Cherry³, Suresh Subramani², Trey Ideker¹

¹Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA

²Section of Molecular Biology, Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA

³Department of Genetics, Stanford University, Stanford, CA 94304, USA

*Co-first authors

ABSTRACT

We have developed a general progressive procedure, Active Interaction Mapping, to guide assembly of the hierarchy of functions encoding any biological system. Using this process, we assemble a comprehensive ontology of functions comprising autophagy, a central recycling process implicated in numerous diseases. A first-generation model, built from gene networks previously generated in *Saccharomyces*, recovers most known autophagy components and places them in broad relation to the cellular response to stress, cell cycle and general vesicle transport. Systematic analysis identifies synthetic-lethal genetic interactions as the most informative data for further experiments; consequently, we saturate the model with ~156,000 such measurements across autophagy-activating conditions. These targeted interactions provide more information about autophagy than all previous datasets, leading to a second-generation ontology of 220 discrete, hierarchically related functions. Among multiple new findings, we confirm a role for Atg26 in transport of aminopeptidase I by the cytoplasm-to-vacuole targeting pathway, a biosynthetic autophagic process.

INTRODUCTION

A key promise of Systems Biology has been to advance our understanding of biological systems by combining genome-scale data with model building in a virtuous cycle (Ideker et al., 2001a; Kitano, 2002). New genomics data inform the biological model and, in turn, analysis of the model informs collection of new datasets. Systems Biology studies typically begin with extensive survey of the literature to assemble knowledge of the specific molecules, interactions and reactions involved in a biological system and to use this knowledge to initialize a computational model, for instance of metabolism, cell signaling or transcriptional regulation (Covert et al., 2004; Davidson et al., 2002; Ideker et al., 2001b; Karr et al., 2012; King et al., 2004; Malleshaiah et al., 2010). Once the model is built, model predictions are systematically compared to laboratory measurements, with discrepancies used to improve the model and design further experiments.

Formulating a model from literature can be laborious, requires that much is already known about the system, and may be biased towards well-studied components. Alternatively, many studies have sought to construct models of biological systems directly from systematic datasets (Ergun et al., 2007; Ho et al., 2002). Genome-scale data, including profiles of mRNAs, proteins or metabolites and gene and protein interactions, are analyzed to infer a set of genes, gene interactions, and/or clusters or modules of genes describing the system (Behrends et al., 2010; Havugimana et al., 2012; Kim et al., 2011; Lefebvre et al., 2012; Picotti et al., 2013). In recent years, such network models have been increasingly used to sketch the molecular relationships underlying key biological systems, including tissue specificity (Greene et al., 2015) and stem cell pluripotency and differentiation

(Cahan et al., 2014), as well as diseases such as autism (Parikshak et al., 2013) and cancer (Carro et al., 2010; Creixell et al., 2015; Leiserson et al., 2015).

At this juncture, we see two challenges. First, biological systems do not merely consist of a flat network of genes and gene modules – they are exquisitely hierarchical in nature. Models should seek to capture this hierarchy, which extends over multiple scales from nucleotides (1nm scale) to proteins (1-10nm), protein complexes (10-100nm), cellular processes (100nm), organelles (1 μ m), cells (1-10 μ m), tissues (100 μ m-100mm) and complex organisms (>1m). Second, while the field has taken a foundational step by enabling the construction of models directly from data rather than literature, a next step will be to determine how to iteratively improve these models with the most appropriate new data. The ability to rationally motivate and design new -omics data sets is sorely needed in genomics at large; far too many large-scale experiments are performed simply because one can, without formal justification or guidance for why those experiments might lead to an improved biological understanding.

Towards the first challenge, a hierarchical biological model in very widespread use is the Gene Ontology (GO) (Ashburner et al., 2000). Through extensive literature curation efforts, GO aims to factor the cell into a complete hierarchy of cellular components and processes, each represented by a GO “term”. Recently, we showed that an ontology of the cell very much like GO could be automatically constructed from diverse genome-scale datasets, bypassing manual curation of literature (Dutkowski et al., 2013; Kramer et al., 2014). Using this method we built a data-driven ontology of yeast, the Network eXtracted Ontology (NeXO), which recapitulated approximately 60% of known cellular components in GO. Analysis of new terms in NeXO but not in GO revealed many functions that were

subsequently added to GO by its curators, illustrating the potential for biological discovery through the construction of ontologies directly from data. However, it was also clear that our hierarchical understanding of cell biology is far from complete, leaving open the second challenge: how to most efficiently improve cell function hierarchies with high value experiments. Thus far, methods have been devised for selecting the optimal next experiment considering a particular kind of model (e.g. a gene network), data type (e.g. growth phenotyping), and a constrained repertoire of treatments (e.g. single or double gene knockouts) (Atias et al., 2014; Barrett and Palsson, 2006; Ideker et al., 2000; King et al., 2009; King et al., 2004). However, such questions of experimental design have not been approached for building ontologies, and given the plethora of -omics data types that can now be generated, guidance is especially needed on which type of data one should collect.

Here, we describe a general program for elucidating the hierarchy of functions underlying a cellular process, based on progressive cycles of mapping and modeling which we call Active Interaction Mapping (AI-MAP, **Figure 1**). We apply this approach to develop a hierarchical model of autophagy, the process of “self-eating” by which cells respond to starvation and other stress by degrading macromolecules into amino and fatty acids for synthesis of new proteins and ATP (Levine and Yuan, 2005). During autophagy, which is highly conserved from yeast to mammals (Feng et al., 2014), cellular contents are enclosed in a double-walled membrane and delivered to the yeast vacuole or mammalian lysosome for degradation; such contents can include general bulk cytoplasm (macroautophagy) or selective targets such as peroxisomes (pexophagy), mitochondria (mitophagy), ribosomes (ribophagy), protein aggregates (aggregophagy), lipid droplets (lipophagy), pathogens (xenophagy) or parts of the nucleus (nucleophagy). All general and

selective autophagy pathways share a core set of molecular machinery (Jin and Klionsky, 2013), which has been implicated in tumor suppression (Liang et al., 1999), neurodegeneration (Rubinsztein et al., 2005), pathogen defense (Talloczy et al., 2006), development (Wu et al., 2013), growth control (Wang and Levine, 2010), and aging (Rubinsztein et al., 2011).

In what follows, we 1) construct a first-generation ontology model of autophagy from publicly available interaction networks for the budding yeast *Saccharomyces cerevisiae*, 2) identify synthetic-lethal genetic interactions as the most useful type of data for systematically improving the model, 3) generate static and differential genetic networks involving growth measurements of approximately 156,000 yeast genotypes across autophagy-relevant conditions, and 4) integrate these new with previous datasets to arrive at a substantially improved understanding of autophagy. The Active Interaction Mapping process creates a ‘living’ ontology of cell function which actively suggests new data types for self-improvement and highlights new functions discovered as data are added.

RESULTS

Integrating diverse ‘omics data constructs a first-generation model of autophagy

To initialize the AI-MAP process, we sought to construct a ‘first-generation’ hierarchical model of autophagy, called the Autophagy Ontology (AtgO), based on publicly available molecular interaction network datasets in *S. cerevisiae*. Although not all datasets have been specifically designed to study autophagy, general or selective autophagy pathways are thought to have some activity in nearly all conditions (Reggiori et al., 2012). In total, we obtained a large compendium of 76 networks of 7 different data types, including protein interactions, genetic interactions, gene co-expression and gene-gene

similarity based on shared protein sequence and structural information, as previously curated for YeastNet v3 (Kim et al., 2014) (**Figure 1A**).

To build an ontology from these heterogeneous datasets, we applied a previously published approach (Kramer et al., 2014) which integrates all data into a single pairwise gene similarity network (**Figure 1B**) and then analyzes the hierarchical structure of this network to infer a gene ontology (**Figure 1C**). Pairwise gene-gene similarities in the data-derived similarity network were determined using a statistical regression procedure that combines available interaction datasets into a single quantitative similarity score. To learn how the different data types should be combined, the regression was trained to model a ‘standard reference’ gene similarity network derived from GO (gene-gene similarity scores based on relatedness in GO of the two genes, see **Experimental Procedures**). In this way, the data-derived similarity network was able to capture known biological processes and components in GO when supported by data (Pearson $r = 0.4$ versus GO, $p < 10^{-300}$, out-of-bag prediction); similar data combinations then identify new biological processes, components, and relationships not previously documented (**Figure 2A**).

For assembly of the ontology, we selected 492 candidate genes with potential relation to autophagy. First, we included 20 genes designated as having core functions common to general and selective autophagy (Jin and Klionsky, 2013). Beyond these core genes, another 102 genes were annotated to “autophagy” (GO:0006914) by GO and were also included. Finally, 370 additional genes were identified based on network similarity to the core genes (**Figure 2B**). A broad net was intentionally cast to allow later steps of the AI-MAP process to dictate which of these 492 genes are most related to autophagy.

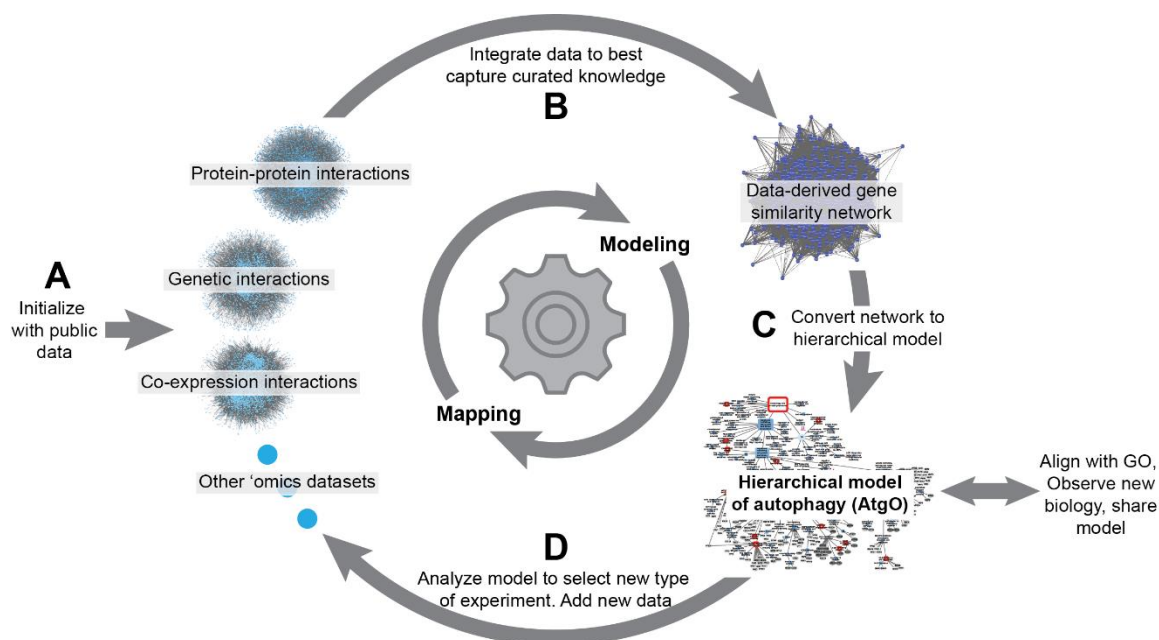


Figure 1. Program for Active Interaction Mapping. After initialization with public data sets (A), the approach proceeds by progressive iterations of data integration (B), hierarchical model assembly (C), and adaptive generation of new data (D).

The CliXO algorithm (Clique Extracted Ontologies, Kramer et al., 2014) was then applied to analyze the hierarchical structure of the data-derived gene similarity network among these 492 genes (**Figure 1C**). By this method, nested communities of genes apparent in the pairwise similarity data were identified and represented by a hierarchy of terms and term relations (inferring a directed acyclic graph, **Figure 2C**). The result was an initial Autophagy Ontology (AtgO 1.0) of 218 terms and 310 term relations. To determine which terms represent known biological functions and which represent potential new biology, AtgO 1.0 was aligned with GO according to a previously described alignment procedure (Dutkowski et al., 2013). This process seeks a one-to-one mapping between terms in AtgO and terms in GO; for significantly aligned terms, term names and descriptions are transferred from the GO reference. Turning to the section of the AtgO

hierarchy closest to known autophagy genes, we focus on term AtgO:9, which aligned to the GO term “autophagy”. This term encapsulated 234 genes organized into 82 sub-terms (**Figure 3A-C**), including 19 of the 20 core autophagy genes and 67 of the other 102 genes annotated to GO:“autophagy” (**Figure 3D**). Also included were 148 genes newly associated via the similarity networks; within these we observed a significant enrichment for functions in cell cycle (51 genes, $p < 10^{-12}$), cellular response to stress (42 genes, $p < 10^{-12}$) and general vesicle transport (47 genes, $p < 10^{-24}$), placing these functions in broad relation to autophagy. Of the 82 sub-terms, 6 were significantly aligned to GO terms in the GO:“autophagy” branch, 32 were significantly aligned to other GO terms and therefore brought these functions under the umbrella of autophagy, and the rest represented potential new sub-components or processes of autophagy discovered from data.

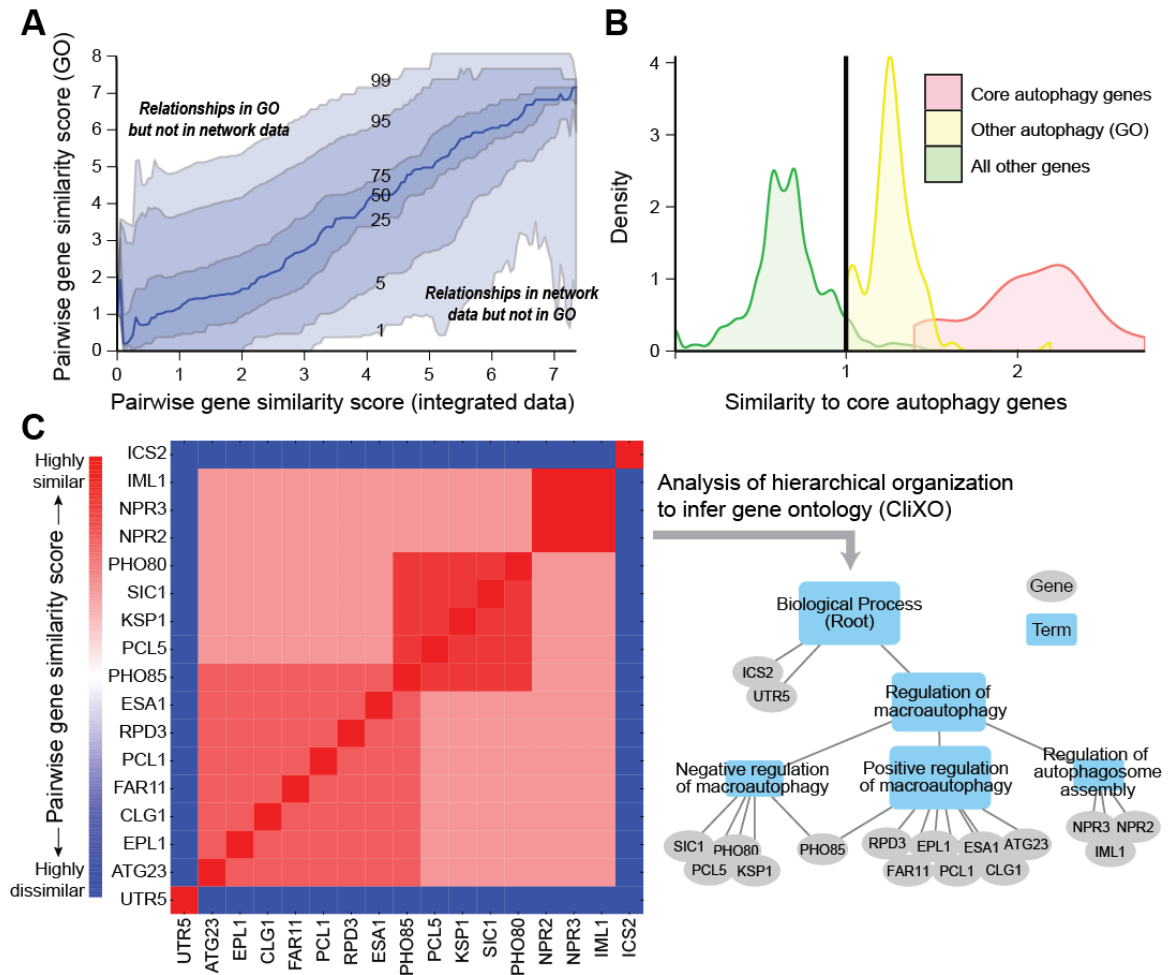


Figure 2. Assembly of hierarchical model from integrated data. (A) Pairwise gene similarity scores from curated knowledge (GO) versus the integrated data. Curves track percentiles in GO similarity scores at a given level of data similarity. **(B)** Selection of autophagy-relevant genes. The similarity of each yeast gene to a seed set of 20 core autophagy genes (Jin and Klionsky, 2013) is calculated by integrating information in the data- and GO-derived similarity networks. Kernel-smoothed probability densities of these similarities are plotted for the 20 core autophagy genes (red), 102 other genes annotated to GO:“autophagy” (yellow), and all other genes (green). A similarity threshold (vertical line) was set to select genes at least as similar to core genes as those in GO:“autophagy”. **(C)** Inference of the hierarchical model from (idealized) pairwise similarity scores using CliXO.

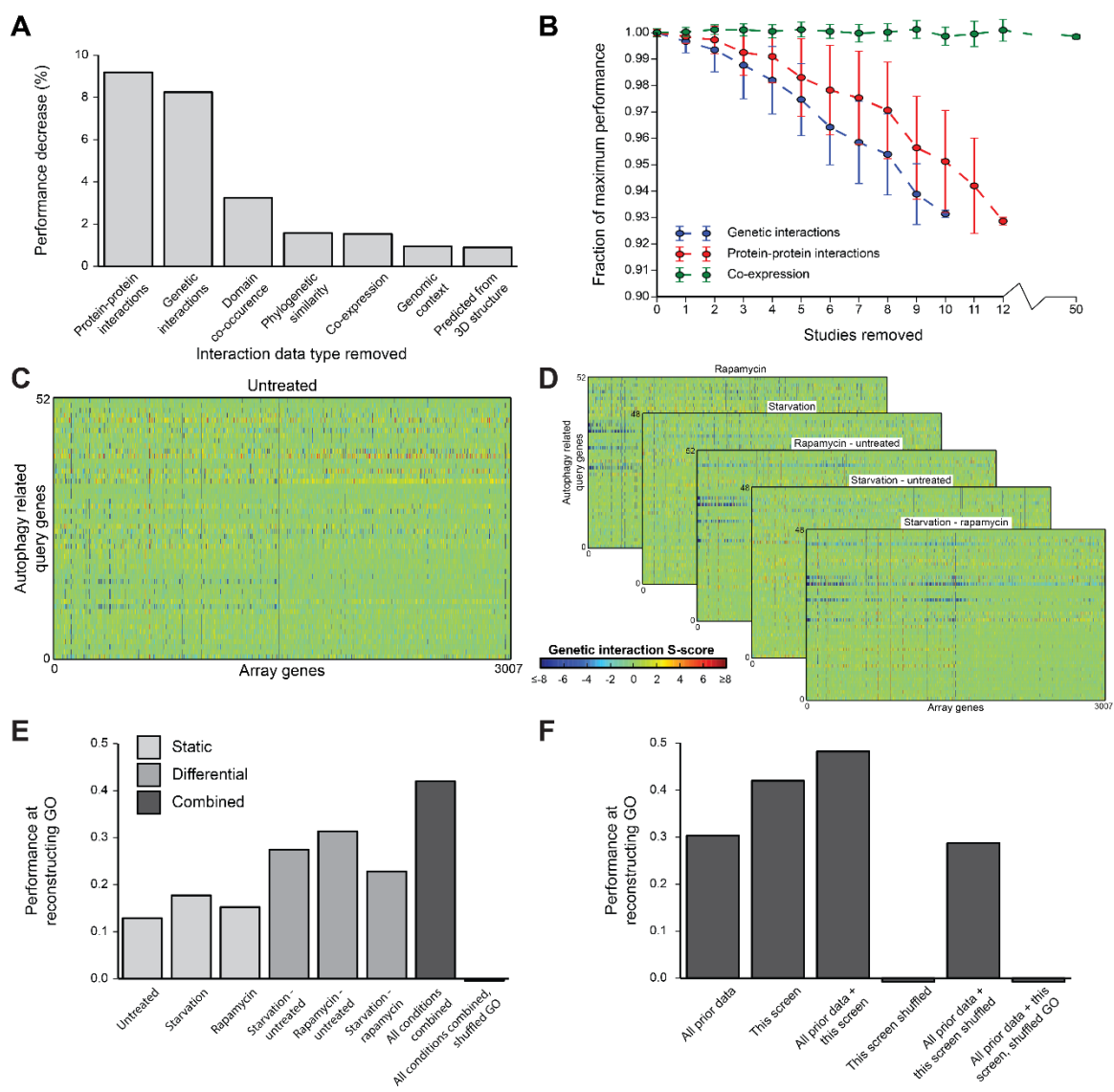
Figure 3. First-generation autophagy ontology. (A-C) AtgO 1.0 model of terms (rectangles), genes (ovals), and hierarchical interrelations among these (links). Rectangle size shows number of genes annotated to that term; its color shows term alignment to GO: darker blue indicates higher similarity to GO, white terms / red outlines do not align. Oval color indicates gene status. For compactness, gene annotations are displayed only for terms AtgO:15 (**B**, no GO alignment) and AtgO:18 (**C**, aligned to GO: “macroautophagy”). (**D**) Comparison to the Gene Ontology as curated from literature, showing GO terms and annotations for GO:0006914:“autophagy” and its descendants. Links of the same color are relations with the same parent term.

Analysis of model reveals genetic interactions as most effective new data

Next, we determined which new network data would be most capable of improving the AtgO model in further experiments (**Figure 1D**). For this purpose, we removed all data of a given type and evaluated the resulting decrease in model performance, measured as the ability to capture the GO hierarchy (**Figure 4A**). The largest contribution was from protein-protein interactions, closely followed by genetic interactions; that is, recovery of GO depends most strongly on data of these two types. Beyond this absolute importance measure, we were also interested in the importance of data relative to the present time, i.e., the extent to which adding another dataset of a given type is expected to improve the model. Therefore, we measured the decrease in model performance as individual studies of a particular type were sequentially removed (**Figure 4B**). The unit of one ‘study’ was used to normalize for different temporal and financial costs of generating each type of data – it roughly estimates the amount of data one ‘omics lab can produce in one study on average (i.e. data per lab-study). We found that removing single studies of genetic interactions caused the steepest instantaneous decrease in performance, suggesting that such studies may continue to provide the most information. A similar result was obtained when individual interactions rather than whole studies were progressively removed (**Figure S1**). Interestingly, gene expression data provided very little information not captured by other data types, either when removed as a group (**Figure 4A**) or one study at a time (**Figure 4B**).

Figure 4. Active analysis of data types leads to conditional genetic interaction maps.

(A) Performance decrease in AtgO 1.0 when all data of one type are excluded from the model. Performance is measured as the Pearson correlation between pairwise similarity scores derived from integrated data versus GO, focusing on gene pairs within the 492 autophagy-related genes (see main text). **(B)** Performance degradation as single studies of the indicated type (color) are cumulatively excluded. The mean (points) and standard deviation (error bars) are calculated over 50 random sets of removed studies. See also **Figure S1**. **(C-D)** New genetic interaction maps between 52 autophagy query genes and 3007 non-essential array genes in untreated conditions (rich media, **C**) as well as rapamycin and starvation conditions (**D**). Differential maps between each pair of conditions are also displayed. **(E-F)** Performance of reconstructing GO, using new data only (**E**) or integrating new with prior data (**F**). Performance is measured over gene pairs within the set of 52 autophagy query genes, in comparison to randomized data or GO (mean of 100 trials).



Conditional genetic interaction maps markedly improve the ontology

Guided by the above analysis, we designed a systematic screen for genetic interactions targeted at genes and conditions relevant to autophagy (**Figure 4C,D**). Synthetic Genetic Array (SGA) technology in *S. cerevisiae* (Tong and Boone, 2006) was used to query 52 autophagy-related genes for genetic interactions against a panel of 3,007 genes, covering approximately two-thirds of the non-essential yeast genome. SGA uses high-throughput robotic colony pinning on agar to create and score growth of many double gene deletion strains in parallel, here yielding $52 \times 3,007 = 156,364$ tests for gene-gene interaction. SGA was performed in three conditions: rapamycin, which pharmacologically induces autophagy; nitrogen starvation, which metabolically induces autophagy; and an untreated control. The resulting growth measurements were analyzed using an established computational workflow (Bean et al., 2014) to assign quantitative S-scores to all gene pairs: S-scores near zero indicate that the growth of the double deletion mutant is as expected (no interaction); positive S-scores indicate faster than expected growth (epistatic or suppressive interaction); and negative S-scores indicate slower than expected growth (synthetic-sick or lethal interaction). We computed the correlation in genetic interaction profiles between every pair of query genes in each condition, yielding three new networks (the genetic interaction profile of a query gene is its complete set of S-scores over all array genes). Correlations were also computed in three ‘differential’ configurations, based on the differences in S-scores between 1) rapamycin and untreated, 2) nitrogen starvation and untreated, and 3) nitrogen starvation and rapamycin.

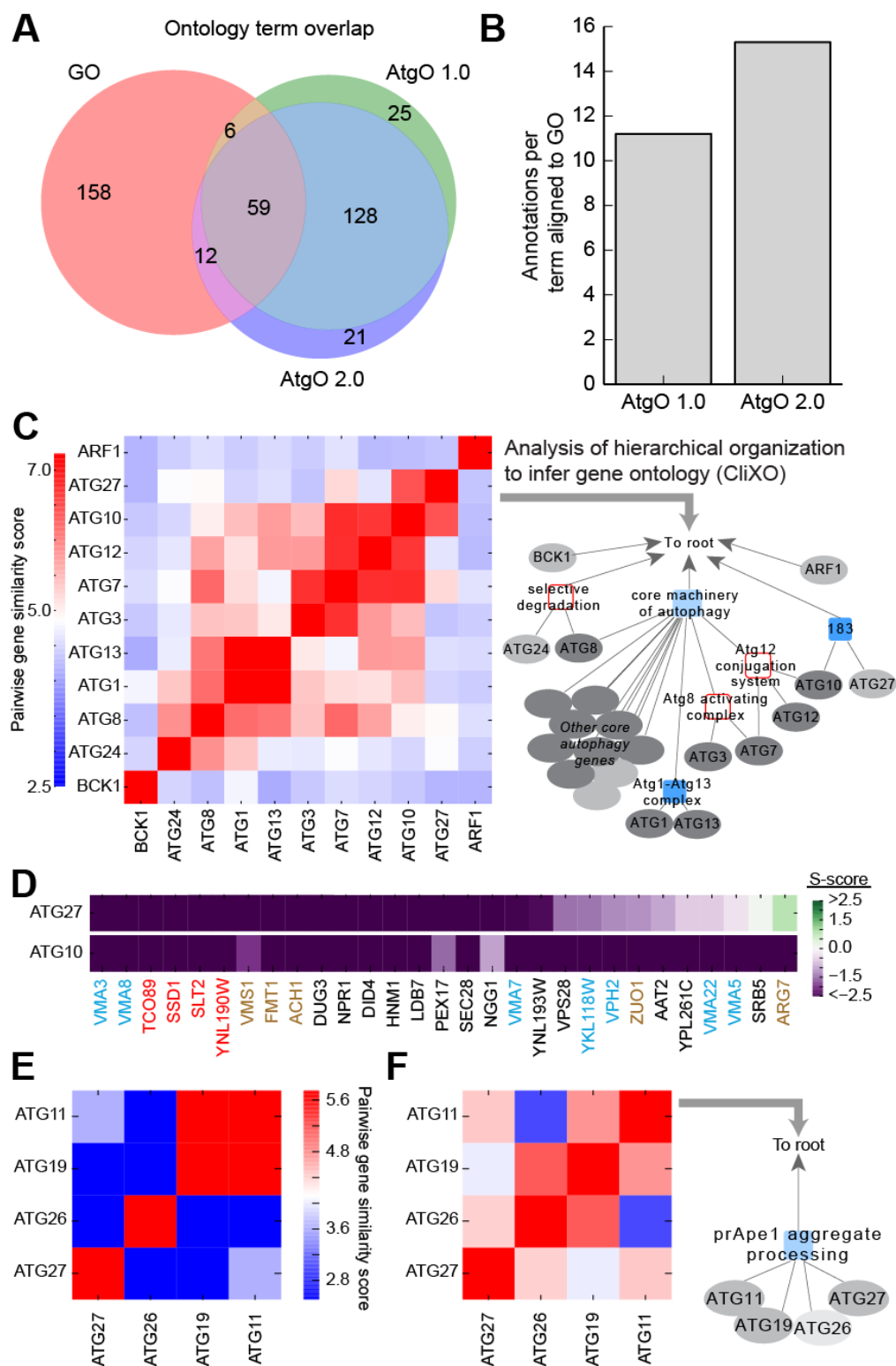
Next, we evaluated the extent to which these new networks improved the AtgO ontology. When used as the sole source of data, networks from each of the three static

conditions showed some ability to reconstruct GO ($r = 0.13 - 0.18$, **Figure 4E**). The differential networks showed better performance ($r = 0.23 - 0.31$), consistent with previous reports on the utility of differential interaction mapping (Bandyopadhyay et al., 2010; Ideker and Krogan, 2012). Integrating the three static and three differential conditions into a single network yielded the best correspondence with GO ($r = 0.42$), suggesting that multiple conditions reveal different aspects of cell biology. Remarkably, this performance was better than all previous data combined ($r = 0.30$; **Figure 4F**). Integrating previous with new data performed best of all ($r = 0.48$); we designate the ‘second-generation’ ontology of 220 terms resulting from these integrated data as AtgO 2.0 (**Figure 5**). The terms defined by the new AtgO 2.0 were similar to AtgO 1.0 (**Figure 6A**), suggesting convergence of ontology assembly over progressive AI-MAP iterations. The substantial increase in recovery of GO (**Figure 4F**) was driven by a modest increase in number of terms that aligned to GO (**Figure 6A**) and a larger increase in number of genes that could now be annotated to those common terms (**Figure 6B**).

For example, term AtgO:35 aligned to GO (“late nucleophagy”, GO:0044805, **Figure 3D i**) and was included in both AtgO versions, but expanded from 9 genes in AtgO 1.0 to 17 genes in AtgO 2.0 (**Figure 3B i** versus **Figure 5B i**). Of these, 15 are core autophagy genes (Jin and Klionsky, 2013) and the remaining 2, Atg11 and Atg15, have known core-related functions: Atg11 is a scaffold that links cargo receptors with core machinery and partially substitutes for the core function of Atg17 (Feng et al., 2014; Kamber et al., 2015); Atg15 is required for release of autophagy cargos in all general and selective pathways (Teter et al., 2001). Although 14 of the 17 AtgO:35 genes are annotated to GO:“late nucleophagy”, none is nucleophagy-specific. Hence, we name AtgO:35 “core

Figure 5. Second-generation ontology. (A) AtgO 2.0 hierarchical model. Rectangle color shows term alignment to the previous AtgO 1.0; otherwise, display follows Fig. 3 convention. AtgO:249:“autophagy and related processes” and its descendants are displayed, showing gene annotations for branches beneath AtgO:15:“ late secretory pathway / vacuolar membrane dynamics, cell growth and stress responses” **(B)** and AtgO:18:“vesicle trafficking, docking and fusion (especially golgi, some autophagy / vacuole), glucose sensing and response” **(C)**. All term names have been manually curated by our team, thus may differ from previous versions of AtgO or GO.

Figure 6. Characterization of AtgO 2.0. (A) Overlap in terms identified by GO, AtgO 1.0 or AtgO 2.0 calculated by alignment of ontologies. The Venn diagram includes GO terms for which >50% of annotated genes are in the 492 autophagy-related genes defined in the main text; note this includes a much larger set of terms than are included under GO:“autophagy” (**Figure 3D**). **(B)** Gene annotations per term aligned to GO in AtgO 1.0 and AtgO 2.0. **(C)** Pairwise gene similarity matrix and resulting AtgO 2.0 hierarchy for a subset of genes. **(D)** Genetic interaction profiles for Atg10 and Atg27, highlighting array genes with rapamycin-dependent negative interactions. Red names indicate genes involved in cellular integrity, blue indicates vacuolar-ATPase and brown indicates mitochondria. **(E-F)** The refinement of pairwise gene similarities from AtgO 1.0 **(E)** to AtgO 2.0 **(F)**, resulting in expansion of AtgO:185 which we name “prApe1 aggregate processing”.



machinery of autophagy” (**Figure 5B i**) and propose Atg11 and Atg15 as core autophagy genes.

AtgO terms capture both uncured knowledge and new discoveries

Additions in AtgO 2.0 relative to GO represent two classes of novel cellular componentry: [1] those that are supported by prior studies missed in GO curation, and [2] those that are potential new biological discoveries. Towards the first class, AtgO:138 recovered the Ubp3p-Bre5p ubiquitin protease complex implicated in ribophagy (**Figure 5C i**) (Kraft et al., 2008); AtgO:160 recovered the Pho85-Pho80 CDK-cyclin complex which regulates autophagy (**Figure 5B ii**) (Yang et al., 2010); AtgO:106 recovered Skm1, Ste20 and Cla4, part of a complex that downregulates sterol uptake (**Figure 5A i**) (Lin et al., 2009). We submitted each of these terms to GO curators, and given the ample prior evidence all were accepted for inclusion in the Gene Ontology (GO:1990861, GO:1990860, and GO:1990872 respectively). Other terms missing from GO but supported by literature include AtgO:221, which captured an E1/E2 enzyme complex that activates Atg8 in a ubiquitin-like cascade (**Figure 5B iii**, **Figure 6C**), and AtgO:224, a second ubiquitin-like cascade activating Atg12 (**Fig 5B iv**, **Figure 6C**) (Kaiser et al., 2012). Numerous new AtgO gene annotations were also supported by prior studies: for instance term AtgO:96 was assigned Vac8, Nvj1, and Swb1, which form the nucleus-vacuole junction (**Figure 5B v**) during piecemeal microautophagy of the nucleus (Dawaliby and Mayer, 2010; Kvam and Goldfarb, 2004). Two of these genes (Nvj1, Swb1) were not annotated as such in GO and thus at our request were annotated to this term (GO:0034727). As another example, we found that eight genes annotated to autophagy by AtgO but not

GO already had literature support for autophagy phenotypes: Hog1 (Prick et al., 2006), Pep12 (Kanki et al., 2009), Snf1 (Wang et al., 2001), Stt4 (Wang et al., 2012), Tpk1 (Budovskaya et al., 2004), Vps4 (Nebauer et al., 2007), Vps51 (Reggiori et al., 2003), and Vps52 (Reggiori et al., 2003). Here too, these new gene annotations were accepted for inclusion in GO. Thus, our confidence in AtgO is increased by the discovery that much of it already has literature support, even when discrepant with the current GO. Furthermore, although the above knowledge had been previously reported, such knowledge is valuably captured, structured, and unified by AtgO as a central research resource.

We now turn to the second class of novelty in AtgO – newly discovered terms with strong data support in AtgO but not previously reported. For example, AtgO:183 involved two genes: Atg27, a membrane protein involved in Atg9 vesicle trafficking; and Atg10, which activates Atg12 (**Figure 5B vi**, **Figure 6C**). This term was present in AtgO 1.0 due to correlation in previously measured genetic interaction profiles (Costanzo et al., 2010) and strengthened in AtgO 2.0 due to consistent correlations in all static (0.73 – 0.81) and differential (0.54 – 0.67) networks. Atg27 and Atg10 exhibited significantly overlapping rapamycin-dependent negative interactions with genes related to cellular integrity and potentially other functions (**Figure 6D**, $p < 10^{-30}$); loss of either gene causes rapamycin-dependent sensitivity to cell wall insults via an unknown mechanism. As another example, AtgO:219 was a new term in AtgO 2.0 that grouped Atg8, a ubiquitin-like protein in bulk autophagy and selective cargo recruitment (Noda et al., 2010), and Atg24, a sorting nexin with roles in selective autophagy (Kanki and Klionsky, 2008). This relationship emerged due to the synergy of new with old data, resulting in a strong and specific integrated similarity score between these two genes (**Figure 6C**). We have provisionally named this

term “selective degradation” (**Figure 5B vii**), as Atg8 and Atg24 are among those genes necessary for most selective degradation. It remains to be seen what distinguishes these two genes from other selective autophagy genes.

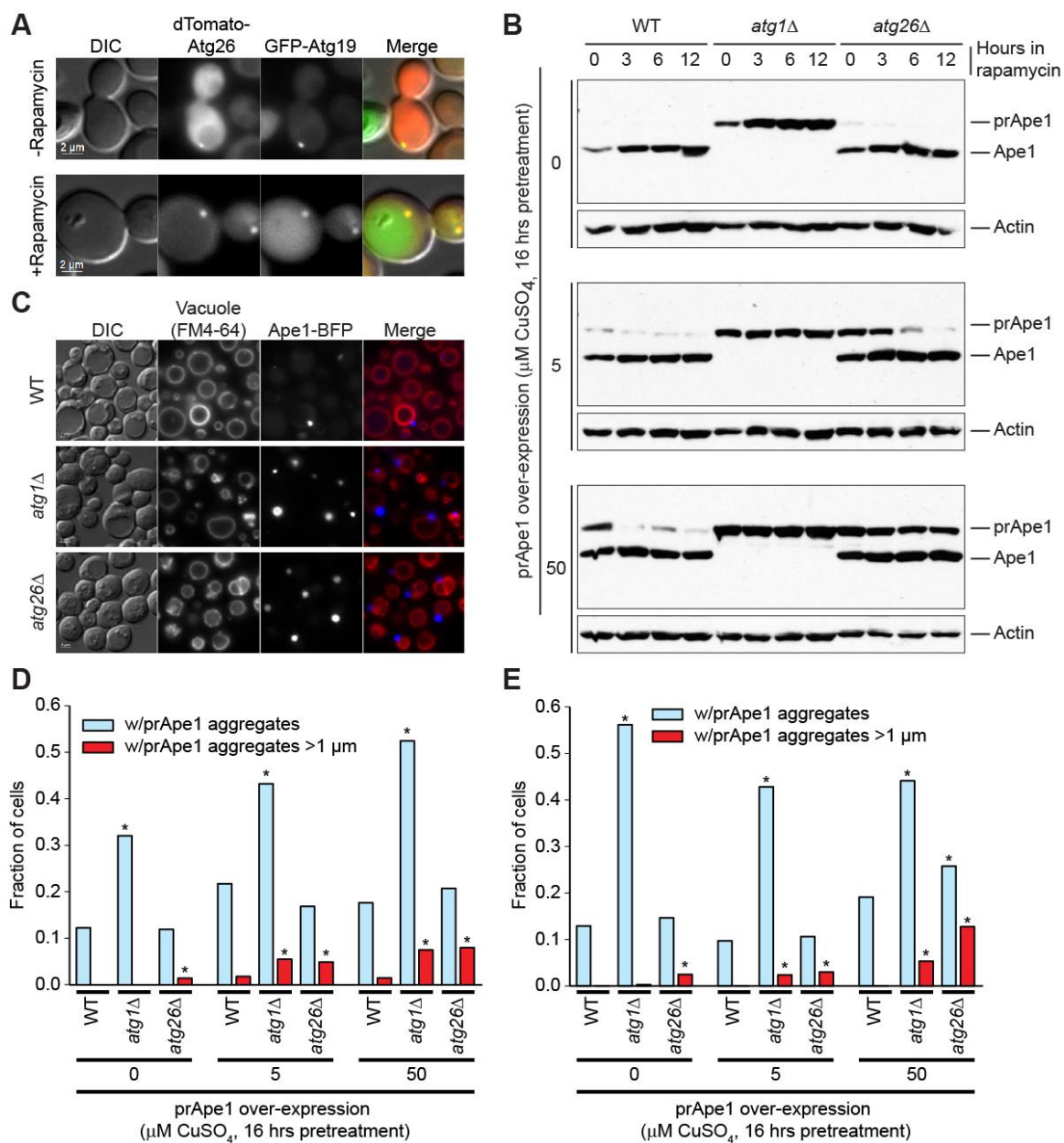
Atg26 is required for processing of large Ape1 aggregates in the Cvt pathway

One novel addition in AtgO 2.0 was the expansion of term AtgO:185 from two genes, Atg11 and Atg19 (**Figure 3C i**, **Figure 6E**), to include Atg26 and Atg27 (**Figure 5A ii**, **Figure 6F**). Three of these genes (Atg11, Atg19, Atg27; Yen et al., 2007; Yorimitsu and Klionsky, 2005) have known roles in the Cytoplasm-to-vacuole targeting pathway (Cvt pathway), by which aggregates of the aminopeptidase precursor prApe1 are transported to the vacuole and processed to mature Ape1. AtgO:185 implied involvement of the remaining gene, Atg26, in transport of prApe1 aggregates to the vacuole. Previous work, however, had failed to show a defect in prApe1 maturation in an *atg26Δ* knockout strain and also found no involvement of Atg26 in any autophagy-related pathway in *S. cerevisiae* (Cao and Klionsky, 2007). By contrast, in the yeast *Pichia pastoris*, the Atg26 ortholog participates in pexophagy, where it is required for the degradation of large peroxisomes; hence, its designation as an ‘Atg’ gene (Nazarko et al., 2009; Oku et al., 2003).

Based on these observations, we hypothesized that Atg26 is required for efficient processing of prApe1, especially larger prApe1 aggregates. In support of this hypothesis, we observed that Atg26 co-localizes in distinct, overlapping puncta with Atg19, the known receptor for prApe1 aggregates (**Figure 7A**). To test if Atg26 plays a role in processing large prApe1 aggregates, we moderately overexpressed prApe1 from a copper-inducible promoter to create larger aggregates of prApe1 (Geng et al., 2008; Suzuki et al., 2013) and

monitored its processing in *wild-type* and *atg26Δ* strains, as well as *atg1Δ* as a positive control for disruption of prApe1 processing (Cheong et al., 2008). Normal maturation of prApe1 to Ape1 was observed in *wild type* but not *atg1Δ* cells (**Figure 7B**). Atg26 was not required when prApe1 expression was normal (0 μ M CuSO₄), but *atg26Δ* mutants showed inefficient prApe1 processing, especially in cells with larger prApe1 aggregates (50 μ M CuSO₄, **Figure 7B**). Next, prApe1-BFP was expressed in the same system and cells were visualized by microscopy, revealing many large prApe1 aggregates in *atg26Δ* as compared to *wild type* cells (**Figure 7C**). To quantify large complex formation, >250 cells per strain were observed at increasing levels of prApe1 expression; cells with prApe1 aggregates were counted and large prApe1 aggregates >1 μ m were noted. Even in the absence of CuSO₄, *atg26Δ* cells contained significantly more large prApe1 aggregates than did *wild-type* cells ($p = 0.03$, **Figure 7D**); overexpression exacerbated this defect (5 μ M CuSO₄, $p = 0.02$; 50 μ M CuSO₄, $p = 0.0003$). Using rapamycin to induce macroautophagy, *wild-type* cells successfully processed all large prApe1 aggregates even when it was greatly overexpressed (50 μ M CuSO₄, **Figure 7E**). In contrast, many aggregates remained in the *atg26Δ* mutant after rapamycin treatment (**Figure 7E**). Taken together, these data support a new role for Atg26 in the processing of large prApe1 aggregates and validate term AtgO:185, which we hereby name “prApe1 aggregate processing”.

Figure 7. Atg26 is a component of the Cvt pathway. (A) Co-localization of dTomato-Atg26 and GFP-Atg19 observed by fluorescence microscopy. **(B)** Western blot of prApe1 and Ape1 across genetic backgrounds, with prApe1 expressed from a copper-inducible promoter and incubated 16 hours in SD-Cu media plus copper (CuSO_4) followed by removal from CuSO_4 and treatment in rapamycin for indicated times. **(C)** Fluorescence micrographs of cells expressing endogenous prApe1, prApe1-BFP from the endogenous promoter and prApe1 from the copper-inducible promoter. Cells incubated 16 hours in SD-Cu + $50\mu\text{M}$ CuSO_4 followed by 14 hours in SD-Cu media + rapamycin. **(D)** Analysis of cells with constructs as in **(C)**. Fraction of cells with observed prApe1 aggregates (light blue) and large prApe1 aggregates $>1\mu\text{M}$ (red) after 16 hours incubation in SD-Cu media + indicated CuSO_4 . >250 cells observed for each condition. Asterisk indicates significant difference from wildtype in same condition by Fisher's exact test. **(E)** Same as **(D)** but pre-treatment in CuSO_4 followed by 14 hours incubation in SD-Cu + rapamycin.



DISCUSSION

This study provides a prescription for how to progressively elucidate the ontology of cellular functions underlying any biological process. Studying a new process is straightforward, requiring only starting knowledge of some of the genes involved and access to public ‘omics data sets. Other information is provided by the Gene Ontology: importantly, GO is used only as a general guide for how heterogeneous data should be combined to build a functional hierarchy and does not require specific coverage of functions in the process of interest. At all points of the study, virtually any genomic data can be analyzed since almost any data linked to genes or proteins can inform gene-gene similarity networks; here, we combined gene and protein interactions, gene expression data, sequence features, and information about evolutionary conservation. Flexibility is also provided by the way integration of the various data sets is trained to reconstruct current knowledge. Here, we used the current state of the *S. cerevisiae* GO for training, but other species or knowledge bases could be specified.

Regardless of this training, the current AtgO 2.0 model, assembled entirely from network data, is larger and has differences in content and structure from the literature-curated GO autophagy branch. Close inspection reveals that the majority of AtgO terms not in GO are consistent with known biology (**Figure 5**) while others represent reasonable new biological hypotheses (**Figure 5, Figure 6C,D,F**). Some differences between the models may arise due to differences in policy between the AI-MAP process and human curation. For example, AtgO groups core autophagy genes under the single term “core machinery of autophagy”, whereas GO repetitively annotates each core gene to all general and selective autophagy pathways including “macroautophagy,” “microautophagy,” and

“nucleophagy.” Furthermore, GO generally does not have terms grouping paralogous genes of similar function. By contrast, AtgO treats functionally similar paralogs the same as any functional gene set and creates a term (e.g., “PDK homolog kinases,” “MARK homolog kinases,” “ADP ribosylation factors”). Also of interest are GO terms not currently recapitulated in AtgO (**Figure 3D, 6A**). Some of these terms may be included in future AtgO versions as new network data arise.

One compelling result was the identification of a new gene functioning in autophagy in *S. cerevisiae*, Atg26. Because AtgO:185 (**Figure 5A ii, 6F**) grouped Atg26 with other genes involved in prApe1 aggregate processing (Atg11, Atg19 and Atg27), we were compelled to investigate whether it plays a role in prApe1 transport to the vacuole in *S. cerevisiae* analogous to its role in pexophagy in *P. pastoris*; indeed this was the case (**Figure 7**). The commonality between the roles of Atg26 in *P. pastoris* and *S. cerevisiae* is that engulfment of large cargos, whether prApe1 aggregates or peroxisomes, requires larger double-membrane vesicles. These larger vesicles, in turn, require increased membrane delivery to the phagophore assembly site, likely facilitated by Atg26 in addition to Atg11, Atg19 and Atg27 (Backues et al., 2015).

An important aspect of Active Interaction Mapping is that heterogeneous new and existing data are interwoven in the same modeling framework. This enables a comprehensive assessment of the current support for a biological model (**Figure 4A**), as well as what type of information is most needed now (**Figure 4B**). After data generation, the value of the new dataset is systematically evaluated in the context of all existing data (**Figures 4E,F**). Here, such assessments pointed us to genetic interaction maps, leading us to generate a new targeted dataset which was more powerful at recovering knowledge about

autophagy in GO than all previous public datasets combined (**Figure 4F**). Given that large genetic interaction networks were already available for *Saccharomyces* (Bandyopadhyay et al., 2010; Costanzo et al., 2010; Fiedler et al., 2009), it is perhaps remarkable that further genetic interaction data turned out to be so useful. While we do not rule out other possibilities, the significant information gained by including multiple conditions (**Figure 4E**) argues that condition-specific interaction maps, and especially differential interactions between conditions, provide a particularly worthwhile bolus of data.

In future many research communities, each focused on a cellular process of interest, may find it useful to organize around shared hierarchical models as a means for representing cell biological knowledge, communicating new findings, and designing targeted datasets. Once a baseline model of a process is created, the value of any new data can be evaluated by its ability to improve this baseline. As we have demonstrated, such evaluation can involve a quantitative score which represents the impact of the new experiments in light of all prior studies. Thus, the value of new interaction mapping efforts can be rigorously evaluated rather than assumed, and the design of these experiments can be guided rationally, based on current knowledge and data. This work moves systems biology towards a future in which most new ‘omics data can contribute to an organized effort of hierarchical modeling and discovery.

EXPERIMENTAL PROCEDURES

Public data

Data from YeastNet v3 (Kim et al., 2014) were obtained from <http://www.inetbio.org/yeastnet/>, consisting of the following numbers of datasets classified by type: 50 co-expression, 1 domain co-occurrence, 1 genomic neighbor, 10 genetic interaction, 12 high-throughput protein-protein interaction, 1 phylogenetic profile and 1 protein network tertiary structure.

Data integration

Input datasets were provided as features to a random forest regression system for prediction of gene-gene pairwise similarity, trained using pairwise gene similarities in GO as a bronze standard. The network of predicted similarities we call the ‘data-derived gene similarity network’ See **Supplemental Experimental Procedures**.

Selection of autophagy-relevant genes

To focus on the process of autophagy, we selected a ‘seed’ of 20 genes designated as having core functions common to multiple autophagic processes (Jin and Klionsky, 2013). We identified additional genes based on similarity to this seed set in prior knowledge (GO) or data (**Figure 2B**). See **Supplemental Experimental Procedures**.

Construction of Autophagy Ontology (AtgO) 1.0

We analyzed the data-derived gene similarity network using the Clique Extracted Ontology (CliXO) algorithm, version 0.3 (http://mhk7.github.io/clixo_0.3/; Kramer et al., 2014), with parameters $\alpha=0.1$ and $\beta=0.5$. We identified significantly aligned terms between AtgO and GO using a previously described ontology alignment procedure with FDR threshold of 10% and minimum alignment score of 0.24

(<http://mhk7.github.io/alignOntology/>; Dutkowski et al., 2013). Alignment was performed first against the sub-branch of GO rooted at GO:“autophagy” and then the entire GO, prepared as in Dutkowski et al., 2013; significant alignments to the autophagy sub-branch were given first priority.

Genetic interaction mapping

Standard protocols used. Details described in **Supplemental Experimental Procedures**.

Construction of Autophagy Ontology (AtgO) 2.0

The data-derived gene similarity network was recalculated incorporating both prior data and the new genetic interaction screen. The CliXO algorithm was applied with previous parameters, and the resulting AtgO 2.0 was aligned to AtgO 1.0 as described above. Term names in AtgO 2.0 were manually curated by our team, considering the names of aligned terms in GO, when available, as well as an extensive literature review to find, where possible, cohesive biological functions to explain the set of genes annotated to each term.

Biochemical studies of prApe1 processing and Fluorescence microscopy

Standard protocols used. Details described in **Supplemental Experimental Procedures**.

AUTHOR CONTRIBUTIONS

MHK, MKY and TI designed the study and developed the conceptual ideas. MHK constructed and analyzed AtgO. KM, JCF and KL designed and performed the genetic interaction screen. MHK processed genetic interaction screen data. MHK and JCF curated AtgO 2.0 term names. RB and JMC investigated and curated new Gene Ontology terms

and annotations. JCF designed and performed prApe1 processing experiments. MHK and TI wrote the manuscript with input from other authors.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institutes of Health to TI (R01-ES014811, R01-GM084279, P50-GM085764) and to SS (R01-DK41737). Additional funding in support of trainees was provided to MHK (F30-HG007618, T32-GM007198) and MY (T32-GM008806). We wish to thank Dr. Claudine Kraft for plasmids pCK782, pDP105, pDP103 and pCK770. The authors wish to declare no competing financial interests related to this work.

Chapter 4, in full, is being revised and prepared for resubmission for publication. Kramer M*, Farre JC*, Mitra K, Yu MK, Licon K, Balakrishnan R, Cherry JM, Subramani S, Ideker T. “Active Interaction Mapping reveals the hierarchical organization of autophagy.” 2016. The dissertation author was the primary investigator and co-first author of this paper.

REFERENCES

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25, 25-29.
- Atias, N., Gershenzon, M., Labazin, K., and Sharan, R. (2014). Experimental design schemes for learning Boolean network models. *Bioinformatics* 30, i445-452.
- Backues, S.K., Orban, D.P., Bernard, A., Singh, K., Cao, Y., and Klionsky, D.J. (2015). Atg23 and Atg27 act at the early stages of Atg9 trafficking in *S. cerevisiae*. *Traffic* 16, 172-190.
- Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Geunole, A., van Attikum, H., Shokat, K.M., Kolodner, R.D., Huh, W.K., Aebersold, R., Keogh, M.C., Krogan, N.J., Ideker, T. (2010). Rewiring of genetic networks in response to DNA damage. *Science* 330, 1385-1389.
- Barrett, C.L., and Palsson, B.O. (2006). Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. *PLoS computational biology* 2, e52.
- Bean, G.J., Jaeger, P.A., Bahr, S., and Ideker, T. (2014). Development of ultra-high-density screening tools for microbial "omics". *PloS one* 9, e85177.
- Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. *Nature* 466, 68-76.
- Budovskaya, Y.V., Stephan, J.S., Reggiori, F., Klionsky, D.J., and Herman, P.K. (2004). The Ras/cAMP-dependent protein kinase signaling pathway regulates an early step of the autophagy process in *Saccharomyces cerevisiae*. *The Journal of biological chemistry* 279, 20663-20671.
- Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. *Cell* 158, 903-915.
- Cao, Y., and Klionsky, D.J. (2007). Atg26 is not involved in autophagy-related pathways in *Saccharomyces cerevisiae*. *Autophagy* 3, 17-20.
- Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., Lasorella, A., Aldape, K., Califano, A., Iavarone, A. (2010). The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 463, 318-325.

Cheong, H., Nair, U., Geng, J., and Klionsky, D.J. (2008). The Atg1 kinase complex is involved in the regulation of protein recruitment to initiate sequestering vesicle formation for nonspecific autophagy in *Saccharomyces cerevisiae*. *Molecular biology of the cell* 19, 668-681.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., Prinz, J., St. Onge, R.P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F.J., Alizadeh, S., Bahr, S., Brost, R.L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.Y., Liang, W., Marback, M., Paw, J., San Luis, B.J., Shuteriqi, E., Tong, A.H., van Dyk, N., Wallace, I.M., Whitney, J.A., Weirauch, M.T., Zhong, G., Zhu, H., Houry, W.A., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F.P., Giaever, G., Nislow, C., Troyanskaya, O.G., Bussey, H., Bader, G.D., Gingras, A.C., Morris, Q.D., Kim, P.M., Kaiser, C.A., Myers, C.L., Andrews, B.J., Boone, C. (2010). The genetic landscape of a cell. *Science* 327, 425-431.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., and Palsson, B.O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92-96.

Creixell, P., Schoof, E.M., Simpson, C.D., Longden, J., Miller, C.J., Lou, H.J., Perryman, L., Cox, T.R., Zivanovic, N., Palmeri, A., Wesolowska-Anderson, A., Helmer-Citterich, M., Ferkinghoff-Borg, J., Itamochi, H., Bodenmiller, B., Erler, J.T., Turk, B.E., Linding, R. (2015). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* 163, 202-217.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Caletani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Rust, A.G., Pan, Z., Schilstra, M.J., Clarke, P.J., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., Bolouri, H. (2002). A genomic regulatory network for development. *Science* 295, 1669-1678.

Dawaliby, R., and Mayer, A. (2010). Microautophagy of the nucleus coincides with a vacuolar diffusion barrier at nuclear-vacuolar junctions. *Molecular biology of the cell* 21, 4173-4183.

Dutkowski, J., Kramer, M., Surma, M.A., Balakrishnan, R., Cherry, J.M., Krogan, N.J., and Ideker, T. (2013). A gene ontology inferred from molecular networks. *Nature biotechnology* 31, 38-45.

Ergun, A., Lawrence, C.A., Kohanski, M.A., Brennan, T.A., and Collins, J.J. (2007). A network biology approach to prostate cancer. *Molecular systems biology* 3, 82.

Feng, Y., He, D., Yao, Z., and Klionsky, D.J. (2014). The machinery of macroautophagy. *Cell research* 24, 24-41.

Fiedler, D., Braberg, H., Mehta, M., Chechik, G., Cagney, G., Mukherjee, P., Silva, A.C., Shales, M., Collins, S.R., van Wageningen, S., Kemmeren, P., Holstege, F.C., Weissman, J.S., Keogh, M.C., Koller, D., Shokat, K.M., Krogan, N.J. (2009). Functional organization of the *S. cerevisiae* phosphorylation network. *Cell* 136, 952-963.

Geng, J., Baba, M., Nair, U., and Klionsky, D.J. (2008). Quantitative analysis of autophagy-related protein stoichiometry by fluorescence microscopy. *The Journal of cell biology* 182, 129-140.

Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., CHasman, D.I., FitzGerald, G.A., Dolinski, K., Grosser, T., Troyanskaya, O.G. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* 47, 569-576.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., Babu, M., Craig, S.A., Hu, P., Wan, C., Vlasblom, J., Dar, V.U., Bezginov, A., Clark, G.W., Wu, G.C., Wodak, S.J., Tillier, E.R., Paccanaro, A., Marcotte, E.M., Emili, A. (2012). A census of human soluble protein complexes. *Cell* 150, 1068-1081.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sørensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., Tyers, M. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180-183.

Ideker, T., Galitski, T., and Hood, L. (2001a). A new approach to decoding life: systems biology. *Annual review of genomics and human genetics* 2, 343-372.

Ideker, T., and Krogan, N.J. (2012). Differential network biology. *Molecular systems biology* 8, 565.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. (2001b). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929-934.

Ideker, T.E., Thorsson, V., and Karp, R.M. (2000). Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, 305-316.

Jin, M., and Klionsky, D.J. (2013). The Core Molecular Machinery of Autophagosome Formation. In *Autophagy and cancer*, H.G. Wang, ed. (New York: Springer), pp. 25-45.

Kaiser, S.E., Mao, K., Taherbhoy, A.M., Yu, S., Olszewski, J.L., Duda, D.M., Kurinov, I., Deng, A., Fenn, T.D., Klionsky, D.J., Schulman, B.A. (2012). Noncanonical E2 recruitment by the autophagy E1 revealed by Atg7-Atg3 and Atg7-Atg10 structures. *Nature structural & molecular biology* 19, 1242-1249.

Kamber, R.A., Shoemaker, C.J., and Denic, V. (2015). Receptor-Bound Targets of Selective Autophagy Use a Scaffold Protein to Activate the Atg1 Kinase. *Molecular cell* 59, 372-381.

Kanki, T., and Klionsky, D.J. (2008). Mitophagy in yeast occurs through a selective mechanism. *The Journal of biological chemistry* 283, 32386-32393.

Kanki, T., Wang, K., Baba, M., Bartholomew, C.R., Lynch-Day, M.A., Du, Z., Geng, J., Mao, K., Yang, Z., Yen, W.L., Klionsky, D.J. (2009). A genomic screen for yeast mutants defective in selective mitochondria autophagy. *Molecular biology of the cell* 20, 4730-4738.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jr., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389-401.

Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S., Shim, J.E., and Lee, I. (2014). YeastNet v3: a public database of data-specific and integrated functional gene networks for *Saccharomyces cerevisiae*. *Nucleic acids research* 42, D731-736.

Kim, Y.A., Wuchty, S., and Przytycka, T.M. (2011). Identifying causal genes and dysregulated pathways in complex diseases. *PLoS computational biology* 7, e1001095.

King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., Sparkes, A., Whelan, K.E., Clare, A. (2009). The automation of science. *Science* 324, 85-89.

King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G., Bryant, C.H., Muggleton, S.H., Kell, D.B., and Oliver, S.G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 427, 247-252.

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662-1664.

Kraft, C., Deplazes, A., Sohrmann, M., and Peter, M. (2008). Mature ribosomes are selectively degraded upon starvation by an autophagy pathway requiring the Ubp3p/Bre5p ubiquitin protease. *Nature cell biology* 10, 602-610.

Kramer, M., Dutkowski, J., Yu, M., Bafna, V., and Ideker, T. (2014). Inferring gene ontologies from pairwise similarity data. *Bioinformatics* 30, i34-42.

Kvam, E., and Goldfarb, D.S. (2004). Nvj1p is the outer-nuclear-membrane receptor for oxysterol-binding protein homolog Osh1p in *Saccharomyces cerevisiae*. *Journal of cell science* 117, 4959-4968.

Lefebvre, C., Rieckhof, G., and Califano, A. (2012). Reverse-engineering human regulatory networks. *Wiley interdisciplinary reviews Systems biology and medicine* 4, 311-325.

Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M.S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G.A., Lopez-Bigas, N., Getz, G., Ding, L., Raphael, B.J. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics* 47, 106-114.

Levine, B., and Yuan, J. (2005). Autophagy in cell death: an innocent convict? *The Journal of clinical investigation* 115, 2679-2688.

Liang, X.H., Jackson, S., Seaman, M., Brown, K., Kempkes, B., Hibshoosh, H., and Levine, B. (1999). Induction of autophagy and inhibition of tumorigenesis by beclin 1. *Nature* 402, 672-676.

Lin, M., Unden, H., Jacquier, N., Schneider, R., Just, U., and Hofken, T. (2009). The Cdc42 effectors Ste20, Cla4, and Skm1 down-regulate the expression of genes involved in sterol uptake by a mitogen-activated protein kinase-independent pathway. *Molecular biology of the cell* 20, 4826-4837.

Malleshaiah, M.K., Shahrezaei, V., Swain, P.S., and Michnick, S.W. (2010). The scaffold protein Ste5 directly controls a switch-like mating decision in yeast. *Nature* 465, 101-105.

Nazarko, T.Y., Farre, J.C., and Subramani, S. (2009). Peroxisome size provides insights into the function of autophagy-related proteins. *Molecular biology of the cell* 20, 3828-3839.

Nebauer, R., Rosenberger, S., and Daum, G. (2007). Phosphatidylethanolamine, a limiting factor of autophagy in yeast strains bearing a defect in the carboxypeptidase Y pathway of vacuolar targeting. *The Journal of biological chemistry* 282, 16736-16743.

Noda, N.N., Ohsumi, Y., and Inagaki, F. (2010). Atg8-family interacting motif crucial for selective autophagy. *FEBS letters* 584, 1379-1385.

Oku, M., Warnecke, D., Noda, T., Muller, F., Heinz, E., Mukaiyama, H., Kato, N., and Sakai, Y. (2003). Peroxisome degradation requires catalytically active sterol glucosyltransferase with a GRAM domain. *The EMBO journal* 22, 3231-3241.

Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155, 1008-1021.

Picotti, P., Clement-Ziza, M., Lam, H., Campbell, D.S., Schmidt, A., Deutsch, E.W., Rost, H., Sun, Z., Rinner, O., Reiter, L., Shen, Q., Michaelson, J.J., Frei, A., Alberti, S., Kusebauch, U., Wollscheid, B., Moritz, R.L., Beyer, A., Aebersold, R. (2013). A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* 494, 266-270.

Prick, T., Thumm, M., Haussinger, D., and Vom Dahl, S. (2006). Deletion of HOG1 leads to Osmosensitivity in starvation-induced, but not rapamycin-dependent Atg8 degradation and proteolysis: further evidence for different regulatory mechanisms in yeast autophagy. *Autophagy* 2, 241-243.

Reggiori, F., Komatsu, M., Finley, K., and Simonsen, A. (2012). Autophagy: more than a nonselective pathway. *International journal of cell biology* 2012, 219625.

Reggiori, F., Wang, C.W., Stromhaug, P.E., Shintani, T., and Klionsky, D.J. (2003). Vps51 is part of the yeast Vps fifty-three tethering complex essential for retrograde traffic from the early endosome and Cvt vesicle completion. *The Journal of biological chemistry* 278, 5009-5020.

Rubinsztein, D.C., DiFiglia, M., Heintz, N., Nixon, R.A., Qin, Z.H., Ravikumar, B., Stefanis, L., and Tolkovsky, A. (2005). Autophagy and its possible roles in nervous system diseases, damage and repair. *Autophagy* 1, 11-22.

Rubinsztein, D.C., Marino, G., and Kroemer, G. (2011). Autophagy and aging. *Cell* 146, 682-695.

Suzuki, K., Akioka, M., Kondo-Kakuta, C., Yamamoto, H., and Ohsumi, Y. (2013). Fine mapping of autophagy-related proteins during autophagosome formation in *Saccharomyces cerevisiae*. *Journal of cell science* 126, 2534-2544.

Talloczy, Z., Virgin, H.W.t., and Levine, B. (2006). PKR-dependent autophagic degradation of herpes simplex virus type 1. *Autophagy* 2, 24-29.

Teter, S.A., Eggerton, K.P., Scott, S.V., Kim, J., Fischer, A.M., and Klionsky, D.J. (2001). Degradation of lipid vesicles in the yeast vacuole requires function of Cvt17, a putative lipase. *The Journal of biological chemistry* 276, 2083-2087.

Tong, A.H., and Boone, C. (2006). Synthetic genetic array analysis in *Saccharomyces cerevisiae*. *Methods in molecular biology* 313, 171-192.

Wang, K., Yang, Z., Liu, X., Mao, K., Nair, U., and Klionsky, D.J. (2012). Phosphatidylinositol 4-kinases are required for autophagic membrane trafficking. *The Journal of biological chemistry* 287, 37964-37972.

Wang, R.C., and Levine, B. (2010). Autophagy in cellular growth control. *FEBS letters* 584, 1417-1426.

Wang, Z., Wilson, W.A., Fujino, M.A., and Roach, P.J. (2001). Antagonistic controls of autophagy and glycogen accumulation by Snf1p, the yeast homolog of AMP-activated protein kinase, and the cyclin-dependent kinase Pho85p. *Molecular and cellular biology* 21, 5742-5752.

Wu, X., Won, H., and Rubinsztein, D.C. (2013). Autophagy and mammalian development. *Biochemical Society transactions* 41, 1489-1494.

Yang, Z., Geng, J., Yen, W.L., Wang, K., and Klionsky, D.J. (2010). Positive or negative roles of different cyclin-dependent kinase Pho85-cyclin complexes orchestrate induction of autophagy in *Saccharomyces cerevisiae*. *Molecular cell* 38, 250-264.

Yen, W.L., Legakis, J.E., Nair, U., and Klionsky, D.J. (2007). Atg27 is required for autophagy-dependent cycling of Atg9. *Molecular biology of the cell* 18, 581-593.

Yorimitsu, T., and Klionsky, D.J. (2005). Atg11 links cargo to the vesicle-forming machinery in the cytoplasm to vacuole targeting pathway. *Molecular biology of the cell* 16, 1593-1605.

SUPPLEMENTARY MATERIALS AND METHODS

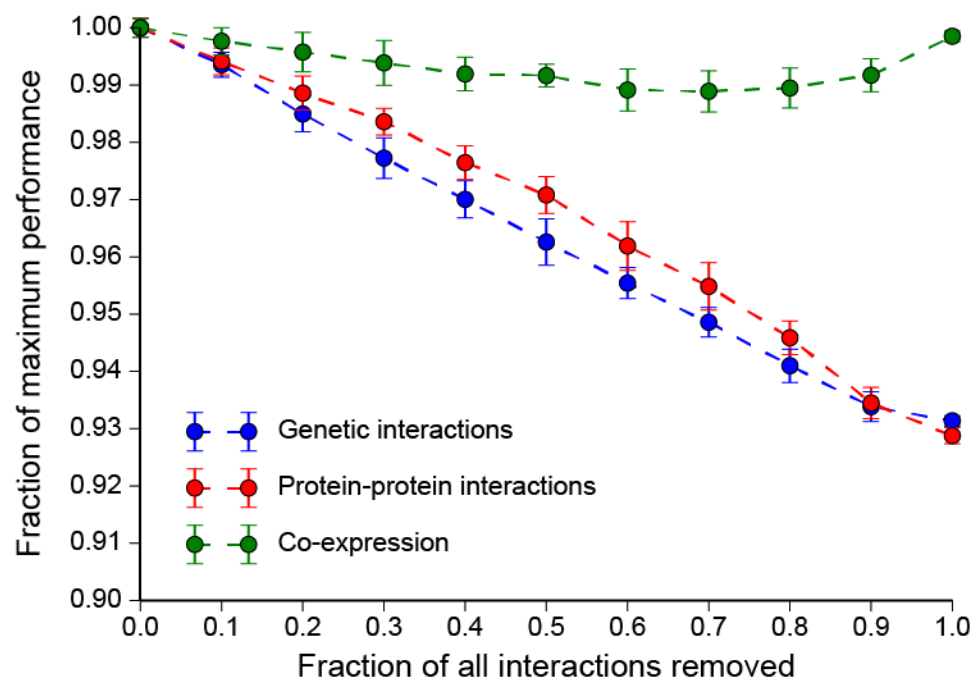


Figure S1. Performance degradation as interactions are removed. Performance degradation as more interactions of the indicated type (color) are cumulatively excluded. The mean degradation (point) and standard deviation (error bars) are calculated over 50 random sets of interactions. See also **Figure 4B**.

Supplemental Experimental Procedures

Data integration

We aimed to use the Gene Ontology (GO) as a standard reference for learning to integrate data in a supervised fashion. Previously, we showed that a weighted network of gene-gene similarities derived from a gene ontology can be analyzed to reconstruct the full hierarchy of terms and term relations with near perfect precision and recall (Kramer et al., 2014). Because the network and the ontology can be interconverted, both contain the same information. Based on this equivalence, we guided our integration of data using a gene similarity network derived from GO instead of using GO directly. We derived this network by calculating the Resnik semantic similarity for each pair of genes (Resnik, 1995). Semantic similarities are calculated across the Biological Process and Cellular Component branches of GO, downloaded on 6/2/2015 from www.geneontology.org.

The 76 input datasets were integrated into a single network by using them as features in a supervised learning of the GO gene similarity network. Learning was performed using random forest regression (Breiman, 2001) from the Python scikit-learn package (Pedregosa et al., 2011). Predictions were made “out of bag”, i.e., the similarity of a gene pair was predicted based on information learned from other gene pairs. In effect, the random forest learns patterns in the networks which recapitulate information in GO. Hence, only relations in GO that can be systematically explained from data are included, any relations not justified by the data are excluded, and new relations not in GO are added when the network data support them (**Fig 2A**).

Selection of autophagy-relevant genes

To focus on the process of autophagy, we selected a ‘seed’ of 20 genes designated as having core functions common to multiple autophagic processes (Jin and Klionsky, 2013). We identified additional genes based on similarity to this seed set in prior knowledge (GO) or data (**Figure 2B**). For each candidate gene, we computed three scores: (1) average similarity to the seed set in the GO gene similarity network, (2) average similarity to the seed set in the data-derived gene similarity network, and (3) maximum similarity to the seed set in the data-derived gene similarity network. Score (1) was normalized so that the lowest score across genes annotated to the GO:“autophagy” term is 1. Scores (2) and (3) were each normalized so that the lowest score across seed genes is 1. For each candidate gene, we estimated its relevance to autophagy by taking the maximum of these three individual scores. Genes for which this maximum is at least 1 were kept for further analyses.

Genetic interaction screening

Strain construction, plating of mutants, mutant selection, and scoring of genetic interactions in each condition were performed using a previously defined protocol (Collins et al., 2006; Schuldiner et al., 2006). Using a replica pinning robot (Singer Instruments; UK), haploid double mutants representing crosses of 52 autophagy related query genes and 3007 array mutants were grown on agar plates that were either untreated, rapamycin treated or nitrogen deficient. Plates were photographed and colony sizes normalized, spatially corrected and quantified using the Colony Analyzer Toolkit (Bean et al., 2014). For the three replicates per double mutant, the resulting experimental data were used to assign a quantitative S-score based on a modified T-test that compares the

observed double mutant growth rate to that expected assuming no interaction exists, again using the Colony Analyzer Toolkit. Differential interactions were calculated by subtracting the S-scores for the same double mutant pair across conditions.

Biochemical studies of prApe1 processing

Cells were grown for 16 h in SD-Cu media (1x yeast nitrogen base [YNB] without amino acids and copper, 2% glucose plus any required amino acid, nucleotide, or vitamin supplement) without exceeding exponential phase (OD₆₀₀ above 1) and transferred to SD-Cu medium with 200 nM Rapamycin and the indicated CuSO₄ concentration at an OD₆₀₀ of 1. One ml of cells was collected at different times as described in the **Figure 7B**, TCA precipitated and analyzed by Western blot probing it with anti-Ape1 and anti-Actin antibodies.

Fluorescence microscopy

Microscopy was performed as described for biochemical studies and pictures were acquired. To quantify prApe1 aggregates, the fluorescence microscopy parameters such as exposition time, gain and binning were kept constant. Images were captured on a motorized fluorescence microscope (Axioskop 2 MOT, Carl Zeiss MicroImaging, Thornwood, NY) coupled to a monochrome digital camera (AxioCam MRm, Carl Zeiss MicroImaging) and processed using the AxioVision 4.8.2 software.

Supplemental References

Breiman, L. (2001). Random forests. *Mach Learn* 45, 5-32.

Collins, S.R., Schuldiner, M., Krogan, N.J., and Weissman, J.S. (2006). A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome biology* 7, R63.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12, 2825-2830.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Int Joint Conf Artif*, 448-453.

Schuldiner, M., Collins, S.R., Weissman, J.S., and Krogan, N.J. (2006). Quantitative genetic analysis in *Saccharomyces cerevisiae* using epistatic miniarray profiles (E-MAPs) and its application to chromatin functions. *Methods* 40, 344-352.

CHAPTER 5

Reprint of “Translation of Genotype to Phenotype by a Hierarchy of Cell Systems.”
Published February 24, 2016 in *Cell Systems*.

Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems

Michael Ku Yu,^{1,2} Michael Kramer,^{2,3} Janusz Dutkowski,^{2,4} Rohith Srivas,^{2,5} Katherine Licon,² Jason F. Kreisberg,² Cherie T. Ng,⁶ Nevan Krogan,⁷ Roded Sharan,⁸ and Trey Ideker^{2,*}

¹Bioinformatics and Systems Biology Program

²Department of Medicine

³Biomedical Sciences Program

University of California, San Diego, La Jolla, CA 92093, USA

⁴Data4Cure, La Jolla, CA 92037, USA

⁵Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA

⁶aTyr Pharmaceuticals, San Diego, CA 92121, USA

⁷Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94143, USA

⁸Blavatnik School of Computer Science, Tel-Aviv University, Tel Aviv 69978, Israel

*Correspondence: ideker@ucsd.edu

<http://dx.doi.org/10.1016/j.cels.2016.02.003>

SUMMARY

Accurately translating genotype to phenotype requires accounting for the functional impact of genetic variation at many biological scales. Here, we present a strategy for genotype-phenotype reasoning based on existing knowledge of cellular subsystems. These subsystems and their hierarchical organization are defined by the Gene Ontology or a complementary ontology inferred directly from previously published datasets. Guided by the ontology's hierarchical structure, we organize genotype data into an “ontotype,” that is, a hierarchy of perturbations representing the effects of genetic variation at multiple cellular scales. The ontotype is then interpreted using logical rules generated by machine learning to predict phenotype. This approach substantially outperforms previous non-hierarchical methods for translating yeast genotype to cell growth phenotype, and it accurately predicts the growth outcomes of two new screens of 2,503 double gene knockouts affecting DNA repair or nuclear lumen. Ontotypes also generalize to larger knockout combinations, setting the stage for interpreting the complex genetics of disease.

INTRODUCTION

A central problem in genetics is to understand how different variations in DNA sequence, dispersed across a multitude of genes, can nonetheless elicit similar phenotypes (Waddington, 1942). In recent years, it has been repeatedly observed that different genetic drivers of a trait can be recognized by their aggregation in networks of pairwise protein or gene interactions (Califano et al., 2012; Greene et al., 2015; Hanahan and Weinberg, 2011; Kim and Przytycka, 2012; Ramanan et al., 2012; Wang et al., 2010). Rather than associating genotype with phenotype

directly, variations in genotype are first mapped onto knowledge of gene networks; affected subnetworks are then statistically associated with phenotype. This approach can greatly increase our power to identify relevant associations between genotype and phenotype. This principle of “network-based” or “pathway-based” association (Califano et al., 2012) is now being applied to effectively map the genetics underlying complex phenotypes, including cancer and other common diseases (Hofree et al., 2013; Lee et al., 2011; Leiserson et al., 2014; Ng et al., 2012; Pe'er and Hachohen, 2011; Skafidas et al., 2014; Sullivan, 2012; Willsey et al., 2013).

In these studies, network knowledge is represented as a set of genes and pairwise gene interactions. In reality, however, genotype is transmitted to phenotype not only through gene-gene interactions but through a rich hierarchy of biological subsystems at multiple scales: genotypic variations in nucleotides (1 nm scale) give rise to functional changes in proteins (1–10 nm), which in turn affect protein complexes (10–100 nm), cellular processes (100 nm), organelles (1 μ m), and, ultimately, phenotypic behaviors of cells (1–10 μ m), tissues (100 μ m to 100 mm), and complex organisms (>1 m). What has been less well studied in genotype-phenotype association is how to leverage our extensive pre-existing knowledge across these scales or how to identify the scales most relevant to a set of genetic variants (Deisboeck et al., 2011; Eissing et al., 2011; Walpole et al., 2013).

In many fields, knowledge across scales is modeled by ontologies: a factorization of prior knowledge about the world into a hierarchy of increasingly specific concepts (Brachman and Levesque, 2004). For instance, intelligent systems like Apple's Siri and IBM's Watson carry out logical reasoning using a large collection of world knowledge represented by ontologies (Carvunis and Ideker, 2014). In molecular and cellular biology, extensive knowledge of the hierarchy of subsystems in a cell has been represented by the Gene Ontology (GO), a community standard reference database that documents interrelationships among thousands of intracellular components, processes, and functions in a large hierarchy of terms (Gene Ontology Consortium, 2015). Thus far, genotype-phenotype association methods have sometimes used prior knowledge in the GO by flattening the term hierarchy to a network, in which pairwise interactions

connect genes annotated with the same GO term (Pesquita et al., 2009). This flattening, however, may discard important information about the rich hierarchy of biological systems connecting genotype to phenotype. Moreover, a hierarchical model is highly complementary, and in some ways orthogonal, to flat networks: the GO is primarily concerned with “deep” connectivity up and down a hierarchy of cellular processes spanning dozens of scales, whereas network models typically focus on horizontal flow of signaling, transcriptional, or metabolic information among genes or reactions at the same scale (Lee et al., 2010, 2011). Another advantage of the GO is that it is continuously improved by a very large community of dozens of curators and editors, who update GO from new knowledge published in thousands of peer-reviewed papers each year (Balakrishnan et al., 2013; Huntley et al., 2014). To complement this process of manual curation, recently we and others have shown that a large hierarchy of cellular systems can be systematically assembled directly from analysis of genome-wide datasets, including molecular interactions and gene expression profiles; we call this assembly NeXO (Dutkowski et al., 2013; Gligorijević et al., 2014; Kramer et al., 2014). This “data-driven” ontology closely resembles, and in some cases greatly revises and expands, the literature-curated GO.

Here we report a general approach for using deep hierarchical knowledge of the cell, represented by an ontology, to translate genotype to phenotype. This approach recursively aggregates the effects of genetic variation upward through the hierarchy; in this way, genetic variants comprising genotype are converted to effects on the cell subsystems affected by those variants. We call the set of all such effects an “ontotype,” representing variation at intermediate scales between nanoscopic changes in genes and macroscopic changes in phenotype.

Here, we focus on yeast genetic interactions, in which the deletion of two or more genes results in an unexpectedly slow or fast cellular growth phenotype. Genetic interactions have previously been screened systematically using synthetic genetic arrays in yeast (Costanzo et al., 2010); these experiments comprise ~3 million different genetic backgrounds and are among the largest genotype-phenotype compendia in existence. We integrate these data with the GO to produce a multi-scale computational model, the functionalized ontology. The model accurately predicts growth phenotypes of 2,503 previously untested double-deletion genotypes, and it is also capable of predicting the phenotypes that result from larger combinations of gene disruptions. Similar predictive power is achieved by substituting the GO with NeXO, our data-driven ontology of cellular systems. In aggregate, this work suggests a strategy for building hierarchical models of the cell whose structure and function are learned completely from data.

RESULTS

Association between Genetic Interactions and Hierarchical Relations among Cellular Systems

As preparation for modeling, we identified patterns by which genetic interactions are associated with, and thus biologically explained by, the structure of gene ontologies. We observed that sets of genes assigned to the same GO term tended to be highly enriched for genetic interactions ($p < 10^{-5}$), for both positive

genetic interactions (double gene disruptions with better than expected growth, e.g., epistasis) and negative genetic interactions (double gene disruptions with worse than expected growth, e.g., synthetic lethality) (Figure 1A). Such interaction enrichment within GO terms occurred over a wide range of term sizes—the number of genes annotated to a term—suggesting that genetic interactions emerge from both broad and specific cellular mechanisms at multiple scales.

Because of the hierarchical structure of the cell, genetic interactions among genes annotated to a term can potentially be re-interpreted as interactions between the genes of different terms at a lower scale in the GO. For example, the “parent” term “microtubule-associated complex” displays strong within-term interaction enrichment, which factors into strong between-term interaction enrichment across two of its “children” terms, kinesin and dynactin (Figure 1B). We found that such hierarchical relationships were widespread in the GO: approximately half of within-term enrichments could be factored into between-term enrichments among their descendants (Figure 1C). Occurrences of interactions within or between biological pathways have been previously investigated as separate biological interpretations (Bandyopadhyay et al., 2008; Bellay et al., 2011; Collins et al., 2010; Kelley and Ideker, 2005; Leiserson et al., 2011; Ma et al., 2008; Qi et al., 2008; Ulitsky et al., 2008). Here, both types of explanations can be applied to the same interaction, as they are related hierarchically within the unified structure of the cell. Overall, approximately 40,000 interactions were involved in 1,661 within- or between-term enrichments, representing a 24:1 compression of information (Figure 1D). Thus, the GO integrates genetic interactions in an overarching hierarchy capturing multiple scales of cell biology. As one moves upward in this hierarchy, separate disruptions to multiple systems converge to multiple disruptions to a single system, with the scale of this transition indicated naturally by the hierarchical structure.

The Ontotype: An Intermediate between Genotype and Phenotype

Guided by this concordance between the GO hierarchy and genetic interactions, we developed a general system for ontology-based translation of genotype to phenotype that involves three general steps. First, the genotype is described according to convention by the set of genes that have been disrupted relative to wild-type (e.g., *b1d1*; Figure 2A). These disruptions are propagated recursively up the ontology, such that every term is assigned the disrupted genes annotated to that term plus all of those assigned to its children. For example, because the gene *KIP1* encodes a subunit of the kinesin complex (Figure 1B), its deletion in a *kip1Δ* strain propagates upward in the ontology to affect the parent term “kinesin complex” and continues to propagate upward to affect ancestor terms at higher scales such as “microtubule associated complex” and “cytoskeleton.”

Second, every term is assigned a functional state, representing the aggregate impact of gene disruptions on the activity of the component or process that term represents. Although it is possible to envisage many ways one might compute this functional impact, as proof of principle, we explored a simple and parameter-free computation, the number of disrupted genes associated with the term. This general approach is iterated across all terms; we call the profile of states across all terms

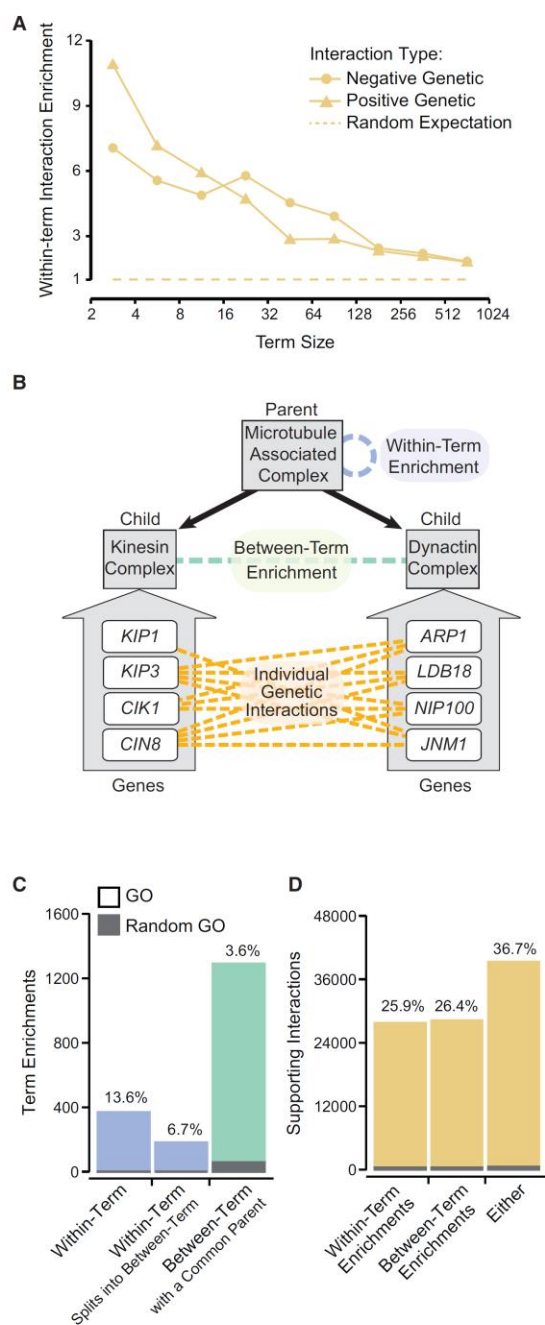


Figure 1. Patterns of Genetic Interaction Reflect the Hierarchical Structure of the GO

(A) Enrichment for negative (circle) or positive (triangle) genetic interactions among genes annotated to the same GO term as a function of term size, measured by the number of genes annotated to that term or its descendants.

the “ontotype.” In this way, the ontotype provides a complete picture of cell function and spans scales between genotype and phenotype. Whereas genotype describes the states of genes, and phenotype describes the states of observable traits, ontotype describes the states of all known biological objects. Many of these objects exist at scales bigger than genes but too small to be classically observable by eye, such as protein complexes and other subcellular structures, or too diffuse, such as signaling pathways (Figure 2A). In its most general definition, ontotype encompasses both genotype and phenotype, with genes and observable traits positioned at lower and higher levels of the hierarchy of objects encoding life.

A Functionalized Gene Ontology Integrating Cell Structure and Functional Prediction

Third, once genotypes are transformed to ontotypes, a supervised learning approach based on the technique of random forests regression (Breiman, 2001) is used to learn rules by which term states predict phenotypes. Rules are organized as a collection, or “forest,” of decision trees (Experimental Procedures), with a typical decision tree describing a series of logical true-or-false tests to evaluate the states of several terms (e.g., T4, T5, and T7 in Figure 2A). Making decisions on the states of terms rather than nucleotide variants or genes enables machine learning across a range of scales, so that different genotypes converging on similar ontotypes (e.g., *aΔdΔ* and *bΔdΔ* in Figure 2B) can yield the same phenotype. Decision tree logic was trained to predict quantitative genetic interaction scores from ~3 million tests for pairwise genetic interactions (Costanzo et al., 2010) (Experimental Procedures). This hierarchical structure of the ontology, when coupled to the decision logic described above, forms a “functionalized” ontology, that is, a computational cell model that defines both the sub-structures of the cell and how these sub-structures hierarchically translate genotype to phenotype.

Separate functionalized ontologies were trained using either the GO curated from the *Saccharomyces* literature (Cherry et al., 2012) (F_{GO}) or a data-driven ontology assembled from *Saccharomyces* datasets using the method of network-extracted ontologies (Dutkowski et al., 2013; Kramer et al., 2014) (F_{NeXO}). Whereas the GO represents knowledge of published cell biology, application of NeXO yielded an ontology whose hierarchy of cell systems was learned directly from publicly available data,

Enrichment is normalized as the fold change over expected for randomized GO annotations.

(B) Genetic interactions are propagated up the GO hierarchy to support “between-term enrichment” between the dynactin and kinesin complexes and “within-term enrichment” within the parent “microtubule associated complex.”

(C) Number of within-term and between-term enrichments highlighted by current genetic interaction data. Approximately half of within-term enrichments can be factored into one or more between-term enrichments that occur lower in the GO hierarchy. Percentages are calculated with respect to the total possible tests for within-term (2,719) and between-term (36,210) enrichments. (D) Number of genetic interactions involved in a within-term, between-term, or either type of enrichment. Percentages are calculated with respect to the total number of genetic interactions (107,133). The expected numbers of enrichments (C) and supporting interactions (D) were also calculated over randomized GO annotations (dark gray bars).

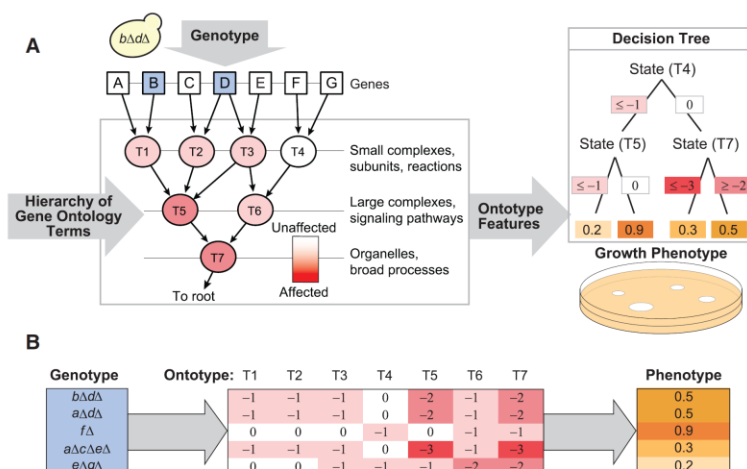


Figure 2. The Ontotype Method of Translating Genotype to Phenotype

(A) The relationship between genotypic and phenotypic variation is modeled through an intermediate “ontotype,” defined as the profile of states corresponding to the effect of genotype on each cellular component, biological process, and molecular function represented as a term in GO. To generate an ontotype, perturbations to genes are propagated hierarchically through the ontology, altering term states. A random forest regresses to predict a phenotype using the ontotype as features. An example decision tree from the forest is shown.

(B) Example genotype, ontotype, and phenotype associations from the ontology in (A). Different genotypes (e.g., *bΔdΔ* and *aΔdΔ*) give rise to similar or identical phenotypes by influencing similar or identical combinations of terms.

including protein-protein interactions, gene expression profiles, and protein sequence properties, but excluding any prior information about genetic interactions (datasets taken from the YeastNet v3 study; Kim et al., 2014). NeXO (4,805 terms) was tuned so that the resulting ontology was approximately similar in size to the GO (5,125 terms). Alignment of these two ontologies revealed 1,614 significantly overlapping terms. Thus, NeXO represents a distinct hierarchy of cellular systems that provides an alternative to the hierarchy maintained by GO curators.

Quantitative Assessment of Performance for Genotype-Phenotype Translation

F_{GO} accurately predicted growth phenotypes across a range of genetic interaction scores (Figures 3A and 3B). The correlation between predicted and measured scores was highly significant (Figure 3C; Pearson’s $r = 0.35$, $p < 2.2 \times 10^{-16}$) and reduced substantially when a randomized version of the ontology was used ($r = 0.04$); the maximum achievable correlation, as previously determined by experimental genetic interaction replicates (Bar-yshnikova et al., 2010), was $r = 0.67$. Progressively removing either small or large terms from the model degraded the correlation (Figures 3D and 3E), indicating that all scales in the hierarchy aid in prediction. F_{NeXO} achieved nearly the same correlation (Figure 3C; $r = 0.32$) and was also sensitive to randomization ($r = 0.03$).

Both functionalized ontologies compared favorably with non-hierarchical approaches for predicting genetic interactions (Boucher and Jenna, 2013; Lehner, 2013). We evaluated three state-of-the-art methods: flux balance analysis (FBA), which uses a mechanistic model of yeast metabolic pathways to simulate the impact of gene deletions on cell growth (Szappanos et al., 2011); guilt by association (GBA), which predicts the phenotype of pairwise gene deletions on the basis of the phenotypes of their network neighbors (Lee et al., 2010); and the multi-network multi-classifier (MNMC), a “black box” supervised learning system that uses many different lines of experimental evidence as features to predict genetic interactions (Pandey et al., 2010) (Experimental Procedures). In comparison with all

of these approaches, the functionalized ontologies achieved substantially greater correlation between predicted and measured interaction scores (Figure 3C) as well as better trade-offs in precision versus recall (Figure 3F) in 4-fold cross-validation. We also assessed prediction performance in a challenging validation scenario in which the training set of genotypes does not disrupt any genes in the test set (Park and Marcotte, 2012) (Supplemental Experimental Procedures). In this scenario, any genotype-phenotype logic that applies to individual genes is no longer generalizable; for example, promiscuous genes with a high degree of genetic interactions (Gillis and Pavlidis, 2012; Mackay, 2014) could be used to explain training data but not test data. Despite this challenge, F_{GO} still outperformed predictions made with a randomized GO or with the non-hierarchical methods (Figure S1).

We found that the accuracy of growth phenotype prediction depends significantly on the degree to which cellular systems have been characterized in the GO. F_{GO} was especially accurate at modeling genotypes for which the disrupted genes are well characterized by GO annotations; conversely, it was far less able to model genotypes for which the genes are poorly characterized (Figure S2). Moreover, many genes that are poorly characterized in the GO are better characterized in NeXO, such that genotypes involving these genes lead to better phenotypic predictions by F_{NeXO} than by F_{GO} (Figures S2A–S2C). These differences demonstrate the utility of data-driven ontologies for translating genotype to phenotype, especially in species that are lacking in GO curation but have “omics” datasets from which a gene ontology can nonetheless be built.

Finally, we investigated whether hierarchical features (i.e., the ontotype) were essential or if equally good predictions could be made from “flat” features derived from the same ontologies. The GO was flattened by computing the semantic similarity (Resnik, 1995), which scores every pair of genes by their functional relatedness in GO. As a non-hierarchical representation of NeXO, we directly considered the data on which it had been based: pairwise gene-gene similarities derived from different types of experimental evidence in YeastNet. Use of these flat datasets derived

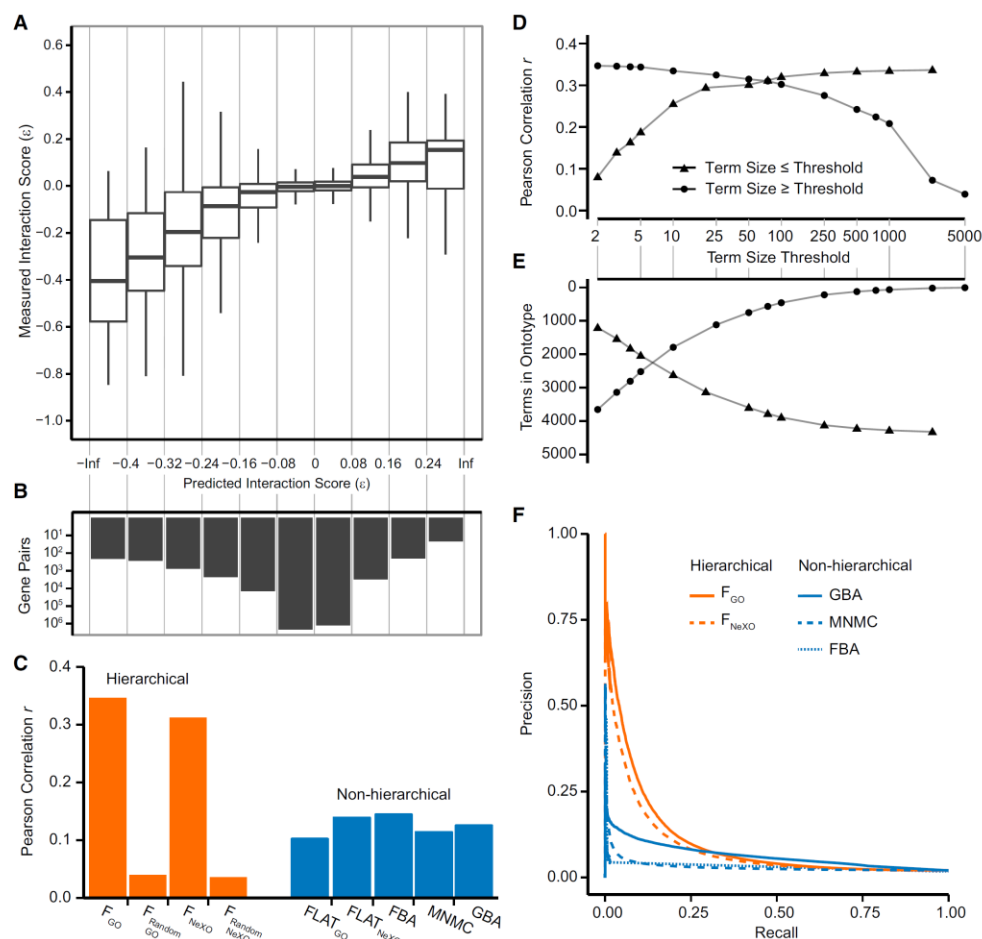


Figure 3. Genome-wide Prediction of Pairwise Genetic Interactions in Yeast

(A) Measured genetic interaction scores versus those predicted from ontotypes constructed from GO using 4-fold cross-validation. For each bin of predicted scores, box plots summarize the distribution of measured scores by its median (central horizontal line), interquartile range (box), and an additional 1.5σ (whiskers). (B) Number of gene pairs in each bin of predicted scores.

(C) Method performance, as represented by the correlation of measured versus predicted interaction scores across gene pairs that meet an interaction significance criterion of $p < 0.05$ in Costanzo et al. (2010). Comparison is made with ontotypes constructed from a randomized gene ontology or NeXO and to previous non-hierarchical methods for predicting genetic interactions. FBA correlation is reported for the set of 104,826 gene pairs considered by this model and for which gene annotations are available in the GO. The ontotype correlations do not fluctuate greatly ($<4\%$) whether computed over all gene pairs (shown) or the FBA gene pairs. See also Figures S1 and S2.

(D) Method performance when the ontotype is constructed from only GO terms that are no larger than (triangles) or no smaller than (circles) a size threshold.

(E) The number of GO terms that meet each size threshold criterion.

(F) Precision-recall curves for classification of negative genetic interactions.

from the two ontologies resulted in a substantial degradation in prediction performance (FLAT_{GO} and FLAT_{NeXO}; Figure 3C), even though the same random-forests regression procedure was used as for the functionalized ontologies.

Simulating Growth Phenotypes for “New” Genotypes Not Yet Observed or Examined

We next used F_{GO} to simulate growth for all 12,512,503 pairwise deletions of non-essential yeast genes, 73% of which

had not yet been tested in the laboratory (Figure 4A; Data S1). A total of 41,605 genetic interactions were predicted. These predictions were concentrated within and between particular terms and term pairs (Figures 4A and 4B), covering a total of 1,367 unique terms and indicating where in the ontology the logic of F_{GO} takes place. For example, F_{GO} predicted many genetic interactions within “oxidative phosphorylation” (Figure 4C), with negative interactions linking the sub-systems of electron and proton transport and positive

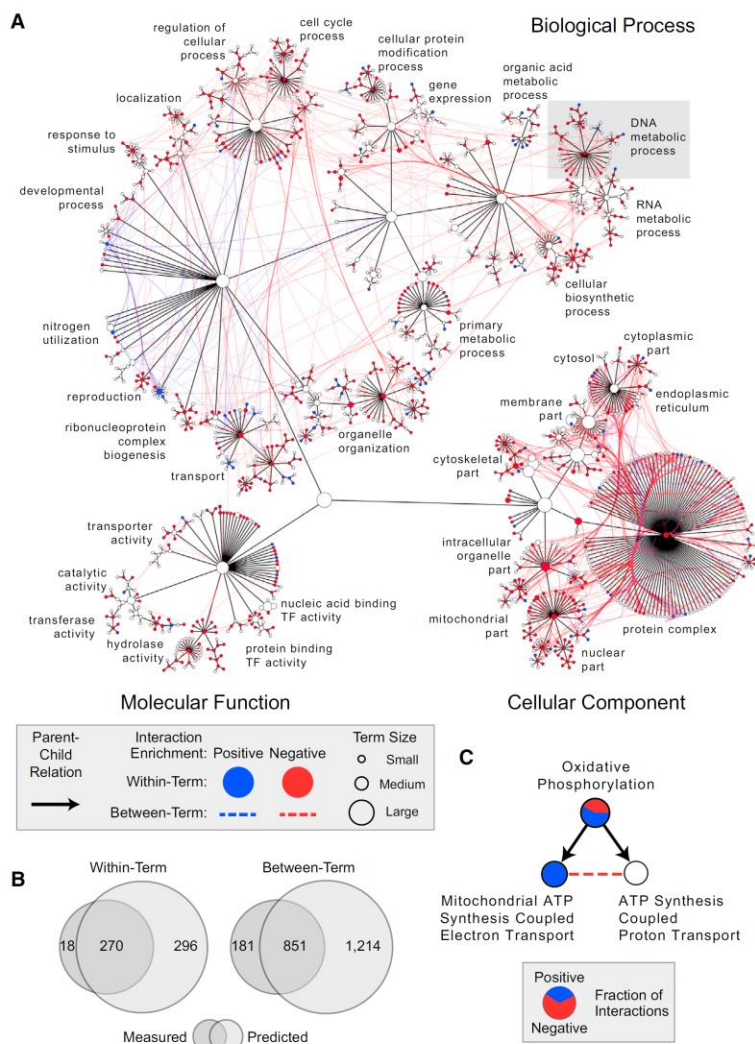


Figure 4. The Functionalized GO

(A) Visualization of F_{GO} structure and function. Terms and hierarchical parent-child relations are represented by nodes and black edges. Colored nodes and edges denote within- and between-term interaction enrichments, illustrating how terms and term combinations are used for prediction.

(B) Venn diagrams showing number of term enrichments identified for measured interactions, predicted interactions, or both.

(C) Example term “oxidative phosphorylation,” which factors into the transport of electrons (left child) versus protons (right child). Although both positive and negative genetic interactions are predicted within the oxidative phosphorylation genes (represented by a pie with both blue and red slices), positive interactions segregate within electron transport (blue pie) while negative interactions segregate between electron and proton transport (dotted red edge).

See also Figure S3.

Validation and Expansion of the Functionalized Ontology of DNA Repair and Nuclear Lumen

Key terms in F_{GO} were “DNA repair” and “nuclear lumen,” which featured prominently in the decision tree logic, leading to a high concentration of predicted interactions (9.0 and 7.6 times the expected interaction density, respectively) according to particular patterns of disruption (Figures 5A and S4). Genetic perturbations within each term led to particularly accurate growth phenotypes in cross-validation, as the correlation between predicted interactions and those measured by Costanzo et al. (2010) was noticeably better for gene pairs in DNA repair or nuclear lumen (both $r = 0.61$) than for gene pairs in other terms (average $r = 0.35$; Figure S2G; Table S2). To test whether this

performance generalized to new data, we experimentally measured growth phenotypes for 1,218 pairwise deletions of DNA repair genes and 1,600 pairwise deletions of nuclear lumen genes and scored these mutants for genetic interactions (Table S3; Supplemental Experimental Procedures). Of these, 1,345 mutants had also been scored previously by Costanzo et al. (2010). Surprisingly, we observed that the new measurements were better predicted by F_{GO} than by the previous measurements of those same genotypes (i.e., experimental replicates; Figure 5B). Such improvement suggests that functionalized ontologies may be able to reduce experimental noise by learning the overarching patterns of cellular subsystems that translate genotype to phenotype.

We next tested F_{GO} ’s ability to generalize to unseen mutant genotypes. For this purpose we constructed a “limited” F_{GO} , trained only on those genotypes that had been tested earlier

interactions segregating entirely within electron transport. These distinct patterns of positive and negative segregation were observed broadly across F_{GO} (Figure S3). Of particular interest were predicted interactions between 71 term pairs, as these terms were only distantly related in the GO (Table 1; Table S1; Supplemental Experimental Procedures). For example, all ten genes in “intron homing” had negative interactions with all four genes in the “Phosphatidylinositol-3-kinase complex,” although neither these terms nor their parents shared any genes, and these terms were in entirely separate branches of the GO (biological process versus cellular component). Thus, F_{GO} makes predictions guided by, but not rigidly confined to, known hierarchical relations among cellular subsystems. The unexpected connections point to potential new cellular functions and functional relationships important for regulating cell growth.

performance generalized to new data, we experimentally measured growth phenotypes for 1,218 pairwise deletions of DNA repair genes and 1,600 pairwise deletions of nuclear lumen genes and scored these mutants for genetic interactions (Table S3; Supplemental Experimental Procedures). Of these, 1,345 mutants had also been scored previously by Costanzo et al. (2010). Surprisingly, we observed that the new measurements were better predicted by F_{GO} than by the previous measurements of those same genotypes (i.e., experimental replicates; Figure 5B). Such improvement suggests that functionalized ontologies may be able to reduce experimental noise by learning the overarching patterns of cellular subsystems that translate genotype to phenotype.

Table 1. Top New Functional Relationships in F_{GO}

	Term A (Number of Genes)	Term B (Number of Genes)	Interactions/Total (%)		p Value ^a
			Predicted	Measured	
Negative Interactions	intron homing (10)	phosphatidylinositol 3-kinase complex II (4)	40/40 (100.0%)	2/2 (100.0%)	6.74E-96
	negative regulation of chromatin silencing at silent mating-type cassette (8)	protein import into mitochondrial inner membrane (3)	24/24 (100.0%)	14/14 (100.0%)	3.56E-55
	pre-mRNA binding (5)	RNA pol II transcription coactivator activity in preinitiation complex assembly (3)	15/15 (100.0%)	2/2 (100.0%)	2.86E-32
	protein lipoylation (4)	carbon-oxygen lyase activity, acting on phosphates (3)	12/12 (100.0%)	9/9 (100.0%)	1.23E-24
	Swr1 complex (8)	U6 snRNP (3)	22/24 (91.7%)	6/6 (100.0%)	1.20E-47
	alpha-1,6-mannosyltransferase complex (6)	negative regulation of chromatin silencing involved in replicative cell aging (4)	21/24 (87.5%)	1/4 (25.0%)	3.08E-44
	tubulin complex assembly (5)	maintenance of DNA trinucleotide repeats (3)	13/15 (86.7%)	12/12 (100.0%)	3.67E-25
	inositol phosphate biosynthetic process (5)	minus-end-directed microtubule motor activity (3)	12/15 (80.0%)	5/6 (83.3%)	5.56E-22
	regulation of ARF GTPase activity (6)	phosphatidylinositol-3,5-bisphosphate 5-phosphatase activity (4)	19/24 (79.2%)	2/4 (50.0%)	7.92E-38
	positive regulation of RNA elongation from Pol I promoter (5)	HIR complex (4)	14/20 (70.0%)	8/10 (80.0%)	3.82E-25
Positive Interactions	negative regulation of chromatin silencing at silent mating-type cassette (8)	U6 snRNP (3)	19/24 (79.2%)	10/11 (90.9%)	7.92E-38
	tubulin complex assembly (5)	DNA-directed RNA polymerase I complex (4)	15/20 (75.0%)	8/8 (100.0%)	4.37E-28
	RSC complex (8)	inactivation of MAPK activity (4)	19/32 (59.4%)	3/4 (75.0%)	6.33E-34
	vacuolar proton-transporting V-type ATPase, V1 domain (8)	free ubiquitin chain polymerization (3)	14/24 (58.3%)	7/14 (50.0%)	1.91E-23
	alpha-1,6-mannosyltransferase complex (6)	dynactin complex (5)	16/30 (53.3%)	8/12 (66.7%)	1.14E-26
	vacuolar proton-transporting V-type ATPase, V0 domain (7)	AP-3 adaptor complex (4)	13/28 (46.4%)	7/16 (43.8%)	1.26E-19
	SLIK (SAGA-like) complex (14)	positive regulation of stress-activated MAPK cascade (3)	14/42 (33.3%)	12/21 (57.1%)	4.92E-19
	histone exchange (9)	minus-end-directed microtubule motor activity (3)	9/27 (33.3%)	7/22 (31.8%)	2.38E-10
	histone methyltransferase activity (H3-K4 specific) (7)	snoRNA transcription from an RNA polymerase II promoter (3)	7/21 (33.3%)	4/6 (66.7%)	7.33E-07
	glycerol transport (4)	transcription-coupled nucleotide-excision repair (4)	5/16 (31.3%)	1/4 (25.0%)	3.42E-03

See also Table S1.

^aCalculated with respect to predicted interactions. Bonferroni corrected for family wise error rate.

(Costanzo et al., 2010) but not by our new screens. This limited F_{GO} achieved high sensitivity versus specificity (Figure 5C) and precision versus recall (Figure 5D) in predicting the new interactions measured for DNA repair and nuclear lumen genes. Given this validation, we combined the genetic interaction scores from both new screens with previous data (Costanzo et al., 2010) and re-trained the ontotype decision logic on this more complete dataset. The structure of this improved F_{GO}, with the accompanying ontotype-phenotype logic, is available online on the Network Data Exchange (<http://goo.gl/cY1XWJ>; UUID: 01b46d52-c3a5-11e5-8fbc-06603eb7f303; Pratt et al., 2015) and as a Cytoscape file in Data S2.

Toward More Complex Genotypes

Although the ontotype had been trained using double-deletion genotypes, we hypothesized that once trained, it might be capable of predictions for genotypes involving mutations to larger numbers of genes. Although few studies have examined three-way or higher order genetic interactions, a recent study (Haber et al., 2013) showed proof of principle for a three-way gene deletion methodology, representing one of the few systematic screens for triple mutants to date. This work reported that deletion of *CAC1* in combination with any gene in the HIR complex (*HIR1*, *HIR2*, *HIR3*, and *HPC2*) results in a synthetic growth defect (negative genetic interaction); however, the

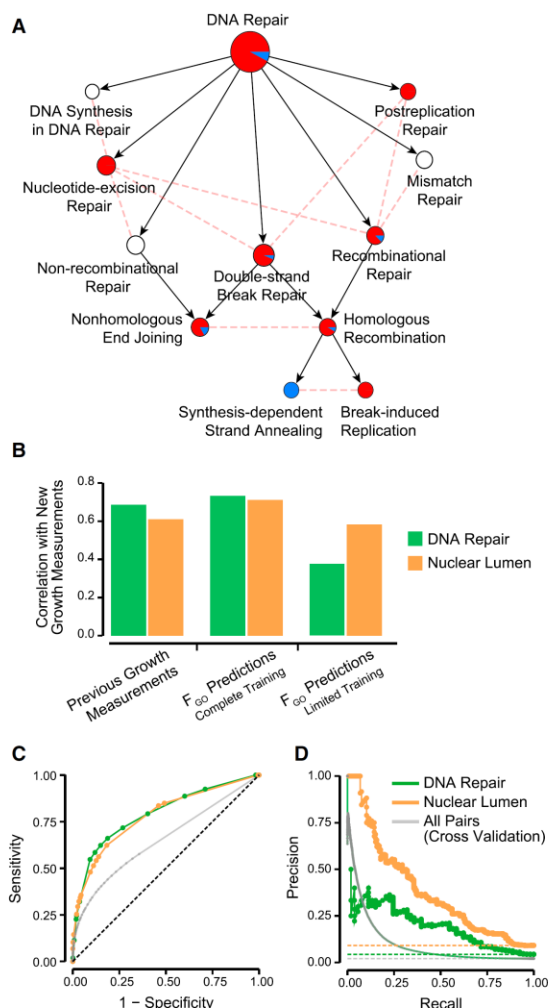


Figure 5. Elucidating the Genetic Logic of DNA Repair and the Nuclear Lumen

(A) DNA repair has a rich structure of predicted genetic interactions among specific repair processes. Coloring and visual style of panels follow the convention of Figure 4. See also Figure S4.

(B–D) Yeast growth was experimentally measured for double gene deletion strains in which both genes are involved in either DNA repair (green) or nuclear lumen (orange). See also Tables S2 and S3. (B) The new measurements are correlated with previous data by Costanzo et al. (2010) as well as predictions of a F_{GO} trained with all previous data or predictions of a “limited” F_{GO} trained with all previous data excluding genotypes tested in the new screen. In all cases, correlation is computed among the genotypes tested by both the new screen and Costanzo et al. (2010). Among all genotypes in the new screen, we calculated receiver-operating (C) and precision-recall curves (D) for predicting negative genetic interactions in DNA repair and the nuclear lumen using the limited F_{GO} . The corresponding curves across all gene pairs in the previous screen are reproduced for comparison (gray; see Figure 3F).

additional deletion of a third gene, *ASF1*, suppresses this phenotype. Consistent with these findings, F_{GO} predicted both the synthetic sickness of the double mutants and phenotypic suppression by the triple mutant (Figure 6A). Visual inspection of the model (Figure 6B) implicated decision logic on the basis of the functional activities of two related processes, DNA replication-independent nucleosome assembly and nucleosome organization. Deleting a single gene in DNA replication-independent nucleosome assembly leads to a state in which the deletion of another gene functioning elsewhere in nucleosome organization causes synthetic sickness. In contrast, the triple mutants include deletion of two genes in DNA replication-independent nucleosome assembly (*asf1* Δ HIR Δ), leading to a neutral phenotype. This effect probably occurs because the double mutant impairs growth to such an extent that additional perturbations have no detectable effect. Indeed, whereas *CAC1* is primarily involved in regulating DNA replication, *ASF1* and the HIR complex have been linked to other chromatin-related processes, including transcriptional elongation (Formosa et al., 2002; Schwabish and Struhl, 2006) and mRNA export (Pamblanco et al., 2014). This triple-mutant case study illustrates the complexity of logic in interpreting genetic interactions, underscoring the utility of a knowledge representation and reasoning system for unraveling such combinatorial genetic effects.

DISCUSSION

Many years of work by the Gene Ontology Consortium have established an extensive description of cell structure spanning a hierarchy of biological scales. Here, we have shown that the ontology structure can also be used functionally for interpretation of genetic variants to make phenotypic predictions. The ability to systematically map and then integrate these two aspects, structure and function, outlines a general strategy for development of computational cell models. First, a knowledge base of the cell’s hierarchical structure is acquired, either through literature curation (GO) or data-driven methods (NeXO). In a second step, mathematical relations are learned by algorithms that translate how the functional states of these subsystems—the ontology—give rise to a phenotype of interest. Thus, cell structure is determined from physical information derived from literature or systematic data, and cell function is learned from genetic data such as synthetic lethal interactions and genome-wide association studies.

Functionalized ontologies substantially outperformed previous phenotypic predictors (Figures 3C and 3F), a notable finding given the simplicity of the ontology and its use as the sole feature set for learning. We believe this success follows from several key aspects of implementation. First and most important, the utility of hierarchical organization in genotype-phenotype translation cannot be overstated. Indeed, the functionalized ontologies also outperformed predictors based on non-hierarchical versions of the same information (Figure 3C) or truncated versions of the ontology (Figures 3D and 3E). From the perspective of the ontology, all mutations or variants in a genotype coalesce to the same cellular module, provided one looks at a high enough level (Figure 1B). A genotype may include some mutations that map to the same gene, others to

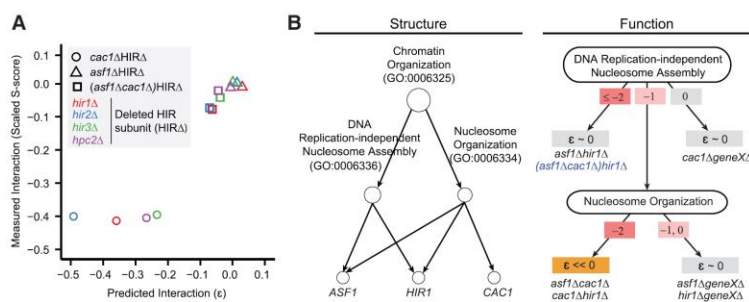


Figure 6. Prediction of Triple Mutants

(A) Measured versus predicted interaction scores for genotypes involving pairwise and three-way deletions involving *ASF1*, *CAC1*, and genes in the HIR complex (*HIR1*, *HIR2*, *HIR3*, and *HPC2*) (Haber et al., 2013).

(B) Relevant GO structure (left) and corresponding functional decision tree (right) for predicting the two- and three-way interactions in (A). At left, arrows represent parent-child relations and gene annotations in GO. At right, arrows represent decisions based on ontology: numbers on arrows are term states; arrows point to predicted interaction scores (ϵ).

the same protein complex, still others to the same broad process or organelle, with all mutations falling within the highest scale represented by the cell itself. Propagating mutations upward through terms of increasing scale enables subsequent selection of the “right” scale for accurate prediction. In this regard, F_{GO} sheds light on previous, partially discrepant, studies of genetic interaction networks. Some analyses have found that negative genetic interactions tend to connect between complementary modules, whereas positive interactions tend to occur within a single module (Bandyopadhyay et al., 2008; Collins et al., 2010; Kelley and Ideker, 2005; Leiserson et al., 2011; Ma et al., 2008; Qi et al., 2008; Ulitsky et al., 2008); a more recent report identified dense patterns of both positive and negative interactions between modules (Bellay et al., 2011). Analysis of F_{GO} suggests that both interpretations can be correct, depending on the scale of the module(s) within the cellular hierarchy.

The second factor in the success of functionalized ontologies is the sustained efforts of biologists at large. The GO is a rich resource of cellular knowledge that is both broad, in its extensive coverage of cell biology, and deep, in its resolution of cell subsystems across many different scales. Although not perfect, this knowledge is continuously refined, updated, and expanded by the sustained efforts of a global community. Given the staggering complexity of the cell, such a collaborative approach incorporating diverse expertise and tools may be instrumental in establishing robust and complete prior knowledge for computational cell modeling. Previously, cellular modeling efforts have typically involved independent curation within a single laboratory or institute.

A last factor that may have worked in our favor is the fact that functionalized ontologies balance rigid modeling constraints imposed by prior knowledge with flexible statistical learning guided by experiments. Computing the ontology requires no parameters and instead leverages the topology of the ontology. Logical rules for predicting phenotype are based on the ontology, but their functional form (i.e., which terms are used and how their states are combined) is learned from data. In contrast, many previous efforts in mechanistic modeling (e.g., see Cahan et al., 2014; Carrera et al., 2014; Deutscher et al., 2006; Karr et al., 2012; Lerman et al., 2012; Machado et al., 2011; O’Brien et al., 2013; Orth et al., 2010; Segrè et al., 2005; Szappanos et al., 2011; Szczurek et al., 2009; Takahashi et al., 2003; Tomita et al., 1999) have been driven by low-level prior knowledge in the form of biophysical equations. Although naturally conferring a

mechanistic explanation when correct, these equations have a known challenge that they are often of preset form and have sensitive parameters (Apgar et al., 2010; Ashyraliyev et al., 2009; Gutenkunst et al., 2007), such that achieving accurate predictions within one dataset risks overfitting.

Extending Functional Ontologies beyond Current Limits

F_{GO} based its predictions principally on 1,367 terms, spread across various biological processes, cellular components, and molecular functions (Figure 4A). Although this coverage of cell biology is substantial (27% of the yeast GO), one might wonder whether it should be more complete. First, some term logic is likely missed because those terms are not frequently disrupted in the current set of genotypes. For example, genes annotated to 783 GO terms were never disrupted in any genotype tested (Costanzo et al., 2010). Second, some biological processes are likely not required for the phenotype tested—growth of cells in rich media—but instead may drive a wide variety of other phenotypes (Dowell et al., 2010; Hillenmeyer et al., 2008; Ideker and Krogan, 2012; Lee et al., 2014). Third, important processes or components may not yet have been curated in the GO, and some existing terms might have errors in gene annotations or relations to other terms. Such false-positive and false-negative information could obscure a term’s utility in prediction. We expect that testing additional genotypes, phenotypes, and environmental conditions will increase the functional coverage of terms and enhance F_{GO} with new and more robust logic.

Complex traits arise from a landscape of genetic variants and mutations, in which it is often challenging to interpret the effects of individual genes because of many multi-gene interactions (Kim and Przytycka, 2012; Zuk et al., 2012). Toward meeting this challenge, we have shown that gene ontologies can be transformed into multi-scale models capable of general genotype-phenotype reasoning. Although based on simple rules of propagation, the model substantially outperforms previous methods for predicting cellular growth phenotypes, whether based on mechanistic modeling of pathways or “black-box” machine learning methods. It also generalizes in ways that previous predictors are incapable of doing, including the ability to analyze genotypes of arbitrary complexity. These advances are important steps toward building intelligent systems that can one day interpret the complex genetics underlying human health and disease.

In moving forward, special consideration should be given to the mathematical functions that govern each term state. Here,

we found success with a surprisingly straightforward and parameter-free function that counts the disrupted genes assigned to a term and its sub-terms. More generally, this function might be tailored to each term according to specific knowledge about the inner workings of that cellular component or process. Defining the mathematical relationships between genes within a cellular process has been the focus of “bottom-up” systems biology (Bruggeman and Westerhoff, 2007; Chen et al., 2010). In contrast, defining the broad organization of genes into cellular processes has been the domain of “top-down” systems biology. With its hierarchy of terms and functions spanning many different biological scales, a functionalized ontology may offer a means to bridge this long-standing divide.

EXPERIMENTAL PROCEDURES

Genetic Interaction Data

Experimental genetic interaction scores for >6 million double mutants in yeast, measured using synthetic genetic arrays (SGAs) (Costanzo et al., 2010) (1,711 queries $\times$ 3,885 arrays), were downloaded from <http://drygin.ccrb.utoronto.ca/~costanzo2009/>. Double gene deletion mutants affecting DNA repair and the nuclear lumen were generated on solid agar media using SGA technology as previously described (Collins et al., 2010; Tong and Boone, 2006). See also Supplemental Experimental Procedures.

Preparation of Ontologies

We used all three branches of the GO (biological process, cellular component, and molecular function) by joining them under an artificial root. We removed annotations with the evidence code “inferred by genetic interaction” (IGI) to avoid potential circularity in predicting genetic interactions. We also removed terms that were not annotated with any yeast genes or were redundant with respect to their children terms to construct a gene ontology relevant to yeast (Table S4), following a previously described procedure (Dutkowski et al., 2013; <http://mhk7.github.io/alignOntology/>).

To construct NeXO (Table S5), we integrated the YeastNet v3 networks (Kim et al., 2014), spanning 68 experimental studies across eight data types excluding genetic interactions, into a single network, and then applied the method of Clique Extracted Ontology (Kramer et al., 2014; http://mhk7.github.io/clixo_0.3/). Code for constructing ontotypes is available at <https://github.com/michaelkyu/ontotype>. See also Supplemental Experimental Procedures.

Random Forests Regression

Random forests (Breiman, 2001) were used to regress genetic interaction scores e_{ab} , as described in Results. Because of the very large size of the ontology feature matrix, we optimized the random forests library from the Python scikit-learn package (Pedregosa et al., 2011); the modified code is available at <https://github.com/michaelkyu/scikit-learn-fasterRF>. Although trees grown at approximately 29% (GO) or 37% (NeXO) of the maximal depth did improve performance slightly (<0.02 gain in correlation; Figure S5), we chose to grow trees to maximal depth because it is unclear how significant this gain is and whether it would be reproducible in different random partitions of the data for cross-validation or in different genotype-phenotype datasets. See also Supplemental Experimental Procedures.

Comparison of Methods for Predicting Genetic Interactions

We updated the MNMC method because the original implementation (Pandey et al., 2010) was trained on a set of literature-curated synthetic lethal interactions that was much smaller in size than the set of genetic interactions considered in our study, and because the set of features used by the method to score each gene pair had been updated since the 2010 publication. To train MNMC, we calculated five base features that were identified in the original MNMC as among the most informative for predicting synthetic lethality of a gene pair. This updated MNMC outperformed the original MNMC (Figure S6); this performance difference may have been due to calculating more recent versions of the base features. See also Supplemental Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, five tables, and six data files and can be found with this article online at <http://dx.doi.org/10.1016/j.cels.2016.02.003>.

AUTHOR CONTRIBUTIONS

M.K.Y., J.D., M.K., R. Sharan, and T.I. designed the study and developed the conceptual ideas. M.K. constructed NeXO. M.K.Y. implemented all other computational methods and analysis. M.K.Y. and T.I. wrote the manuscript with input from the other authors. R. Srivas and K.L. performed the DNA repair and nuclear lumen interaction screens.

ACKNOWLEDGMENTS

We gratefully acknowledge helpful discussion and comments from Hannes Braberg, Anne-Ruxandra Carvunis, Manolis Kellis, Benjamin Kellman, Jianzhu Ma, Jenhan Tao, Alex Thomas, members of the Ideker laboratory, and the anonymous referees. This work was funded by the National Institute of General Medical Sciences (P41-GM103504, P50-GM085764) and the National Institute of Environmental Health Sciences (R01-ES014811). M.Y. received first-year support from the University of California, San Diego, Graduate Training Program in Bioinformatics (T32-GM008806). R. Sharan was supported by a research grant from the Israel Science Foundation (grant 241/11). M.K. was supported by the National Human Genome Research Institute (F30-HG007618) and the University of California, San Diego, Medical Scientist Training Program (T32-GM007198). R. Srivas is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRG-2187-14). J.D. is a co-founder, CEO, and shareholder of Data4Cure, Inc.

Received: September 18, 2015

Revised: December 17, 2015

Accepted: February 1, 2016

Published: February 24, 2016

REFERENCES

- Apgar, J.F., Witmer, D.K., White, F.M., and Tidor, B. (2010). Sloppy models, parameter uncertainty, and the role of experimental design. *Mol. Biosyst.* 6, 1890–1900.
- Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J.A., and Blom, J.G. (2009). Systems biology: parameter estimation for biochemical models. *FEBS J.* 276, 886–902.
- Balakrishnan, R., Harris, M.A., Huntley, R., Van Auken, K., and Cherry, J.M. (2013). A guide to best practices for Gene Ontology (GO) manual annotation. *Database (Oxford)* 2013, bat054.
- Bandyopadhyay, S., Kelley, R., Krogan, N.J., and Ideker, T. (2008). Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput. Biol.* 4, e1000065.
- Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.Y., Ou, J., San Luis, B.J., Bandyopadhyay, S., et al. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods* 7, 1017–1024.
- Bellay, J., Atluri, G., Sing, T.L., Toufighi, K., Costanzo, M., Ribeiro, P.S.M., Pandey, G., Baller, J., VanderSluis, B., Michaut, M., et al. (2011). Putting genetic interactions in context through a global modular decomposition. *Genome Res.* 21, 1375–1387.
- Boucher, B., and Jenna, S. (2013). Genetic interaction networks: better understand to better predict. *Front. Genet.* 4, 290.
- Brachman, R.J., and Levesque, H.J. (2004). Knowledge Representation and Reasoning (San Francisco: Morgan Kaufmann).
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.
- Bruggeman, F.J., and Westerhoff, H.V. (2007). The nature of systems biology. *Trends Microbiol.* 15, 45–50.

- Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. *Cell* 158, 903–915.
- Califano, A., Butte, A.J., Friend, S., Ideker, T., and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* 44, 841–847.
- Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A., and Tagkopoulos, I. (2014). An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *Escherichia coli*. *Mol. Syst. Biol.* 10, 735.
- Carvunis, A.-R., and Ideker, T. (2014). Siri of the cell: what biology could learn from the iPhone. *Cell* 157, 534–538.
- Chen, W.W., Niepel, M., and Sorger, P.K. (2010). Classic and contemporary approaches to modeling biochemical reactions. *Genes Dev.* 24, 1861–1875.
- Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., et al. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40, D700–D705.
- Collins, S.R., Roguev, A., and Krogan, N.J. (2010). Quantitative genetic interaction mapping using the E-MAP approach. *Methods Enzymol.* 470, 205–231.
- Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. *Science* 327, 425–431.
- Deisboeck, T.S., Wang, Z., Macklin, P., and Cristini, V. (2011). Multiscale cancer modeling. *Annu. Rev. Biomed. Eng.* 13, 127–155.
- Deutscher, D., Meilijson, I., Kupiec, M., and Ruppin, E. (2006). Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat. Genet.* 38, 993–998.
- Dowell, R.D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D.A., Rolfe, P.A., Heisler, L.E., Chin, B., et al. (2010). Genotype to phenotype: a complex problem. *Science* 328, 469.
- Dutkowski, J., Kramer, M., Surma, M.A., Balakrishnan, R., Cherry, J.M., Krogan, N.J., and Ideker, T. (2013). A gene ontology inferred from molecular networks. *Nat. Biotechnol.* 31, 38–45.
- Eissing, T., Kuepfer, L., Becker, C., Block, M., Coboecken, K., Gaub, T., Goerlitz, L., Jaeger, J., Loosen, R., Ludewig, B., et al. (2011). A computational systems biology software platform for multiscale modeling and simulation: integrating whole-body physiology, disease biology, and molecular reaction networks. *Front. Physiol.* 2, 4.
- Formosa, T., Ruone, S., Adams, M.D., Olsen, A.E., Eriksson, P., Yu, Y., Rhoades, A.R., Kaufman, P.D., and Stillman, D.J. (2002). Defects in SPT16 or POB3 (yFACT) in *Saccharomyces cerevisiae* cause dependence on the Hir/Hpc pathway: polymerase passage may degrade chromatin structure. *Genetics* 162, 1557–1571.
- Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056.
- Gillis, J., and Pavlidis, P. (2012). “Guilt by association” is the exception rather than the rule in gene networks. *PLoS Comput. Biol.* 8, e1002444.
- Glorigorijević, V., Janjić, V., and Pržulj, N. (2014). Integration of molecular network data reconstructs Gene Ontology. *Bioinformatics* 30, i594–i600.
- Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576.
- Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., and Sethna, J.P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* 3, 1871–1878.
- Haber, J.E., Braberg, H., Wu, Q., Alexander, R., Haase, J., Ryan, C., Lipkin-Moore, Z., Franks-Skiba, K.E., Johnson, T., Shales, M., et al. (2013). Systematic triple-mutant analysis uncovers functional connectivity between pathways involved in chromosome regulation. *Cell Rep.* 3, 2168–2178.
- Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674.
- Hillenmeyer, M.E., Fung, E., Wildenhain, J., Pierce, S.E., Hoon, S., Lee, W., Proctor, M., St Onge, R.P., Tyers, M., Koller, D., et al. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* 320, 362–365.
- Hofree, M., Shen, J.P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115.
- Huntley, R.P., Sawford, T., Martin, M.J., and O’Donovan, C. (2014). Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *Gigascience* 3, 4.
- Ideker, T., and Krogan, N.J. (2012). Differential network biology. *Mol. Syst. Biol.* 8, 565.
- Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jell, Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401.
- Kelley, R., and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* 23, 561–566.
- Kim, Y.-A., and Przytycka, T.M. (2012). Bridging the Gap between Genotype and Phenotype via Network Approaches. *Front. Genet.* 3, 227.
- Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S., Shim, J.E., and Lee, I. (2014). YeastNet v3: a public database of data-specific and integrated functional gene networks for *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 42, D731–D736.
- Kramer, M., Dutkowski, J., Yu, M., Bafna, V., and Ideker, T. (2014). Inferring gene ontologies from pairwise similarity data. *Bioinformatics* 30, i34–i42.
- Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A.G., and Marcotte, E.M. (2010). Predicting genetic modifier loci using functional gene networks. *Genome Res.* 20, 1143–1153.
- Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 21, 1109–1121.
- Lee, A.Y., Onge, R.P.S., Proctor, M.J., Wallace, I.M., Nile, A.H., Spagnuolo, P.A., Jitkova, Y., Gronda, M., Wu, Y., Kim, M.K., et al. (2014). Mapping the cellular response to small molecules using chemogenomic fitness signatures. *Science* 344, 208–211.
- Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nat. Rev. Genet.* 14, 168–178.
- Leiserson, M.D.M., Tatar, D., Cowen, L.J., and Hescott, B.J. (2011). Inferring mechanisms of compensation from E-MAP and SGA data using local search algorithms for max cut. *J. Comput. Biol.* 18, 1399–1409.
- Leiserson, M.D.M., Vandin, F., Wu, H.-T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114.
- Lerman, J.A., Hyduke, D.R., Latif, H., Portnoy, V.A., Lewis, N.E., Orth, J.D., Schrimpe-Rutledge, A.C., Smith, R.D., Adkins, J.N., Zengler, K., and Palsson, B.O. (2012). In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* 3, 929.
- Ma, X., Tarone, A.M., and Li, W. (2008). Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS One* 3, e1922.
- Machado, D., Costa, R.S., Rocha, M., Ferreira, E.C., Tidor, B., and Rocha, I. (2011). Modeling formalisms in Systems Biology. *AMB Express* 1, 45.
- Mackay, T.F.C. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* 15, 22–33.
- Ng, S., Collisson, E.A., Sokolov, A., Goldstein, T., Gonzalez-Perez, A., Lopez-Bigas, N., Benz, C., Haussler, D., and Stuart, J.M. (2012). PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* 28, i640–i646.
- O’Brien, E.J., Lerman, J.A., Chang, R.L., Hyduke, D.R., and Palsson, B.O. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* 9, 693.
- Orth, J.D., Thiele, I., and Palsson, B.O. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248.

- Pamblanco, M., Oliete-Calvo, P., Garcia-Oliver, E., Luz Valero, M., Sanchez del Pino, M.M., and Rodríguez-Navarro, S. (2014). Unveiling novel interactions of histone chaperone Asf1 linked to TREX-2 factors Sus1 and Thp1. *Nucleus* 5, 247–259.
- Pandey, G., Zhang, B., Chang, A.N., Myers, C.L., Zhu, J., Kumar, V., and Schadt, E.E. (2010). An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput. Biol.* 6, 6.
- Park, Y., and Marcotte, E.M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods* 9, 1134–1136.
- Pe'er, D., and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell* 144, 864–873.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Pesquita, C., Faria, D., Falcão, A.O., Lord, P., and Couto, F.M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* 5, e1000443.
- Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S., et al. (2015). NDEx, the Network Data Exchange. *Cell Syst.* 1, 302–305.
- Qi, Y., Suhail, Y., Lin, Y.Y., Boeke, J.D., and Bader, J.S. (2008). Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res.* 18, 1991–2004.
- Ramanan, V.K., Shen, L., Moore, J.H., and Saykin, A.J. (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.* 28, 323–332.
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI '95: Proceedings of the 14th International Joint Conference on Artificial Intelligence* (San Francisco: Morgan Kaufmann), pp. 448–453.
- Schwabish, M.A., and Struhl, K. (2006). Asf1 mediates histone eviction and deposition during elongation by RNA polymerase II. *Mol. Cell* 22, 415–422.
- Segrè, D., Deluna, A., Church, G.M., and Kishony, R. (2005). Modular epistasis in yeast metabolism. *Nat. Genet.* 37, 77–83.
- Skafidas, E., Testa, R., Zantomio, D., Chana, G., Everall, I.P., and Pantelis, C. (2014). Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Mol. Psychiatry* 19, 504–510.
- Sullivan, P.F. (2012). Puzzling over schizophrenia: schizophrenia as a pathway disease. *Nat. Med.* 18, 210–211.
- Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., et al. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.* 43, 656–662.
- Szczurek, E., Gat-Viks, I., Tiuryn, J., and Vingron, M. (2009). Elucidating regulatory mechanisms downstream of a signaling pathway using informative experiments. *Mol. Syst. Biol.* 5, 287.
- Takahashi, K., Ishikawa, N., Sadamoto, Y., Sasamoto, H., Ohta, S., Shiozawa, A., Miyoshi, F., Naito, Y., Nakayama, Y., and Tomita, M. (2003). E-Cell 2: multi-platform E-Cell simulation system. *Bioinformatics* 19, 1727–1729.
- Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T.S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., and Hutchison, C.A., 3rd (1999). E-CELL: software environment for whole-cell simulation. *Bioinformatics* 15, 72–84.
- Tong, A.H.Y., and Boone, C. (2006). Synthetic genetic array analysis in *Saccharomyces cerevisiae*. *Methods Mol. Biol.* 313, 171–192.
- Ulitsky, I., Shlomi, T., Kupiec, M., and Shamir, R. (2008). From E-MAPS to module maps: dissecting quantitative genetic interactions using physical interactions. *Mol. Syst. Biol.* 4, 209.
- Waddington, C.H. (1942). Canalization of development and the inheritance of acquired characters. *Nature* 150, 563–565.
- Walpole, J., Papin, J.A., and Peirce, S.M. (2013). Multiscale computational models of complex biological systems. *Annu. Rev. Biomed. Eng.* 15, 137–154.
- Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11, 843–854.
- Willsey, A.J., Sanders, S.J., Li, M., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A., et al. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155, 997–1007.
- Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U S A* 109, 1193–1198.

Supplemental Information

Translation of Genotype to Phenotype

by a Hierarchy of Cell Subsystems

Michael Ku Yu, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason F. Kreisberg, Cherie T. Ng, Nevan Krogan, Roded Sharan, and Trey Ideker

Supplemental Figures and Legends

Figure S1. Prediction of pairwise genetic interactions under a stringent cross-validation setup, related to Figure 3. Gene pairs were partitioned such that genes represented in the training set are absent in the corresponding test set (**Supplemental Experimental Procedures**). Correlation was calculated across gene pairs that meet an interaction significance criterion of $p < 0.05$ in Costanzo et al.

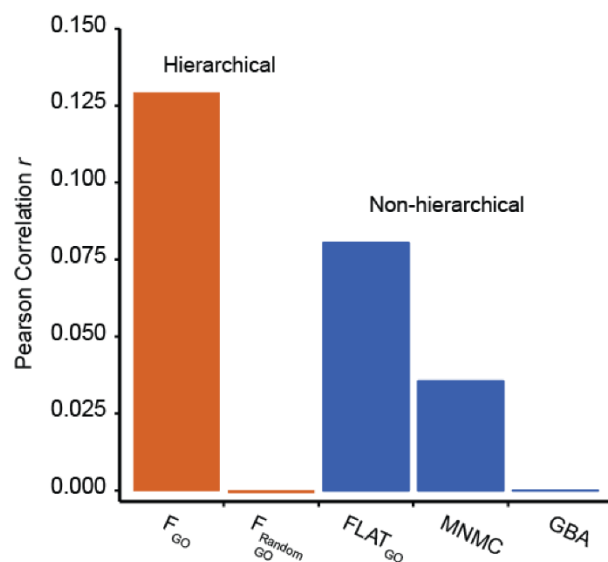


Figure S2. Fuller characterization of genes in an ontology increases prediction accuracy, related to Figure 3. The extent of characterization of a gene in a gene ontology was measured by its Annotation Information Content (AnnIC, **Supplemental Experimental Procedures**). **(A)** Distribution of AnnIC in GO and NeXO across the set of genes annotated to both. **(B)** Distribution of differences between AnnICs computed in NeXO versus GO. **(C,E)** Prediction performance across gene pairs where the GO AnnIC of at least one of the genes is at most a maximum level **(C)** or where the GO AnnICs of both genes are at least a minimum level **(E)**. **(D,F)** The number of gene pairs evaluated in **(C)** and **(E)**. **(G)** Scatterplot of GO terms, comparing the mean AnnIC of genes within the term (x-axis) to the ability to predict within-term genetic interactions (y-axis). Prediction performance is assessed by the correlation between measured and predicted genetic interaction scores across gene pairs that met the interaction significance criterion of $p < 0.05$ in Costanzo et al. Black line indicates the relationship $y = 0.0086x + 0.000091$, fit by a linear regression ($r = 0.39$) where each term is weighted by the number of evaluated gene pairs in the term (area of circles in scatterplot). Only terms with at least 25 and at most 10^5 evaluated gene pairs are shown. The terms “DNA repair” and “nuclear lumen”, explored further in the paper, are indicated.

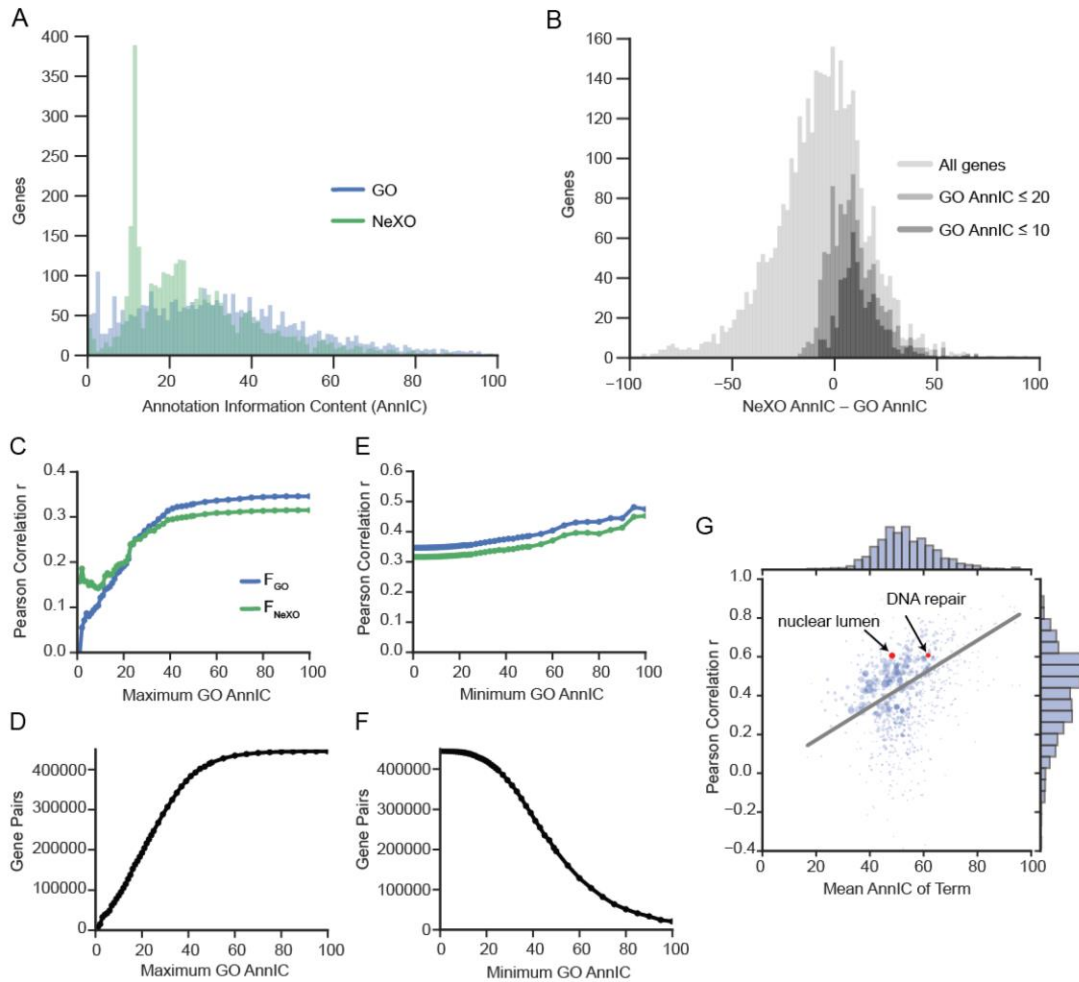


Figure S3: Segregation of positive and negative interactions across F_{GO} , related to Figure 4. For every term enriched for within-term genetic interactions, we calculate the fraction of positive versus negative interactions that continue to fall within one of its children. The distribution of this fraction is shown for all such terms, separately for positive (blue) and negative (red) interactions. Positive interactions falling within a term tend to remain within-term at lower hierarchical levels, whereas negative interactions are more likely to split their genes between two children.

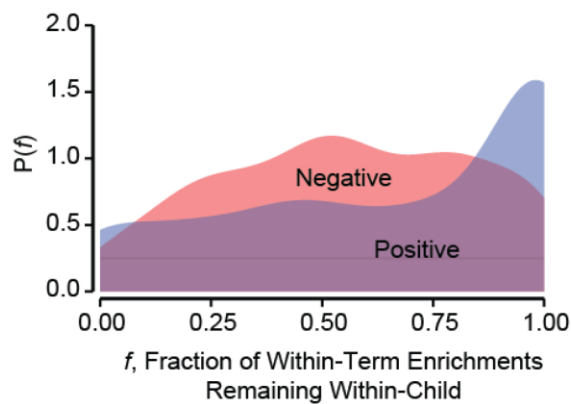


Figure S5: Prediction performance versus the depth of decision trees, related to Experimental Procedures. Decision trees in F_{GO} and F_{NeXO} were pruned to progressively shorter depths and evaluated across one of the cross-validation folds used in **Figure 3C**.

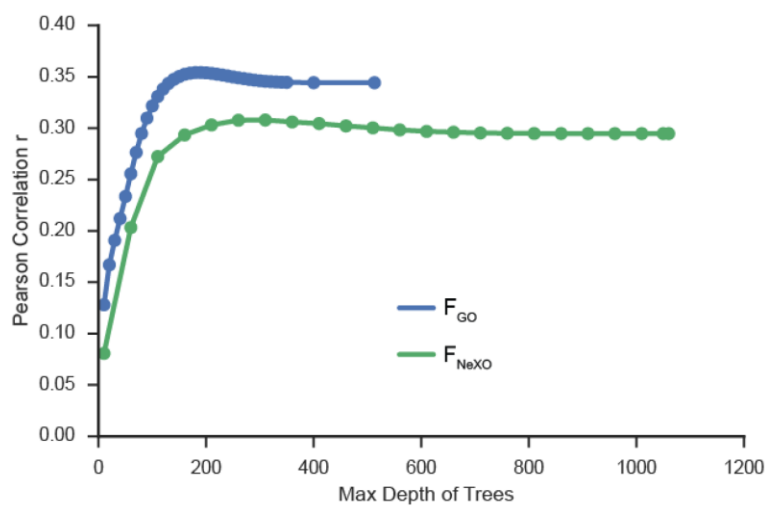
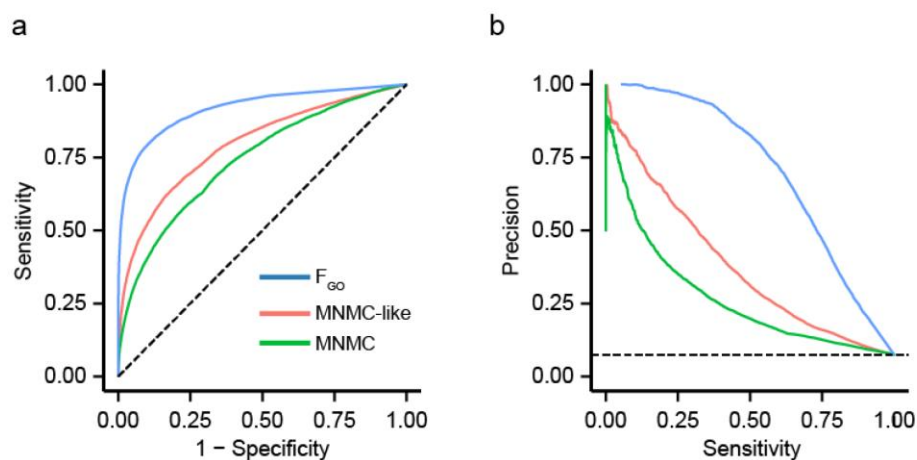


Figure S6: Prediction of synthetic lethal interactions curated in the *Saccharomyces* Genome Database, related to Experimental Procedures. 9,963 synthetic lethal and 124,312 non-synthetic lethal relations, as used in the original MNMC study (Pandey et al., 2010), were taken from the May 2008 version of the *Saccharomyces* Genome Database (SGD). Predictions were made in 10-fold cross validation using F_{GO} , the original MNMC, and a MNMC-like method we constructed based on updates to the original features with alternative classification techniques. **(A)** Receiver-operating and **(B)** Precision-recall curves. Performance of all approaches was far better on this 2008 SGD dataset than on the 2010 Costanzo et al. dataset explored in our main study (**Figure 3C,F**).



Supplemental Tables

Table S1: All 71 new functional relationships found in F_{GO} , related to Table 1.

Table S2: Prediction performance and gene characterization within GO terms, related to Figure 5.

Table S3: New genetic interaction scores for double mutants in DNA repair and the nuclear lumen, related to Figure 5.

Table S4: Gene Ontology structure and gene annotations, related to Experimental Procedures.

Table S5: NeXO structure and gene annotations, related to Experimental Procedures.

Supplemental Files

File S1: Genetic interaction scores predicted by F_{GO} for all pairs of non-essential genes, related to Figure 4.

File S2: Visualization of F_{GO} in Cytoscape format, related to Figure 4 and 5.

Supplemental Experimental Procedures

Previous genetic interaction data

Experimental genetic interaction scores for >6 million double mutants in yeast, measured using synthetic genetic arrays (Costanzo et al., 2010) (SGA, 1,711 queries $\times$ 3,885 arrays), were downloaded from <http://drygin.ccb.utoronto.ca/~costanzo2009/>. We used only measurements where each gene is nonessential (any gene not listed as essential in the *Saccharomyces* Genome Deletion Project, http://www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt) and is mutated as a deletion allele, excluding temperature-sensitive and DAmP alleles. Reciprocal tests for the same gene pair, i.e., query A $\times$ array B versus query B $\times$ array A for two different genes A and B, were merged as previously described (Costanzo et al., 2010). Analysis of these data is based on the principle that the vast majority of strains follow a multiplicative null model in which $f_{ab} = f_a f_b$, where f_{ab} is the measured growth of the double knockout and f_a and f_b are the corresponding single knockout measurements. The genetic interaction score is represented by the deviation $\epsilon_{ab} = f_{ab} - f_a f_b$ from the null model and assigned a p -value of significance p_{ab} using experimental replicates. Following the guidelines of the original data publication (Costanzo et al., 2010), we thresholded ϵ_{ab} and p_{ab} to categorize gene pairs as “negative” or “synthetic sick” interactions ($\epsilon_{ab} < -0.08$ at $p_{ab} < 0.05$), “positive” or “epistatic” interactions ($\epsilon_{ab} > 0.08$ at $p_{ab} < 0.05$), or no significant interaction. For further analysis, we selected measurements for 3,331,172 pairs (69,639 negative and 37,494 positive across 3,686 unique genes) where both genes are annotated to a non-root term in GO, or 3,463,034 pairs (70,353 negative and 37,924 positive across 3,586 unique genes) where both genes are annotated in NeXO.

New genetic interaction mapping of DNA repair and the nuclear lumen

Double gene deletion mutants were generated on solid agar media using SGA technology as previously described (Collins et al., 2010; Tong and Boone, 2006). Briefly, double mutants were allowed to grow for 48 hours at 30°C and imaged using a Canon CCD camera. Colony sizes were quantified using the HT Colony Grid Analyzer (v. 1.1.7), after which genetic interaction scores (S-scores, **Supplemental Table S3**) were computed using the E-MAP toolbox (v. 2.0) (Collins et al., 2006). 55 genes in DNA repair and 81 genes in the nuclear lumen, of which 29 genes are in both terms, were knocked out among the 2,503 measured double mutants. For 315 double mutants, both genes were annotated to both terms. Negative interactions were called with an S-score threshold of -3.70 for gene pairs in DNA repair or -2.47 for gene pairs in the nuclear lumen. These thresholds were determined such that, for gene pairs tested both by these new screens and previously by Costanzo et al., the fraction of negative interactions based on the S-scores was the same as the fraction based on previous measurements (9.0% for DNA repair and 7.5% for the nuclear lumen). Interactions were deposited in the BioGRID database (Stark et al., 2006) and can also be accessed on NDEX (<http://goo.gl/1yyxJT>, UUID: 1515bf44-c398-11e5-8fbc-06603eb7f303, Pratt et al., 2015).

We combined all S-scores with the previous ϵ -scores by Costanzo et al. and retrained F_{GO} (**Supplemental File S2**). In this more complete dataset, we multiplied the S-scores that are < 0 (or > 0) by a factor of 0.02726 (or 0.04251) so that they would be on the same scale as the previous ϵ -scores. In particular, for the 1,345 gene pairs with both types of scores, the fraction of pairs with a rescaled S-score of < -0.08 (or > 0.08) was the same as the fraction of negative (or positive) interactions based on the previous ϵ -scores. For these gene pairs, each type of score was listed as a separate data point for training F_{GO} .

Preparation of the Gene Ontology

The GO structure (OBO) and gene-to-term annotation (GOA) files were downloaded on December 19, 2011 (www.geneontology.org). Annotations with the evidence code “inferred by genetic interaction” (IGI) were removed to avoid potential circularity in predicting genetic interactions. The remaining annotations were propagated up the GO hierarchy in a transitive manner, i.e., if a gene is annotated to a term, then we also annotate the gene to all ancestral terms. We removed terms that were not annotated with any yeast genes or were redundant with respect to their children terms to construct a GO relevant to yeast, following a previously described procedure (Dutkowski et al., 2013, <http://mhk7.github.io/alignOntology/>). In total, 5,124 terms were drawn from all three branches of the Gene Ontology—Cellular Component (707), Biological Process (2,598), and Molecular Function (1,819). Of these, we retained 4,341 terms that were annotated to at least one gene in the genetic interaction data (Costanzo et al., 2010). The three branches were joined under an artificial root term to create a single unified ontology.

It is possible that some annotations were derived from genetic interaction information but were not labeled with the IGI code. Such mislabeling is a curation oversight, and identifying the problematic annotations would likely require a large-scale curation effort similar to that invested into constructing GO. However, while such mislabeled annotations might enable F_{GO} to make circular predictions, they would also enable circularity in $FLAT_{GO}$, in which

the pairwise gene similarities would also reflect the hidden genetic interaction information. The substantially worse performance of FLAT_{GO} versus F_{GO} (**Figure 3C,F**) suggests that the degree of circularity is minimal.

Construction of NeXO: a data-driven ontology assembled from flat gene networks

Our aim was to systematically assemble NeXO from many diverse genome-sized datasets while, at the same time, guiding this assembly based on prior knowledge of a hierarchy of cell systems defined in GO. As input datasets, we used the YeastNet v3 networks (Kim et al., 2014), spanning 68 experimental studies across 8 data types excluding genetic interactions. We integrated these networks together by using them as features in a supervised learning of GO semantic similarity (Resnik, 1995) (random forests regression). Predictions were made “out of bag”, i.e., the similarity of a gene pair was predicted based on information learned from other gene pairs. In effect, the random forest learns patterns in the networks which recapitulate information in GO. Hence, only relations in GO that can be systematically explained from data are included, any relations not justified by the data are excluded, and new relations not in GO are added when the network data support them. Next, we applied the method of Clique Extracted Ontology (CliXO) (Kramer et al., 2014, http://mhk7.github.io/clixo_0.3/), with parameters $\alpha=0.04$ and $\beta=0.5$, on this network. CliXO recursively identifies densely-related groups of genes and hierarchically nests them to form NeXO. We identified the significantly aligned terms between NeXO and GO using a previously described ontology alignment procedure (Dutkowski et al., 2013, <http://mhk7.github.io/alignOntology/>).

Randomization of ontologies

We constructed randomized ontologies by shuffling the names of genes, while preserving gene-to-term annotations and parent-child term relations. In particular, let $\{T_G\}$ refer to the set of terms annotated with gene G . If the names of genes A and B are swapped, then gene B becomes annotated to $\{T_A\}$ and gene A becomes annotated to $\{T_B\}$. This randomization procedure preserves global degree properties of the ontology, i.e., the number of genes, gene pairs, or any k -tuple of genes that span a single term, a pair of terms, or any k -tuple of terms. A total of 10^5 , 10^3 , and 3 randomized ontologies were used in **Figure 1A**, **Figure 1C,D**, and **Figure 3C**, respectively. For these figures, the original mapping of genotypes to phenotypes was used.

Random forests regression

Random forests (Breiman, 2001) from the Python scikit-learn package (Pedregosa et al., 2011) (development version 0.15) were used to regress genetic interaction scores ϵ_{ab} , as described in the **Results**. Due to the very large size of the ontology feature matrix, e.g., 3,331,172 genotypes (rows) by 4,341 terms (columns) for GO, training a random forest on the ontology required an immense amount of computation. To address this challenge, we optimized the scikit-learn source code for both memory consumption and runtime performance, as follows. First, the feature matrix was represented with 8-bit integers instead of the default, 32-bit floating points, such that it consumed ~14GB instead of 58GB of memory. Second, the matrix was stored in Fortran-style order, i.e., columns instead of rows were stored contiguously in memory. Since individual columns are accessed one at a time during training, this ordering improved memory cache hits and resulted in approximately a 1.5-fold decrease in runtime. Third, redundant arithmetic that was being recomputed to find optimal decision splits was eliminated, resulting in a nearly 5-fold decrease in runtime. Together, these optimizations allowed a decision tree in the random forest to be trained in approximately 2 instead of 12 CPU hours. The modified code is available on GitHub (<https://github.com/michaelkyu/scikit-learn-fasterRF>). Following similar thresholds for the measured interaction scores, we categorized gene pairs as being predicted to have a ‘negative’ interaction (predicted $\epsilon_{ab} < -0.08$), a ‘positive’ interaction (predicted $\epsilon_{ab} > 0.08$), or no interaction.

Each random forest consisted of 300 trees, where every tree was learned over a bootstrap sample of gene pairs. At every decision split in a tree, one-third of all unused features were considered for the optimal split, defined by the minimum squared error. Trees were grown to maximal depth, following the default behavior of popular random forests libraries (Liaw and Wiener, 2002; Pedregosa et al., 2011) and the original random forests study (Breiman, 2001). We tested whether pruning these trees to a shorter depth would improve performance. While trees at approximately 29% (GO) or 37% (NeXO) of the maximal depth did improve performance slightly (<0.02 gain in correlation, **Supplemental Figure S5**), we chose to keep the trees at maximal depth because it is unclear how significant this gain is and whether it would be reproducible in different random partitions of the data for cross validation or in different genotype-phenotype datasets.

We used F_{GO} to predict genetic interactions across all pairwise deletions of 5,003 genes that are nonessential and annotated to at least one non-root term in any of the GO branches (**Supplemental File S1**). Performance of these predictions was assessed against the measurements selected from Costanzo et al. using the technique of four-fold cross-validation, resulting in 17,807 negative and 3,246 positive predicted interactions. For the remaining pairs, predictions from the four forests across the cross-validation folds were averaged, resulting in an additional 17,888

negative and 2,666 positive predicted interactions. This averaging was also used to predict scores for the retested mutants using the F_{GO} with complete training in **Figure 5B** and for the triple knockouts involving genes in the HIRA complex.

Comparison of methods for predicting genetic interactions

We updated the MNMC method because the original implementation (Pandey et al., 2010) was trained on a set of literature-curated synthetic lethal interactions that was much smaller in size than the set of genetic interactions considered in our study, and because the set of features used by the method to score each gene pair had been updated since the 2010 publication. To train MNMC, we calculated five base features that, according to the original study, are among the most informative for predicting synthetic lethality of a gene pair. Three of these features corresponded to the semantic similarities of the gene pair in each branch of GO. Following the procedure in the original study, Tao semantic similarities (Tao et al., 2007) were calculated using an unprocessed GO structure and annotations (downloaded on Aug 5, 2012) instead of the processed GO for constructing ontotypes. A fourth feature was the binary co-membership of a gene pair in a protein-protein interaction (PPI) community. Twenty-six communities were identified using the network of curated protein-protein interactions in the BioGRID database (Stark et al., 2006) v3.1.91 and running the 'Graph.community_multilevel(return_levels=False)' method from the Python igraph library v0.6 (Nepusz, 2006). A fifth feature was the number of KEGG pathways (Kanehisa et al., 2014) shared by each gene pair. Annotations of genes to *S. cerevisiae* pathways in KEGG were retrieved on March 16, 2015 using the KEGG REST API (kegg.jp/kegg/rest). Lastly, these five base features were combined in pairs to create $(5 \text{ choose } 2) = 10$ additional 'overlay' features as previously described in the original MNMC method. We trained a random forest using all 15 base and overlay features to classify the synthetic lethals evaluated in the original MNMC study. This updated MNMC outperformed the original MNMC (**Supplemental Figure S6**); this performance difference may have been due to calculating more recent versions of the base features. For **Figure 3C,F**, we trained a separate MNMC on the genetic interactions measurements by Costanzo et al.

We adapted a previous GBA method for predicting genetic interactions (Lee et al., 2010) and applied it on the data similarity network used to construct NeXO. Whereas the original GBA method considered binary genetic interactions and the network neighbors of only one gene in a gene pair, we consider continuous-valued genetic interactions and the neighbors of both genes. Let $S(a, b)$ be the edge weight between genes a and b in the data similarity network. The GBA score between a and b is defined as

$$GBA(a, b) = \sqrt{GBA(a)GBA(b)}$$

$$GBA(x) = \sum_{y \in N(x)} S(x, y) * \epsilon_{xy}$$

where $N(x)$ is the set of neighbors of gene x and ϵ_{xy} is the measured genetic interaction. Conceptually, $GBA(x)$ is the contribution of x 's neighborhood, and $GBA(a, b)$ is the geometric mean of the contributions of both a 's and b 's neighborhoods. We found that taking the geometric mean more accurately predicts genetic interactions than the arithmetic mean does. Using the data similarity network (Pearson's $r = 0.13$; **Figure 3C**) resulted in approximately similar performance as using the YeastNet v3 integrated network ($r = 0.14$).

The same random forest framework as the functionalized ontologies and MNMC was applied to train FLAT_{NeXO}, using the 68 input networks to NeXO as features, and FLAT_{GO}, using the Resnik semantic similarity (Resnik, 1995) between gene pairs in GO as the sole feature. In **Figure 3C,E**, F_{GO} , $F_{Random\ GO}$, FLAT_{GO}, and MNMC were trained over the same cross-validation partitioning of 3,331,172 gene pairs. Similarly, F_{NeXO} , $F_{Random\ NeXO}$, FLAT_{NeXO}, and GBA were trained over the same cross-validation of 3,463,034 pairs.

To generate the strict cross-validation setup in **Supplemental Figure S1**, the set of 3,686 genes in the processed GO were partitioned into 8 approximately equal-sized subsets $G_1, G_2, \dots, G_8$. For every pair of subsets G_i and G_j , the set of all gene pairs where neither gene is in $G_i \cup G_j$ forms a training set, and the set of all gene pairs where both genes are in $G_i \cup G_j$ forms the corresponding test set. This partitioning scheme lets every gene pair be predicted in at least one test set. Gene pairs where both genes are in the same subset G_i are represented in seven test sets, and we take the average prediction across these sets.

FBA predictions were taken from Szappanos et al., 2011. Since the underlying metabolic model (Mo et al., 2009) has been iteratively refined by manual curation to be consistent with genetic and physical data, we were unable to evaluate these predictions in a cross-validation scenario.

Characterization of genes in a gene ontology

To score how well a gene has been characterized in a gene ontology, we introduce the concept of a gene's Annotation Information Content (AnnIC). Formally, the AnnIC of a gene g is defined as

$$AnnIC(g) = \sum_{t \in T_g} -\log_2 \frac{|t|}{N}$$

where T_g is the set of terms that contain g but are not ancestors of any other term that also contains g , $|t|$ is the size of term t , and N is the number of genes in the ontology. Conceptually, $AnnIC(g)$ is high if g is well characterized by being annotated to many small terms, and it is low if g is poorly characterized by being annotated to only a few large terms.

Term enrichments

A term is a least common ancestor (LCA) of a pair of genes if both genes are assigned to it, and if it is not the ancestor of any other term shared by those genes. Note that a pair of genes may have multiple LCAs, since terms may have multiple parents. Let G define a set of gene pairs beneath a term or term pair, as follows:

$$G(T) = \{ \{g_1, g_2\} \mid \text{Term } T \text{ is an LCA of } g_1 \text{ and } g_2 \}$$

$$G(T_1, T_2) = \{ \{g_1, g_2\} \mid \exists P, \text{ a common parent of } T_1 \text{ and } T_2 \text{ that is also an LCA of } g_1 \text{ and } g_2 \}.$$

A term T was identified as a within-term enrichment if the fraction of gene pairs in $G(T)$ that are interacting is at least twice the background rate of interactions, and the number of gene pairs in $G(T)$ that are interacting passes a hypergeometric test with $p < 0.05$ using a Bonferroni correction for family-wise error rate. A pair of terms T_1 and T_2 was identified as a between-term enrichment according to a similar test on $G(T_1, T_2)$. We did not test for within-term enrichment when $G(T) = \emptyset$, or for between-term enrichment when $G(T_1, T_2) = \emptyset$, including the situation where T_1 and T_2 do not have a common parent. These situations were not counted in the total number of tests computed for Bonferroni correction. Note that it is possible for a parent term to have within-term enrichment even if none of its children have between-term enrichments, and vice versa. Enrichments were identified separately for negative and positive interactions. In **Figure 1C**, we counted the number of unique terms and unique term pairs that formed within-term and between-term enrichments for both types of interactions. Together, these enrichments involved 1,050 unique terms (~24% of the 4,341 GO terms that contained at least one gene among the measured genetic interactions).

In addition to between-term enrichments that share a common parent, we also tested enrichment among pairs of terms that were distantly related in GO, i.e., the genes contained by one of the term's parents did not overlap with those of the other term. We only considered pairs of terms that contained at least 12 gene pairs and for which each term contained at least 3 genes. Enrichment was based on the number and ratio of interactions according to a similar criteria for $G(T)$ and $G(T_1, T_2)$. Most of the resulting enrichments were partially redundant such that terms in different enrichments shared some genes. To remove this redundancy, we first sorted the enrichments in descending order according to the fraction of gene pairs that had an interaction, breaking ties by sorting according to the number of gene pairs. Using this sorted ordering, we greedily selected enrichments so that the genes contained by the terms in an enrichment were not already covered by a previously selected enrichment. This selection procedure resulted in a total of 71 enrichments (53 negative and 18 positive) (**Table 1**, **Supplemental Table S1**).

Visualization of the functionalized ontology

F_{GO} was visualized using Cytoscape version 3.2 (Saito et al., 2012). Only a spanning tree of the GO structure is shown in order to remove multi-parent relationships that complicate visualization. To form the spanning tree, the edges from each term to all but one of its parents were removed. The tree was laid out using the yFiles Circular layout tool, and edges representing between-term enrichments with a common parent were bundled together using the automatic edge bundling tool. The visualizations of the structure and logic of F_{GO} shown in **Figure 4A** (<http://goo.gl/ViTztH>, UUID: 2f0cd18a-c3a4-11e5-8fbc-06603eb7f303), **Figure 4C** (<http://goo.gl/iWmyNk>, UUID: 420e4e54-c3a2-11e5-8fbc-06603eb7f303), **Figure 5A** (<http://goo.gl/mPkMVs>, UUID: 555cd256-c3a3-11e5-8fbc-06603eb7f303), and **Supplemental Figure S4** (<http://goo.gl/oQWh0A>, UUID: 575ee3e8-c3a3-11e5-8fbc-06603eb7f303) can be accessed online on the Network Data Exchange (NDEx, Pratt et al., 2015) and as a Cytoscape file in **Supplemental File S2**.

Supplemental References

Collins, S.R., Schuldiner, M., Krogan, N.J., and Weissman, J.S. (2006). A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol.* 7, R63.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–205.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2, 18–22.

Mo, M.L., Palsson, B.O., and Herrgård, M.J. (2009). Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* 3, 37.

Nepusz, G.C. and T. (2006). The igraph software package for complex network research. *InterJournal Complex Sy*, 1695.

Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., Pico, A.R., Bader, G.D., and Ideker, T. (2012). A travel guide to Cytoscape plugins. *Nat. Methods* 9, 1069–1076.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–9.

Tao, Y., Sam, L., Li, J., Friedman, C., and Lussier, Y. a (2007). Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics* 23, i529–38.

ACKNOWLEDGEMENTS

Chapter 5, in full, is a reprint of the material as it appears in: Yu MK, Kramer M, Dutkowski J, Srivas R, Licon K, Kreisberg J, Ng CT, Krogan N, Sharan R, Ideker T. “Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems.” *Cell Systems*. 2016;2(2):77-88. The dissertation author was the second author of this work.