UNIVERSITY OF CALIFORNIA, SAN DIEGO

Hierarchical Models of Biological Systems

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

in

Bioinformatics and Systems Biology

by

Michael Ku Yu

Committee in charge:

Professor Trey Ideker, Chair
Professor Christopher Glass, Co-Chair
Professor Vineet Bafna
Professor Bing Ren
Professor Sheng Zhong

2017

The Dissertation of Michael Ku Yu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2017

TABLE OF CONTENTS

SIGNATURE PAGE..... iii

TABLE OF CONTENTS iv

LIST OF FIGURES..... vi

LIST OF SUPPLEMENTAL FILES viii

ACKNOWLEDGMENTS..... ix

VITA xii

ABSTRACT OF THE DISSERTATION..... xiv

INTRODUCTION..... 1

References 5

CHAPTER 1: A SWISS-ARMY KNIFE FOR HIERARCHICAL MODELING OF BIOLOGICAL SYSTEMS 8

1.1 Abstract 8

1.2 Background 9

1.3 Results..... 12

1.4 Discussion 19

1.5 Conclusions..... 21

1.6 Availability of data and materials 21

1.7 Figures..... 22

1.8 Methods..... 28

1.9 Acknowledgments 30

1.10 References 31

CHAPTER 2: TRANSLATION OF GENOTYPE TO PHENOTYPE BY A HIERARCHY OF CELL SUBSYSTEMS..... 35

2.1 Summary 35
2.2 Introduction 36
2.3 Results 40
2.4 Discussion 50
2.5 Experimental Procedures 55
2.6 Figure and Tables 57
2.7 Supplemental Experimental Procedures 68
2.8 Supplemental Figures 80
2.9 Author Contributions 87
2.10 Acknowledgements 88
2.11 References 89
Chapter 3: USING DEEP LEARNING TO MODEL THE HIERARCHICAL STRUCTURE AND FUNCTION OF A CELL 98
3.1 Abstract 98
3.2 Introduction 99
3.3 Results 102
3.4 Discussion 113
3.5 Methods 117
3.6 Figures 124
3.7 Supplemental Figures 131
3.8 Acknowledgments 134
3.9 References 135

LIST OF FIGURES

Figure 1.1. Software architecture of the Data-Driven Ontology Toolkit (DDOT). 22

Figure 1.2. Discovery of known and candidate genes and subsystems involved in Fanconi Anemia. 24

Figure 1.3. The Hierarchical Viewer (HiView) web application. 25

Figure 1.4. Visual transformations of a DAG into a tree. 26

Figure 1.5. A compendium of gene ontologies for 649 diseases. 27

Figure 2.1: Patterns of genetic interaction reflect the hierarchical structure of the Gene Ontology. 57

Figure 2.2: The ontotype method of translating genotype to phenotype. 59

Figure 2.3: Genome-wide prediction of pairwise genetic interactions in yeast. 60

Figure 2.4: The Functionalized Gene Ontology. 62

Figure 2.5: Elucidating the genetic logic of DNA repair and the nuclear lumen. 64

Figure 2.6: Prediction of triple mutants. 66

Figure S2.1. Prediction of pairwise genetic interactions under a stringent cross-validation setup, related to Figure 2.3. 80

Figure S2.2. Fuller characterization of genes in an ontology increases prediction accuracy, related to Figure 2.3. 81

Figure S2.3: Segregation of positive and negative interactions across F_{GO}, related to Figure 2.4. 83

Figure S2.5: Prediction performance versus the depth of decision trees, related to Experimental Procedures. 85

Figure S2.6: Prediction of synthetic lethal interactions curated in the Saccharomyces Genome Database, related to Experimental Procedures. 86

Figure 3.1. Modeling system structure and function with visible learning. 124

Figure 3.2. Prediction of cell viability and genetic interaction phenotypes. 125

Figure 3.3. Interpretation of genotype-phenotype associations. 127

Figure 3.4. Top subsystem states for translation of genotype to growth. 128

Figure 3.5. Analysis of subsystem functional logic. 129

Figure 3.6. Analysis of a new DNA repair subsystem..... 130

Figure S3.1. Precision-recall curves for classification of negative genetic interactions. 131

Figure S3.2. CliXO top subsystem states for translation of genotype to growth. 132

Figure S3.3. Calculating Relative Local Improvement in Predictive Power (RLIPP). .. 133

LIST OF SUPPLEMENTAL FILES

Table S2.1: All 71 new functional relationships found in F_{GO}, related to Table 2.1.

Table S2.2: Prediction performance and gene characterization within GO terms, related to Figure 2.5.

Table S2.3: New genetic interaction scores for double mutants in DNA repair and the nuclear lumen, related to Figure 2.5.

Table S2.4: Gene Ontology structure and gene annotations, related to Experimental Procedures.

Table S2.5: NeXO structure and gene annotations, related to Experimental Procedures.

File S2.1: Genetic interaction scores predicted by F_{GO} for all pairs of non-essential genes, related to Figure 2.4.

File S2.2: Visualization of F_{GO} in Cytoscape format, related to Figure 2.4 and 2.5.

Table S3.1. RLIPP scores for subsystems in the Gene Ontology and CLIXO.

Table S3.2. Boolean logic approximating the states of subsystems in the Gene Ontology and CLIXO.

ACKNOWLEDGMENTS

I would like to thank my adviser Trey Ideker for his extensive mentorship in scientific research and communication. The ideas and writing in this dissertation are the product of countless debates and revisions with Trey.

I would like to thank my dissertation committee for their support and advice. Christopher Glass and Bing Ren provided deep guidance in my exploration of the 3D physical structure of chromatin.

I would like to thank my colleagues, Jianzhu Ma, Michael Kramer, Janusz Dutkowski, Samson Fong, Fan Zheng, and Keiichiro Ono, for their close collaborations and deep insights across several projects. I would like to thank all past and present members of Trey Ideker's and Hannah Carter's labs, with special thanks to Gordon Bean, Daniel Carlin, Anne-Ruxandra Carvunis, Jing Chen, Barry Demchak, Aaron Gary, Andrew Gross, Joris van de Haar, Matan Hofree, Justin Huang, Philipp Jaeger, Jason Kreisberg, Katherine Licon, Dexter Pratt, Rintaro Saito, Rohith Srivas, Tina Wang, and Wei Zhang. The lab assistants, Charlotte Curtis, Anne Maria Viola, Shunya Wade, Stephanie Mirkin, and Cita Trueblood, were critical in helping me finish administrative tasks. I would like to thank like my collaborator and mentor, Roded Sharan, for his guidance.

I would like to thank my family for their never-ending love and support. My parents, Robin H. Yu, M.D., and Estrelita O. Ku, M.D., have always taught me their wisdom, imbued me with their steadfastness, and nurtured me with their love. My older brothers, Robinson Yu, M.D., and Christopher Yu, have always looked after me and encouraged the best out of me.

I would like to thank my girlfriend Kelly Chinh for her overflowing love and encouragement as I treaded through the thick and thin of attaining a PhD. Our adventures and time spent together in San Diego made me finally feel at home in this city.

I would like to thank my close friends from San Diego and UCSD, including Paul Abels & Jing Zhang, Kunal Bhutani, Chris DeBoever, Kuan Fu Ding, James Jensen, Boyko Kakaradov, Eric Levy & Avni Nijhawan, Justin Mih, Nathan Mih, Shamim Mollah, Gabriel Pratt, Anugraha Raman, Siddarth Selvaraj, and Jenhan Tao; from MIT, including Albert Chang, Ana Chen, David Chen, Corinna Hui, Irene Kaplow, Jeremy Lai, Grace Li, Fan Liu, Victoria Lo, Sharon Tam, Jason Trigg, Albert Wang, Jeff Xing, and Di Ye; and from high school, including Vikram Agarwal and Kevin Wilkes. Their friendships have inspired me to never give up on my dreams.

I would like to thank my previous mentors, Jonathan “Steve” Alexander, Merilyn Jennings, Cigdem & Gokhan Yilmaz, Tony Eng, Michael Baym, Bonnie Berger, Frans Schalekamp & Anke van Zuylen, and Sebastian Will. They were integral in cultivating my scientific skillsets before this PhD.

The Introduction is, in part, based on material currently being prepared for submission as “A Swiss-Army Knife for Hierarchical Modeling of Biological Systems”. Michael Ku Yu, Jianzhu Ma, Keiichiro Ono, Fan Zheng, Samson Fong, Aaron Gary, Jing Chen, Barry Demchak, Dexter Pratt, Trey Ideker. The Introduction is also, in part, based on material as it appears as "Translation of genotype to phenotype by a hierarchy of cell subsystems" in *Cell Systems*, 2017. Michael Ku Yu, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason F. Kreisberg, Cherie T. Ng, Nevan Krogan, Roded

Sharan, and Trey Ideker. The Introduction is also, in part, material as it may appear in *Nature Methods*, 2017. “Using deep learning to model the hierarchical structure and function of a cell”. Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, Trey Ideker. The dissertation author was the primary investigator and author of these papers.

Chapter 1, in full, is currently being prepared for submission of the material as “A Swiss-Army Knife for Hierarchical Modeling of Biological Systems”. Michael Ku Yu, Jianzhu Ma, Keiichiro Ono, Fan Zheng, Samson Fong, Aaron Gary, Jing Chen, Barry Demchak, Dexter Pratt, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reformatted reprint of the material as it appears as "Translation of genotype to phenotype by a hierarchy of cell subsystems" in *Cell Systems*, 2017. Michael Ku Yu, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason F. Kreisberg, Cherie T. Ng, Nevan Krogan, Roded Sharan, and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in *Nature Methods*, 2017. “Using deep learning to model the hierarchical structure and function of a cell”. Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

VITA

EDUCATION

2010 Bachelor of Science, Massachusetts Institute of Technology
Mathematics
Electrical Engineering and Computer Science

2011 Master of Engineering, Massachusetts Institute of Technology
Electrical Engineering and Computer Science

2017 Doctor of Philosophy, University of California, San Diego
Bioinformatics and Systems Biology

PUBLICATIONS

Michael Ku Yu, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason F. Kreisberg, Cherie T. Ng, Nevan Krogan, Roded Sharan, and Trey Ideker. "Translation of genotype to phenotype by a hierarchy of cell subsystems." Cell Systems 2, no. 2 (2016): 77-88.

Michael Ku Yu, Jianzhu Ma, Keiichiro Ono, Fan Zheng, Samson Fong, Aaron Gary, Jing Chen, Barry Demchak, Dexter Pratt, Trey Ideker. “A Swiss-army Knife for Hierarchical Modeling of Biological Systems.” (in preparation)

Jianzhu Ma*****, **Michael Ku Yu***, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, Trey Ideker. “Using deep learning to model the hierarchical structure and function of a cell.” (in review)

Sheng Wang, Jianzhu Ma, **Michael Ku Yu**, Fan Zheng, Edward W Huang, Jiawei Han, Jian Peng, Trey Ideker. “Annotating gene sets by mining large literature collections with protein networks.” Biocomputing 2018 [Internet]. 2017. Available from: http://dx.doi.org/10.1142/9789813235533_0055

Michael H. Kramer, Jean-Claude Farre, Koyel Mitra, **Michael Ku Yu**, Keiichiro Ono, Barry Demchak, Katherine Licon, Mitchell Flagg, Rama Balakrishnan, J. Michael Cherry, Suresh Subramani, Trey Ideker. "Active Interaction Mapping Reveals the Hierarchical Organization of Autophagy." Molecular Cell 65, no. 4 (2017): 761-774.

Michael Kramer, Janusz Dutkowski, **Michael Yu**, Vineet Bafna, Trey Ideker. “Inferring gene ontologies from pairwise similarity data.” Bioinformatics 30, i34–42 (2014).

Janusz Dutkowski, Keiichiro Ono, Michael Kramer, **Michael Yu**, Dexter Pratt, Barry Demchak, Trey Ideker. “NeXO Web: the NeXO ontology database and visualization platform.” Nucleic Acids Res. 42, D1269–74 (2014).

Justin K. Huang, Daniel E. Carlin, **Michael Ku Yu**, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo, Trey Ideker. “Systematic evaluation of gene networks for discovery of disease genes.” (in review)

Wei Zhang, Ana Bojorquez, Daniel Ortiz Velez, John Paul Shen, Guorong Xu, Kevin Chen, Katherine Licon, Collin Melton, Katrina Olson, **Michael Ku Yu**, Justin K. Huang, Hannah Carter, Emma Farley, Michael Snyder, Stephanie Fraley, Jason F. Kreisberg, Trey Ideker. “A global transcriptional network connecting noncoding mutations to changes in tumor gene expression.” (in review)

Tina Wang, Brian Tsui, Jason F. Kreisberg, Neil A. Robertson, Andrew M. Gross, **Michael Ku Yu**, Hannah Carter, Holly M. Brown-Borg, Peter D. Adams, and Trey Ideker. “Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment.” Genome Biology. 2017; 18: 57. doi: 10.1186/s13059-017-1186-2

Joris van de Haar, **Yu Michael Ku**, Trey Ideker. “Mutation load explains the vast majority of mutually exclusive interactions in cancer.” (in preparation)

Sebastian Will*****, **Michael Yu***, Bonnie Berger. “Structure-based whole-genome realignment reveals many novel noncoding RNAs.” Genome Research. 1018–1027 (2013).

Sebastian Will*****, **Michael Yu***, Bonnie Berger. “Structure-based whole genome realignment reveals many novel noncoding RNAs.” In 16th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2012).

Frans Schalekamp*****, **Michael Yu***, Anke van Zuylen*****. “Clustering with or without the Approximation.” Journal of Combinatorial Optimization, 21, 1-37. doi: 10.1007/s10878-011-9382-6. (2011)

*** co-first authors**

ABSTRACT OF THE DISSERTATION

Hierarchical Models of Biological Systems

by

Michael Ku Yu

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2017

Professor Trey Ideker, Chair
Professor Christopher Glass, Co-Chair

For biological systems, structure and hence function are hierarchically organized at multiple scales. For example, genetic variation in nucleotides (1nm) gives rise to functional changes in proteins (1–10nm), which in turn affect protein complexes, cellular processes, organelles (10nm–1μm), and, ultimately, phenotypes observed in cells (1–

10 μm), tissues (100μm–100mm), and individuals (>1m). Here, I exploit this principle for biological modeling.

First, I develop a software library that facilitates the assembly, analysis, and visualization of biological hierarchies, represented by a data structure called an ontology. As demonstration, I assemble a compendium of hierarchies describing the molecular mechanisms of 649 diseases, by integrating a set of gene-disease associations with a gene similarity network derived from ‘omics data. For example, the hierarchy for Fanconi Anemia recaptures the disease’s known relation with DNA repair and proposes new relations with orthogonal pathways.

Next, I introduce a strategy for genotype-to-phenotype translation by using existing knowledge of a hierarchy of cellular subsystems. Guided by this structure, I 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. It also accurately predicts the growth outcomes of two new screens of 2,503 double gene knockouts impacting DNA repair or nuclear lumen and generalizes to larger knockout combinations.

Finally, I present a more accurate and interpretable model for translation called DeepCell, a “visible” neural network that couples the model’s inner workings to those of the cell. Outperforming the ontotype approach, DeepCell simulates cellular growth nearly as accurately as laboratory observations. During simulation, genotypes induce patterns on the activities of cellular subsystems, enabling in-silico investigations of the molecular

mechanisms underlying each genotype-phenotype association. These mechanisms can be validated and many are unexpected; some are governed by Boolean logic. Cumulatively, 80% of the importance for growth prediction is captured by 484 subsystems (21%), reflecting the emergence of a complex phenotype.

INTRODUCTION

Hierarchy as an organizing principle of complex systems. Understanding a complex system requires dissecting its structure – the set of parts and connections in the system, and its function – how parts interact to give rise to system behavior. For biological systems, structure and hence function are hierarchically organized at multiple scales. For example, genetic variation in nucleotides (1nm) gives rise to functional changes in proteins (1–10nm), which in turn affect protein complexes, cellular processes, organelles (10nm–1μm), and, ultimately, phenotypes observed in cells (1–10 μm), tissues (100μm–100mm), and individuals (>1m). Although this principle has long existed as a guiding intuition to study biological systems, its full exploitation in modeling remains an open challenge. In this dissertation, I present a software library and new methods for modeling the hierarchical structure and function of biological systems.

Ontologies as representations of hierarchical organization. In many fields, knowledge across multiple scales is modeled by ontologies— a factorization of 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 of over a dozen layers (The Gene Ontology Consortium, 2014). Other types of

biomedical ontologies have been created to hierarchically organize phenotypes (Human Phenotype Ontology (Robinson et al. 2008)), diseases (Disease Ontology (Kibbe et al. 2015)), chemical compounds (ChEBI Ontology (Degtyarenko et al. 2008)), and cell types (Cell Ontology (Bard, Rhee, and Ashburner 2005)).

A major limitation of existing biomedical ontologies is that they are drawn by manual curation of literature. Given that the number of human curators stays relatively constant, the curation process does not easily scale with the exponential rise in published literature, and curators cannot be experts in every biological domain, introducing bias or incomplete interpretation of literature. Moreover, a complex biological pattern may not have been exposed in a study's manuscript but is nonetheless discoverable in the raw data of that study or in an integration of data across multiple studies. Because of these challenges, it is difficult to scale existing ontologies and to create new ontologies for different biological contexts. For example, there is only one Gene Ontology even though cell structure depends on many factors, including cell and tissue type (Greene et al. 2015; Gosselin et al. 2014), growth condition (Ideker and Krogan 2012), disease (Creixell et al. 2015), and species (Gerstein et al. 2014; Sharan and Ideker 2006; Carvunis et al. 2015).

Data-driven ontologies. The proliferation of 'omics datasets has begun to reveal the organizational complexity of a biological system in an unbiased manner, whether through generation of proteomic data (protein-protein interactions and co-localization) (Huttlin et al. 2017; Chong et al. 2015), transcriptomic data (RNA co-expression across conditions and time points) (Sefer, Kleyman, and Bar-Joseph 2016; Saha et al. 2017), or genetic data (epistasis and synthetic lethality) (Costanzo et al. 2016). Many tools are

available to group genes with similar ‘omics profiles into clusters (Mitra et al. 2013) and to analyze these gene clusters for functional enrichment (Huang, Sherman, and Lempicki 2009). Although such techniques begin to capture the modular design of biological systems, they do not explicitly reveal the underlying physical and functional hierarchy giving rise to the data.

In the interest of building models that are both hierarchical and data-driven, I have been involved in the development of methods to hierarchically organizing genes into cellular subsystems based on their similarities in ‘omics data (Dutkowski et al. 2014, 2013; Kramer et al. 2014). In particular, the CliXO algorithm (Kramer et al. 2014) searches for cliques (sets of genes in which all gene pairs have high similarity) or dense subnetworks (sets of genes in which many gene pairs do) above a specified similarity (scale) threshold. By progressively loosening this threshold, CLIXO identifies progressively larger cliques, which subsume the smaller cliques found at earlier thresholds. Each clique defines a cellular subsystem, and all cliques are arranged to form a “data-driven hierarchy” (or ontology) of subsystems. Unlike other hierarchical clustering algorithms, which produce binary trees (a.k.a. dendrograms), CLIXO produces a hierarchy with the flexibility to capture the true structure of a cell, recognizing, for instance, that a subsystem may factor into many subcomponents (not just two) and participate in several higher-order processes. A data-driven hierarchy can also be compared with literature-curated ontologies through a procedure known as ontology alignment. We have shown that these hierarchies not only recapitulate subsystems in the Gene Ontology (GO), including 60% of known cellular components in yeast, but also discover new subsystems.

A swiss army knife for hierarchical modeling. Bioinformatics research often involves a complex maneuvering between heterogeneous formats, model construction, and model interpretation. To facilitate hierarchical modeling of biological systems with data-driven ontologies, I report in Chapter 1 the development of a user-friendly and integrated software library, the Data-driven Ontology Toolkit (DDOT). DDOT consists of a Python package, which handles the assembly and analysis of ontologies, and a web application, which handles visualization.

Exploiting biological structure to model biological function. 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 Chapter 2, I 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 upwards through the hierarchy: in this way, genetic variants comprising genotype are converted to effects on the cell subsystems impacted by those variants. I call the set of all such effects 'ontotype,' representing variation at intermediate scales between nanoscopic changes in genes and macroscopic changes in phenotype.

In Chapter 3, I will improve upon this modeling approach by developing an interpretable or 'visible' neural network (VNN). In this design, the functional state of each subsystem in a biological hierarchy is represented by a bank of neurons. Connectivity of these neurons is set to mirror the biological hierarchy, so that they take input only from neurons of child subsystems and send output only to neurons of parent (super)systems,

with weights determined during training. During simulation of the neural network, predictions of a biological system's behavior are made with respect to the real parts and interactions that generated that behavior. In this way, the model is interpretable, a sought-after but often missing property in machine learning.

References

Bard, Jonathan, Seung Y Rhee, and Michael Ashburner. 2005. "An Ontology for Cell Types." *Genome Biol.* 6 (2):R21.

Brachman, Ronald J, and Hector J Levesque. 2004. *Knowledge Representation and Reasoning*. The Morgan Kaufmann Series in Artificial Intelligence. San Francisco: Morgan Kaufmann. https://doi.org/http://dx.doi.org/10.1016/B978-155860932-7/50103-5.

Carvunis, Anne-Ruxandra, and Trey Ideker. 2014. "Siri of the Cell: What Biology Could Learn from the iPhone." *Cell* 157 (3). Elsevier Inc.:534–38. https://doi.org/10.1016/j.cell.2014.03.009.

Carvunis, Anne-Ruxandra, Tina Wang, Dylan Skola, Alice Yu, Jonathan Chen, Jason F Kreisberg, and Trey Ideker. 2015. "Evidence for a Common Evolutionary Rate in Metazoan Transcriptional Networks." *Elife* 4 (December).

Chong, Yolanda T, Judice L Y Koh, Helena Friesen, Supipi Kaluarachchi Duffy, Kaluarachchi Duffy, Michael J Cox, Alan Moses, Jason Moffat, Charles Boone, and Brenda J Andrews. 2015. "Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis." *Cell* 161 (6):1413–24.

Costanzo, Michael, Benjamin VanderSluis, Elizabeth N Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, et al. 2016. "A Global Genetic Interaction Network Maps a Wiring Diagram of Cellular Function." *Science* 353 (6306).

Creixell, Pau, Erwin M Schoof, Craig D Simpson, James Longden, Chad J Miller, Hua Jane Lou, Lara Perryman, et al. 2015. "Kinome-Wide Decoding of Network-Attacking Mutations Rewiring Cancer Signaling." *Cell* 163 (1):202–17.

Degtyarenko, Kirill, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2008. "{ChEBI}: A Database and Ontology for Chemical Entities of Biological Interest." *Nucleic Acids Res.* 36 (Database issue):D344--50.

Dutkowski, Janusz, Michael Kramer, Michal a Surma, Rama Balakrishnan, J Michael

Cherry, Nevan J Krogan, and Trey Ideker. 2013. “A Gene Ontology Inferred from Molecular Networks.” *Nature Biotechnology* 31 (1):38–45. https://doi.org/10.1038/nbt.2463.

Dutkowski, Janusz, Keiichiro Ono, Michael Kramer, Michael Yu, Dexter Pratt, Barry Demchak, and Trey Ideker. 2014. “NeXO Web: The NeXO Ontology Database and Visualization Platform.” *Nucleic Acids Research* 42 (Database issue):D1269-74. https://doi.org/10.1093/nar/gkt1192.

Gerstein, Mark B, Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B Brown, Carrie A Davis, et al. 2014. “Comparative Analysis of the Transcriptome across Distant Species.” *Nature* 512 (7515):445–48.

Gosselin, David, Verena M. Link, Casey E. Romanoski, Gregory J. Fonseca, Dawn Z. Eichenfield, Nathanael J. Spann, Joshua D. Stender, et al. 2014. “Environment Drives Selection and Function of Enhancers Controlling Tissue-Specific Macrophage Identities.” *Cell* 159 (6). Elsevier Inc.:1327–40. https://doi.org/10.1016/j.cell.2014.11.023.

Greene, Casey S, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene a Zelaya, Daniel S Himmelstein, Ran Zhang, et al. 2015. “Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.” *Nature Genetics* 47 (6). Nature Publishing Group:569–76. https://doi.org/10.1038/ng.3259.

Huang, Da Wei, Brad T Sherman, and Richard a Lempicki. 2009. “Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists.” *Nucleic Acids Research* 37 (1):1–13. https://doi.org/10.1093/nar/gkn923.

Huttlin, Edward L, Raphael J Bruckner, Joao A Paulo, Joe R Cannon, Lily Ting, Kurt Baltier, Greg Colby, et al. 2017. “Architecture of the Human Interactome Defines Protein Communities and Disease Networks.” *Nature* 545 (7655):505–9.

Ideker, Trey, and Nevan J Krogan. 2012. “Differential Network Biology.” *Molecular Systems Biology* 8 (565). Nature Publishing Group:565. https://doi.org/10.1038/msb.2011.99.

Kibbe, Warren A, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, Christopher J Mungall, et al. 2015. “Disease Ontology 2015 Update: An Expanded and Updated Database of Human Diseases for Linking Biomedical Knowledge through Disease Data.” *Nucleic Acids Res.* 43 (Database issue):D1071--8.

Kramer, Michael, Janusz Dutkowski, Michael Yu, Vineet Bafna, and Trey Ideker. 2014. “Inferring Gene Ontologies from Pairwise Similarity Data.” *Bioinformatics (Oxford, England)* 30 (12):i34-42. https://doi.org/10.1093/bioinformatics/btu282.

Mitra, Koyel, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. 2013. “Integrative Approaches for Finding Modular Structure in Biological Networks.” *Nat.*

Rev. Genet. 14 (10):719–32.

Robinson, Peter N, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. “The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease.” *Am. J. Hum. Genet.* 83 (5):610–15.

Saha, Ashis, Yungil Kim, Ariel D H Gewirtz, Brian Jo, Chuan Gao, Ian C McDowell, GTEx Consortium, Barbara E Engelhardt, and Alexis Battle. 2017. “Co-Expression Networks Reveal the Tissue-Specific Regulation of Transcription and Splicing.” *Genome Res.* 27 (11):1843–58.

Sefer, Emre, Michael Kleyman, and Ziv Bar-Joseph. 2016. “Tradeoffs between Dense and Replicate Sampling Strategies for {High-Throughput} Time Series Experiments.” *Cell Syst* 3 (1):35–42.

Sharan, Roded, and Trey Ideker. 2006. “Modeling Cellular Machinery through Biological Network Comparison.” *Nat. Biotechnol.* 24 (4):427–33.

The Gene Ontology Consortium. 2014. “Gene Ontology Consortium: Going Forward.” *Nucleic Acids Research* 43 (November 2014):1049–56. https://doi.org/10.1093/nar/gku1179.

Waddington, C H. 1942. “Canalization of Development and the Inheritance of Acquired Characters.” *Nature* 150 (3811):563–65.

CHAPTER 1: A SWISS-ARMY KNIFE FOR HIERARCHICAL MODELING OF BIOLOGICAL SYSTEMS

1.1 Abstract

Systems biology requires not only genome-scale data but also methods to integrate these data into interpretable models of biological systems. Previously, we developed approaches that organize ‘omics data into a structural hierarchy of cellular components and pathways, called a “data-driven ontology”. Such hierarchies recapitulate known cellular subsystems and discover new ones. They also enable more accurate and interpretable predictions of phenotype from genotype. To broadly facilitate this type of modeling, we report the development of a software library called the Data-driven Ontology Toolkit (DDOT). DDOT consists of a Python package for assembly and analysis of data-driven hierarchies and a web application (http://hiview.ucsd.edu) for visualizing them. To demonstrate the toolkit, we invoke it programmatically to assemble a compendium of ontologies for 649 diseases, by integrating a set of gene-disease mappings from the Monarch database with a gene similarity network from ‘omics data. For example, the resulting ontology for the disease Fanconi Anemia consists of 194 genes arranged in a hierarchy of 74 subsystems and includes most known and many novel functions. To facilitate data sharing and reproducibility, DDOT provides an easy interface for online storage and retrieval of data-driven hierarchies at the Network Data Exchange (NDEx). Source code, documentation, and example Jupyter notebooks are open-source and available at https://github.com/michaelkyu/ontology.

1.2 Background

For biological systems, structure and function are organized hierarchically across multiple scales. For example, genetic variation in nucleotides (1nm) gives rise to functional changes in proteins (1–10nm), which in turn affect protein complexes, cellular processes, organelles (10nm–1μm), and, ultimately, phenotypes observed in cells (1–10μm), tissues (100μm–100mm), individuals (>1m). The proliferation of ‘omics datasets creates the potential to reveal this organizational complexity in an unbiased manner, whether through generation of proteomic data (protein-protein interactions and co-localization) (Chong et al. 2015; Huttlin et al. 2017), transcriptomic data (RNA co-expression across conditions and time points) (Saha et al. 2017; Sefer, Kleyman, and Bar-Joseph 2016), or genetic data (epistasis and synthetic lethality) (Costanzo et al. 2016).

In the interest of building models that are both hierarchical and data-driven, we have previously developed methods for hierarchically organizing genes into cellular subsystems based on their gene-gene pairwise similarities in ‘omics data (Dutkowski et al. 2013; M. Kramer et al. 2014). In particular, the CliXO algorithm (M. Kramer et al. 2014) searches for cliques (sets of genes in which all gene pairs have high similarity) or dense subnetworks (sets of genes in which many gene pairs do) above a specified similarity (scale) threshold. By progressively loosening this threshold, CLIXO identifies progressively larger cliques, which subsume the smaller cliques found at earlier thresholds. Each clique defines a cellular subsystem, and all cliques are arranged to form a “data-driven hierarchy” (or ontology) of subsystems. Unlike other hierarchical clustering algorithms, which produce binary trees (a.k.a. dendrograms), CLIXO produces a

hierarchy with the flexibility to capture the true structure of a cell, recognizing, for instance, that a subsystem may factor into many subcomponents (not just two) and participate in several higher-order processes. A data-driven hierarchy can also be compared with literature-curated ontologies through a procedure known as ontology alignment. We have shown that these hierarchies not only recapitulate subsystems in the Gene Ontology (GO), including 60% of known cellular components in yeast, but also discover new subsystems (Dutkowski et al. 2013).

The data-driven nature of these hierarchies also enables the *de novo* modeling of diseases and biological processes. For instance, previously we created a hierarchy of autophagy-related processes in *S. cerevisiae* (M. H. Kramer et al. 2017) by invoking a multi-step protocol. First, we compiled a set of genes known to have a core function in autophagy and calculated their similarities to other genes based on several types of ‘omics data, including those generated in experimental models of autophagy. The most similar genes were recognized as additional candidates for being autophagy-related genes. Finally, we applied CliXO to hierarchically organize the total set of core and candidate genes based on their gene-gene similarities. The resulting model, an Autophagy Gene Ontology (AtGO), suggested many mechanistic hypotheses, many of which we experimentally confirmed, including a revised understanding of the substructure of known processes such as selective autophagy, the discovery of new subprocesses such as the transport of Atg19-receptor cargos, and the discovery of new functions of genes such as Gyp1 and Atg26.

To broadly facilitate these approaches in the biomedical research community, we now report the development of a software framework, the Data-driven Ontology Toolkit

(DDOT), which enables the construction and analysis of hierarchical models in a Python package and their visualization in a web application.

1.3 Results

Introduction to DDOT with application to Fanconi Anemia Gene Ontology. DDOT implements four major functions:

- Build Data-Driven Ontology: Given a set of genes and a gene similarity network, hierarchically cluster the genes to infer cellular subsystems using the CLIXO algorithm (M. Kramer et al. 2014). The resulting hierarchy of subsystems defines a data-driven ontology.
- Visualize Hierarchical Structure: Browse the full hierarchical structure of a data-driven ontology, including the network of gene similarities used to infer it, in a web application called the Hierarchical Viewer (HiView, http://hiview.ucsd.edu).
- Align Ontologies: Annotate a data-driven ontology by aligning it to a curated ontology such as the Gene Ontology (GO). For instance, if a data-driven subsystem contains a similar set of genes as the GO term for DNA repair, then annotate this subsystem as being involved in DNA repair. Data-driven subsystems with no such matches represent new molecular mechanisms.
- Expand Gene Set: Given a set of genes as a “seed set” and a gene similarity network, identify an expanded set of genes that are highly similar to the seed set. This function can broaden the scope of a data-driven ontology beyond genes that are already well known.

We illustrate the above steps in an example study of Fanconi Anemia (FA), a rare genetic disorder that is associated with bone marrow failure, myeloid dysplasia, and increased cancer risk (Ceccaldi, Sarangi, and D’Andrea 2016). A total of 20 genes have

been classified as FA genes because their germline mutations in patients have been associated with FA clinical phenotypes ("Fanconi Anemia Mutation Database," n.d.). All of these genes have known functions in the repair of DNA damage due to interstrand cross-links. However, beyond these DNA repair functions, the full spectrum of genes and pathways underlying FA remains unclear. For example, recently Sumpter et al. linked 7 of the 20 FA genes to new functional roles in autophagy pathways, separate from their classical roles in DNA repair (Sumpter et al. 2016). Moreover, 127 other genes have been co-cited with FA in at least one study (**Methods**).

Following our previous procedure for constructing AtGO (M. H. Kramer et al. 2017), we applied DDOT in a five-step pipeline to construct a Fanconi Anemia gene ontology (FanGO) as follows (**Figure 1.1**). First, we gathered input data, consisting of the 20 known FA genes as a seed set of genes for modeling and a gene similarity network derived by integrating several types of molecular evidence including protein-protein interactions, co-expression, co-localization, and epistasis (**Methods**). Second, we scored every gene for its involvement in FA by calculating its average functional similarity to the seed genes. The minimum score among the seed genes was used as a threshold to identify an additional set of 174 candidate genes (**Figure 1.2A**). Third, we organized all genes in a hierarchy of 74 cellular subsystems to construct FanGO. Fourth, we aligned it with GO. Finally, we uploaded FanGO to an online database, the Network Data Exchange (NDEx, http://ndexbio.org) (Pratt et al. 2015), and visualized the results in HiView (**Figure 1.3**).

Since the time of constructing FanGO, one of the candidate genes, RFWD3 (a.k.a. FANCW), was independently confirmed as a FA gene (Knies et al. 2017). Among the other candidate genes, 54 have been co-cited with FA (**Figure 1.2B**). An ontology

alignment between FanGO and GO revealed that 43 of FanGO subsystems (58%) had significant overlap with GO terms (**Figure 1.2C**). Consistent with prior knowledge that FA is marked by sensitivity to DNA damage, many overlapping GO terms are cellular complexes or pathways involved in the recognition of DNA lesions, including the GINS and MutSalpha complexes, or the repair of the DNA helix. Canonical FA subsystems, such as the "FA nuclear complex" and the "FANCM-MHF" complex, were also found (**Figure 1.2D**).

The recovery of these known connections suggests that the other 120 genes and 31 subsystems in FanGO are attractive hypotheses for further study in laboratory models or patients with FA phenotypes. In particular the genes RFC4 and RMI, although not currently known to be involved in FA, have higher average similarity scores to the seed set than observed among the seed genes. Several FanGO subsystems involve cellular functions that are not immediately recognizable as related to DNA damage repair, such as the spliceosome (**Figure 1.5E**), the condensin complex, and telomere maintenance.

A suite of functions organized in a Python package. DDOT consists of a Python package, which handles the construction and analysis of ontologies, and the HiView web application, which handles visualization. Beyond these major functions, described above, the Python package provides many other utility functions to analyze an ontology:

- Examine ontology structure. For each subsystem, retrieve its hierarchical connections (genes, child and descendant subsystems, parent and ancestral subsystems) and the subnetwork of gene similarities that supports the subsystem's existence. For each gene, retrieve its set of subsystems.

- Modify ontology structure: Reduce the size of an ontology by removing a set of subsystems or genes. Randomize connections between genes and subsystems to create new ontologies representing a null model for statistical tests.
- Flatten ontology structure. Instead of inferring an ontology from a gene similarity network, perform the reverse process of inferring a gene similarity network from an ontology. In particular, the similarity between two genes is calculated as the size of the smallest common subsystem, known as the Resnik score (Resnik 1999).
- Map genotypes to the ontology. Given a set of mutations comprising a genotype, propagate the impact of these mutations to the subsystems containing these genes in the ontology. In particular, the impact on a subsystem is estimated by the number of its genes that have been mutated. These subsystem activities, which we have called an “ontotype”, enables more accurate and interpretable predictions of phenotype from genotype (Yu et al. 2016).
- Load curated ontologies. Parse Open Biomedical Ontologies (OBO) and gene-association file (GAF) formats that are typically used to describe curated ontologies like GO.

At the core of the Python package is an “Ontology” class (**Figure 1.1**) through which most analysis can be executed. This object-oriented design enables more intuitive software development, as conceptual manipulations to an ontology’s structure can be reflected by programmatic changes to an Ontology object’s attributes. DDOT’s functions have all been design to work together in concise pipelines that involve minimal boilerplate

code. A tutorial and documentation of every function in the Python package is available at http://the-data-driven-ontology-toolkit-ddot.readthedocs.io.

Ontology visualization with HiView. The HiView web application provides an interactive visualization of the two major features of a data-driven ontology: (1) the hierarchical structure relating genes and subsystems and (2) the data supporting the inference of each subsystem (**Figure 1.3**). Visualizing the hierarchical structure is challenging because it is a directed acyclic graph (DAG), in which each node may have multiple parents and multiple children. As drawing a DAG on a two-dimensional canvas often requires having many edges cross, which induce a “hairball” effect, current tools for visualizing hierarchies are typically limited to simplified planar structures, such as a tree, where each node has at most one parent, or a small subgraph of the DAG based on local context. For instance, the QuickGO browser (Binns et al. 2009) for the Gene Ontology shows the subgraph containing the ancestors of a selected term but excludes other close relations, such as sibling terms. In HiView, a user can choose among three visual transformations of a DAG into a tree (**Figure 1.4**). In one transformation, HiView shows only a subset of edges that form a spanning tree that keeps nodes connected; other edges are hidden, enabling the tree to be drawn more compactly while losing some information (**Figure 1.4B**). In the other two transformations, HiView duplicates nodes in a manner that preserve hierarchical relations but makes the visualization less compact. In particular, nodes are duplicated either to represent a top-down traversal of the DAG (**Figure 1.4C**) or to supplement the spanning tree (**Figure 1.4D**). In addition to these transformations, HiView allows the user to interactively zoom between more expansive views of the entire

ontology and more focused views of particular subsystems. Finally, genes and subsystems can be searched based on their names and metadata.

Whereas a subsystem in a curated ontology is explained by a table of citations and evidence codes, a subsystem in data-driven gene ontology is explained by several types of gene-gene interaction networks, which themselves require special graphical visualization. In HiView, these networks are displayed in a side panel when the user selects a subsystem (**Figure 1.3**). Distinct interaction types, such as protein-protein versus co-expression, are distinguished by edge color, and the interaction strength is represented by edge thickness. To filter the amount of data shown, a user can select edges by their interaction type and strength. A user can also visualize why a subsystem was factorized into children subsystems by highlighting the genes belonging to a particular child. In addition, HiView has been designed to be programmatic, visualizing any ontology that has been pre-formatted with the Python package and hosted online at NDEx. Additional metadata about genes or subsystems can be viewed as node attributes, such as color and size.

Reproducible software pipelines and interfaces with other tools. To facilitate shareable and reproducible software pipelines, DDOT has been designed such that both the input data and output ontologies can be stored and retrieved online at NDEx (**Figure 1.1**) (Pratt et al. 2015). This built-in connection with NDEx provides three advantages. First, it enables a standard data format: a user can execute the simpler task of getting data to and from NDEx. Second, data and results can be shared with the biomedical research community through NDEx-based URLs. Third, it inherits other NDEx features

and functionalities, including provenance tracking, access permissions, user profiles, and interfaces with other web tools. By supporting other data formats, DDOT can readily interface with other desktop applications, such as Cytoscape (Shannon et al. 2003), and other programming libraries in Python, such as the Pandas, NetworkX, igraph, and matplotlib.

Programmatic assembly of data-driven gene ontologies for 649 diseases. To demonstrate the accessibility and ease of computational modeling enabled by DDOT, we repeated the modeling procedure used for FA (**Figures 1,2**) to programmatically construct data-driven ontologies for 649 other diseases. These ontologies were based on two types of input data: a set of known gene associations for each disease, curated in the Monarch Initiative database (Mungall et al. 2017), and the same gene similarity network used to construct FanGO (**Figure 1.5A**). By calling DDOT functions, the pipeline for constructing these 649 ontologies was very concise, consisting of 16 lines of code for loading input data and setting parameters and 8 lines for modeling in a single Python script. For each disease, its ontology suggests an association with new genes (**Figure 1.5B**) as well as cellular subsystems, many of which are not found in GO **(Figure 1.5C-D)**. The ontologies are available on NDEx and can be visualized through HiView.

1.4 Discussion

Bioinformatics analysis often involves complex maneuvering between heterogeneous formats, model construction and interpretation. To facilitate these steps in creating hierarchical models of biological systems, DDOT has been engineered with several key design choices worthy of mention. First, we have implemented an Ontology class as a central data structure through which major functions are executed. Both low-level and high-level functions have been implemented to enable flexible and concise software pipelines. Second, we support several types of formats for importing and exporting Ontology objects, including text files (tabular, CX, and OBO formats), in-memory Python objects (Pandas dataframes, iGraph objects, NetworkX objects), and online files stored on NDEx. Third, model construction by the Python package has been seamlessly tied to model interpretation by HiView, enabling faster prototyping and iteration of ideas. Finally, we have provided in-depth documentation of every function and a tutorial of the Python package to minimize the learning curve for using it and to encourage software extensions by others.

This work enables the exploration of hierarchies of cellular subsystems for numerous diseases and biological contexts. We have taken a first step in this exploration by creating hierarchies for 649 diseases. To gain deeper biological insight, these hierarchies might be further expanded in several ways. First, all disease ontologies were created by using the same gene-gene similarity network; the only difference in input data was the seed set of known genes associated to each disease. However, in future studies this similarity network can incorporate data from experimental models or patients of the disease. This strategy greatly improved our previous hierarchy of autophagy-related

processes, by refining the hierarchical relations among genes and subsystems as well as revealing new cellular subsystems (M. H. Kramer et al. 2017). Second, other algorithms (Lee et al. 2011) for identifying candidate genes from a seed set based on a gene similarity network can be evaluated.

While we have demonstrated DDOT's ability to model gene ontologies, this form of modeling might also organize other biological concepts. For instance, with the appropriate input data, DDOT can be used to infer and study the hierarchical organization of symptoms and diseases within disease classes based on clinical co-occurrence or shared molecular mechanisms (Park, Hescott, and Slonim 2017), chemical compounds within classes based on mode of action, patients within patient subtypes based on molecular or phenotypic similarities, and cell types within cell-lineage pathways based on gene regulatory signatures. Literature-curated analogues already exist in the Human Phenotype Ontology (Robinson et al. 2008), Disease Ontology (Kibbe et al. 2015), ChEBI Ontology (Degtyarenko et al. 2008), and Cell Ontology (Bard, Rhee, and Ashburner 2005).

1.5 Conclusions

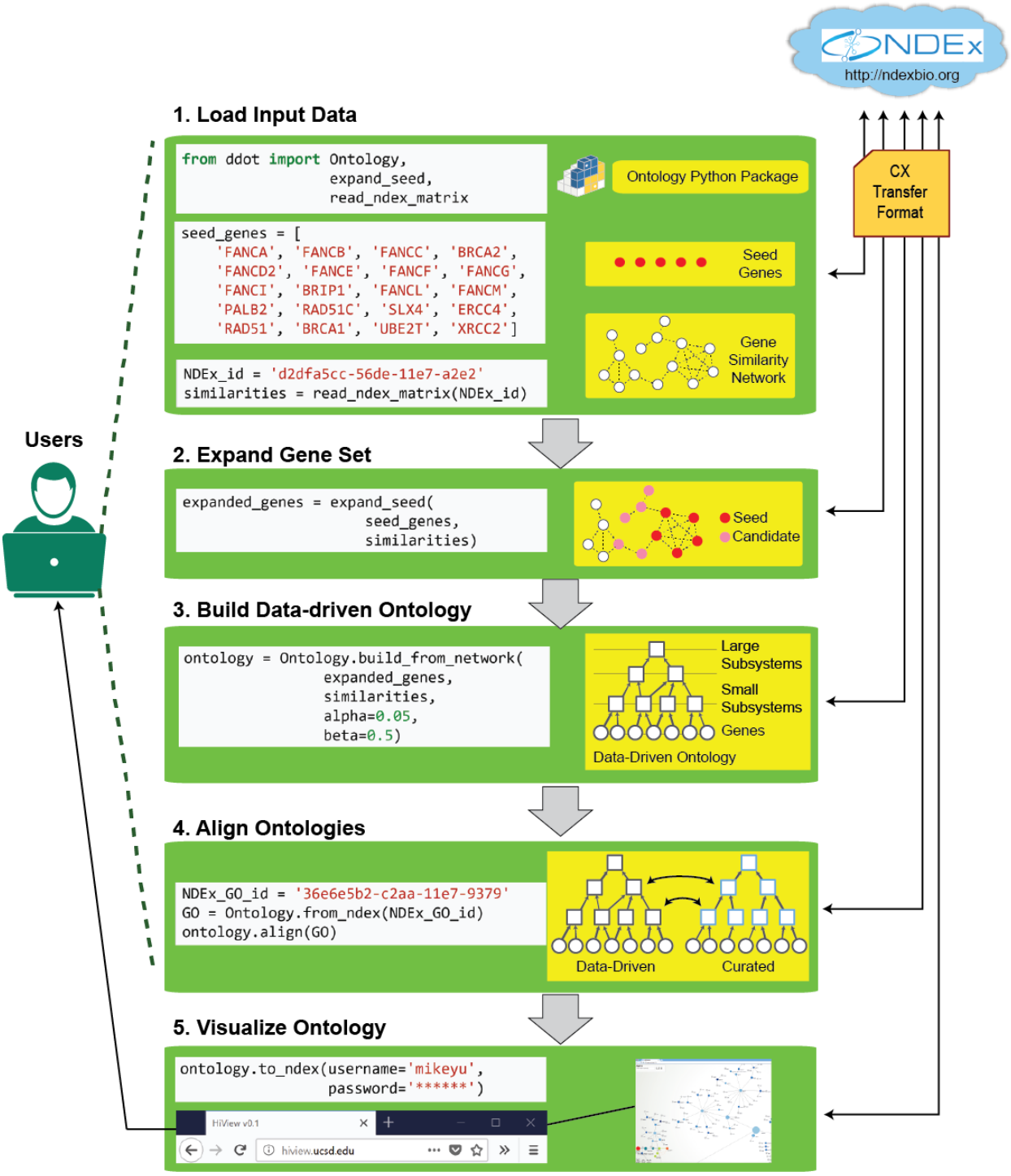
Hierarchical models of the cell and other biological systems have long been curated in the form of biomedical ontologies, but their construction and visualization in a data-driven manner is a more recent endeavor for which no unified software framework yet exists. DDOT aims to fill this software void by providing an implementation of the major required functions in a Python package alongside the HiView web application.

1.6 Availability of data and materials

Source code for the Python package and Jupyter notebooks for reconstructing FanGO and the 649 disease gene ontologies are available at https://github.com/michaelkyu/ontology under a MIT open source license. Documentation and a tutorial for the Python package are available at http://the-data-driven-ontology-toolkit-ddot.readthedocs.io. Source code for HiView is available at https://github.com/idekerlab/hiview under a MIT open source license.

1.7 Figures

Figure 1.1. Software architecture of the Data-Driven Ontology Toolkit (DDOT). DDOT implements several functions that form a workflow to construct, analyze, and visualize data-driven gene ontologies. This workflow is executed in an integrated software framework, consisting of a Python package and a web application called the Hierarchical Viewer. Input and output data can be stored online at the Network Data Exchange (NDEx, http://ndexbio.org), facilitating sharing and reproducibility of results.


NDEx
http://ndexbio.org
1. Load Input Data
from ddot import Ontology,
expand_seed,
read_ndex_matrix
Ontology Python Package
seed_genes = [
'FANCA', 'FANCB', 'FANCC', 'BRCA2',
'FANCD2', 'FANCE', 'FANCF', 'FANCG',
'FANCI', 'BRIP1', 'FANCL', 'FANCM',
'PALB2', 'RAD51C', 'SLX4', 'ERCC4',
'RAD51', 'BRCA1', 'UBE2T', 'XRCC2']
Seed Genes
NDEx_id = 'd2dfa5cc-56de-11e7-a2e2'
similarities = read_ndex_matrix(NDEx_id)
Gene Similarity Network
CX Transfer Format
Users
2. Expand Gene Set
expanded_genes = expand_seed(
seed_genes,
similarities)
Seed
Candidate
3. Build Data-driven Ontology
ontology = Ontology.build_from_network(
expanded_genes,
similarities,
alpha=0.05,
beta=0.5)
Large Subsystems
Small Subsystems
Genes
Data-Driven Ontology
4. Align Ontologies
NDEx_GO_id = '36e6e5b2-c2aa-11e7-9379'
GO = Ontology.from_ndex(NDEx_GO_id)
ontology.align(GO)
Data-Driven
Curated
5. Visualize Ontology
ontology.to_ndex(username='mikeyu',
password='******')
HiView v0.1
hiview.ucsd.edu

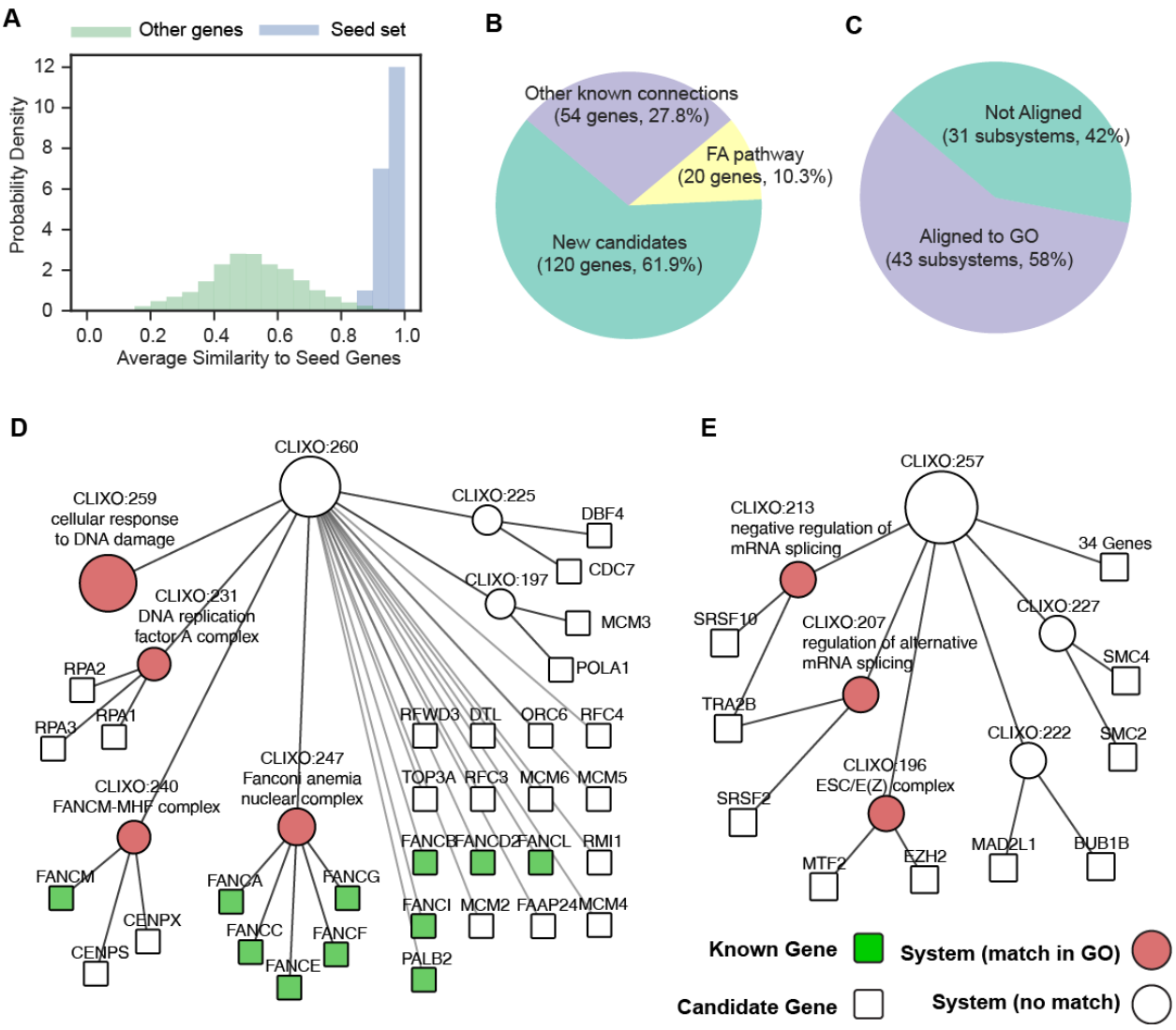


Figure 1.2. Discovery of known and candidate genes and subsystems involved in Fanconi Anemia. A Fanconi Anemia Gene Ontology (FanGO) assembled by combining prior knowledge of a seed set of 20 known FA genes with large-scale 'omics data. **(A)** Every gene was scored for its involvement in FA by calculating its average functional similarity to the seed set. A histogram of these scores is shown for genes in the seed set (blue) and other genes (green). The minimum score among genes in the seed set (0.88) was used as a threshold to identify a candidate set of 174 genes. **(C)** Decomposition of the combined set of 194 genes. **(D)** Decomposition of the 74 subsystems in FanGO. **(D-E)** Focused view on FanGO subsystems related to the Fanconi anemia nuclear complex and FANCM-MHF complex **(D)** and mRNA splicing **(E)**.

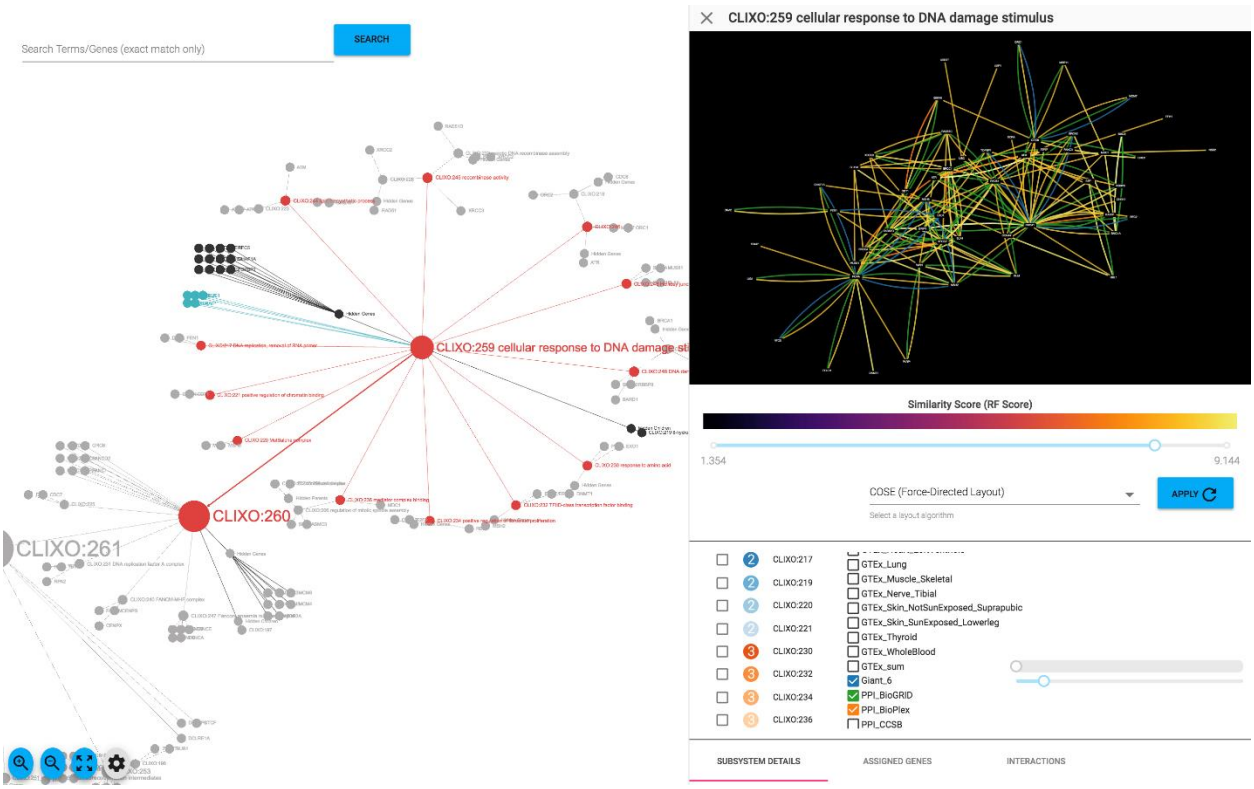


Figure 1.3. The Hierarchical Viewer (HiView) web application. An interactive visualization of a data-driven gene ontology of Fanconi Anemia. HiView visualizes both the hierarchical structure the ontology (left) as well as the ‘omics data (right), in the form of gene interaction networks, that were used to infer a subsystem and its hierarchical relations. Genes and subsystems can be searched by name or metadata (top-left).

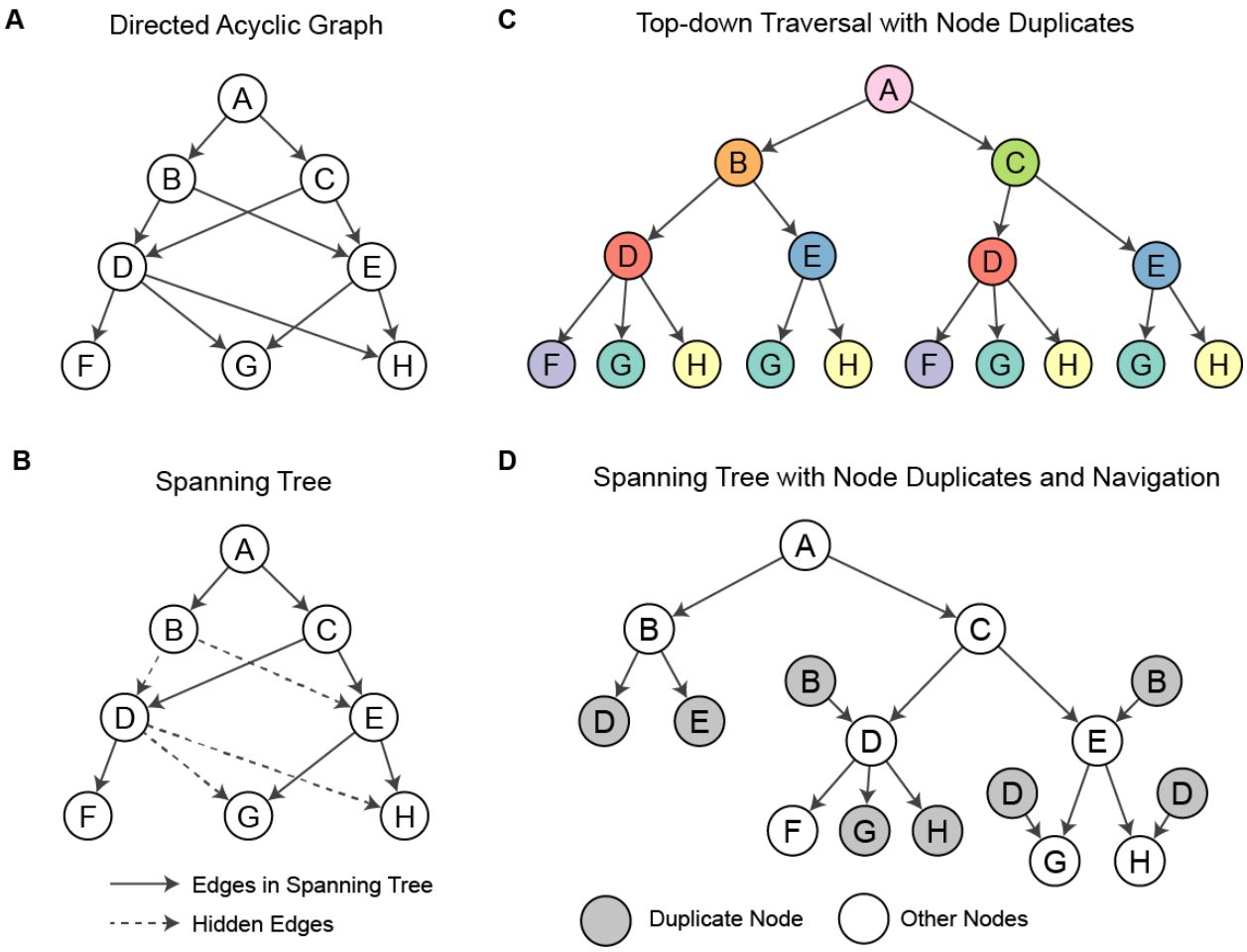


Figure 1.4. Visual transformations of a DAG into a tree. DDOT implements three different transformations to display a directed acyclic graph (DAG) as a tree. **(A)** A toy example of a DAG that cannot be visualized in two dimensions without having edges cross. **(B)** Only edges in a spanning tree of the DAG are shown while other edges are hidden. **(C)** Nodes are duplicated as they are visited in a top-down traversal from the root(s) of the DAG to the leaves. Duplicates of the same node are shown by color. **(D)** Information that is lost by hiding edges from the spanning tree in (B) can be recovered by duplicating parents and children. A duplicate node can be selected to navigate to the location of other duplicates of the same node.

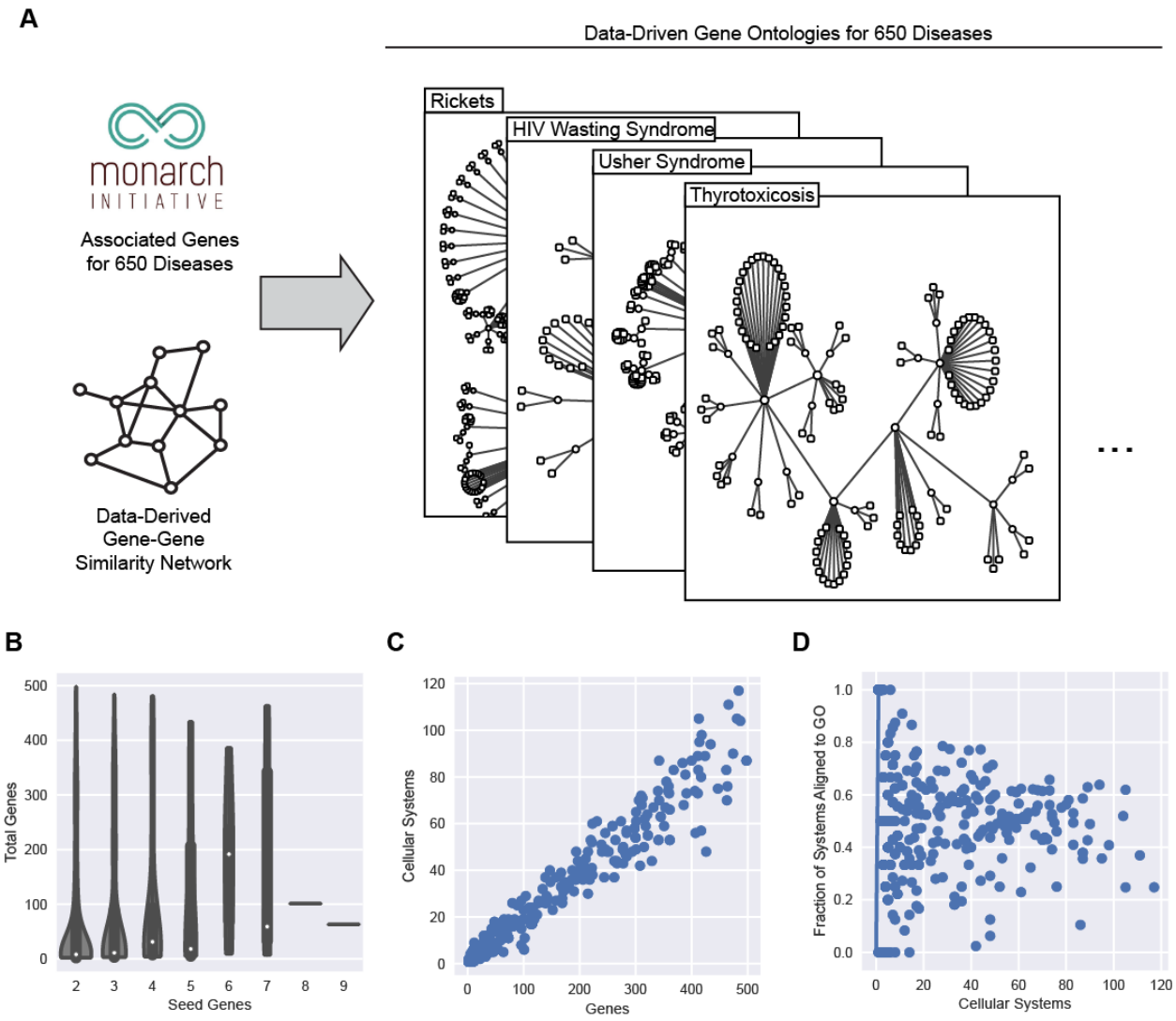


Figure 1.5. A compendium of gene ontologies for 649 diseases. (A) Data-driven gene ontologies were created to discover the genetic and pathway mechanisms of 649 diseases. Each ontology was based on a "seed" set of genes with known associations to a disease (Monarch Initiative (Mungall et al. 2017)) and a network of functional similarities between genes. **(B)** The total number of genes modeled in an ontology as a function of the number of genes in the seed set. **(C)** The number of cellular subsystems inferred in the ontology as a function of the total genes modeled. **(D)** The proportion of cellular subsystems that are aligned to GO.

1.8 Methods

Construction of data-driven gene ontologies. The 20 seed FA genes are FANCA, FANCB, FANCC, BRCA2, FANCD2, FANCE, FANCF, FANCG, FANCI, BRIP1, FANCL, FANCM, PALB2, RAD51C, SLX4, ERCC4, RAD51, BRCA1, UBE2T, and XRCC2. The set of genes that have been co-cited with Fanconi Anemia was gathered using a text mining procedure described in (Wang et al. 2017). Briefly, we gathered all Pubmed abstracts published up to date on April 30th, 2017 and identified all words that appear in an abstract with the phrase “Fanconi Anemia”. The words that are either an official or alias symbol of a human gene, according to the HGNC consortium, were counted (disregarding letter casing).

Gene-disease associations for 8,590 diseases from the Monarch Database was downloaded on May 13, 2017 with the help of Kent Shefchek and Chris Mungall (personal communication). For most diseases, the known and candidate genes together totaled more than 500 genes, suggesting the lack of a coherent molecular signature. Ontologies were constructed and studied for the 649 diseases that had no more than 500 known and candidate genes.

The gene similarity network was based on more than 1000 gene-gene interaction networks from publicly available sources (Zheng et. al., manuscript in preparation). Briefly, we compiled 16 tissue-specific coexpression networks from the GTEx project (GTEx Consortium 2013; Saha et al. 2017b), 10 PCA components of the 980 GEO co-expression networks (Edgar, Domrachev, and Lash 2002) curated by the GIANT (Greene et al. 2015), 1 co-expression network from the Cancer Cell Line Encyclopedia (Barretina et al. 2012), 2 protein domain similarity networks (InterPro (Hunter et al. 2009), PFAM

(Bateman et al. 2004)), 1 genetic interaction network from (Lin et al. 2010), 2 curated protein-protein interaction networks from databases (BioGRID (Stark et al. 2006), InBioMap (Li et al. 2017)), and 4 high-throughput protein-protein interaction networks ((Huttlin et al. 2017b), (Havugimana et al. 2012), (Rolland et al. 2014), (Hein et al. 2015)), and 1 computational predicted human interactome (Zhang et al. 2012). Following the procedure in (M. H. Kramer et al. 2017b), we integrated these networks by using features in a supervised learning (random forests) of the Resnik semantic similarity (Resnik 1999) between genes in the Gene Ontology. The similarity score of a gene pair was defined by 5-fold cross-validation.

The CliXO algorithm (M. Kramer et al. 2014) was ran with alpha=0.05 and beta=0.5. Each data-driven ontology was aligned to the Gene Ontology, downloaded on October 3, 2017, using a FDR cutoff of 0.05, calculated from 100 randomized iterations. For this alignment, we used a reduced version of GO in which terms that do not contain any human genes or contain the same genes as its parents terms were removed using the DDOT function *Ontology.precollapse()*.

The Hierarchical Viewer. In designing HiView, the rendering library was chosen based on a tradeoff between the rendering speed and the visual styling capabilities. The main panel for viewing a hierarchy's structure is rendered with sigma.js with WebGL enabled. The side panel for viewing the supporting interaction networks, on the other hand, was implemented using cytoscape.js (Franz et al. 2016), which renders more slowly than sigma.js but offers more styling capabilities out-of-the-box, such as node shapes and color gradients. The spanning tree used for the transformations in **Figures 1.4B,D**

was calculated by connecting each subsystem to its smallest parent subsystem; all other connections were hidden. All three transformations in **Figure 1.4** are laid out using the Bubble Tree Layout algorithm (Grivet et al. 2006) by calling the Tulip Python library (Auber et al. 2016).

1.9 Acknowledgments

Chapter 1, in full, is currently being prepared for submission of the material as “A Swiss-Army Knife for Hierarchical Modeling of Biological Systems”. Michael Ku Yu, Jianzhu Ma, Keiichiro Ono, Fan Zheng, Samson Fong, Aaron Gary, Jing Chen, Barry Demchak, Dexter Pratt, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

1.10 References

Auber, David, Romain Bourqui, Maylis Delest, Antoine Lambert, Patrick Mary, Guy Melançon, Bruno Pinaud, Benjamin Renoust, and Jason Vallet. 2016. “TULIP 4.” LaBRI - Laboratoire Bordelais de Recherche en Informatique. https://hal.archives-ouvertes.fr/hal-01359308.

Bard, Jonathan, Seung Y. Rhee, and Michael Ashburner. 2005. “An Ontology for Cell Types.” *Genome Biology* 6 (2):R21.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. “The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity.” *Nature* 483 (7391):603–7.

Bateman, Alex, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, et al. 2004. “The Pfam Protein Families Database.” *Nucleic Acids Research* 32 (Database issue):D138–41.

Binns, David, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O’Donovan, and Rolf Apweiler. 2009. “QuickGO: A Web-Based Tool for Gene Ontology Searching.” *Bioinformatics* 25 (22):3045–46.

Ceccaldi, Raphael, Prabha Sarangi, and Alan D. D’Andrea. 2016. “The Fanconi Anaemia Pathway: New Players and New Functions.” *Nature Reviews. Molecular Cell Biology* 17 (6):337–49.

Chong, Yolanda T., Judice L. Y. Koh, Helena Friesen, Supipi Kaluarachchi Duffy, Kaluarachchi Duffy, Michael J. Cox, Alan Moses, Jason Moffat, Charles Boone, and Brenda J. Andrews. 2015. “Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis.” *Cell* 161 (6):1413–24.

Costanzo, Michael, Benjamin VanderSluis, Elizabeth N. Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, et al. 2016. “A Global Genetic Interaction Network Maps a Wiring Diagram of Cellular Function.” *Science* 353 (6306). https://doi.org/10.1126/science.aaf1420.

Degtyarenko, Kirill, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2008. “ChEBI: A Database and Ontology for Chemical Entities of Biological Interest.” *Nucleic Acids Research* 36 (Database issue):D344–50.

Dutkowski, Janusz, Michael Kramer, Michal A. Surma, Rama Balakrishnan, J. Michael Cherry, Nevan J. Krogan, and Trey Ideker. 2013. “A Gene Ontology Inferred from Molecular Networks.” *Nature Biotechnology* 31 (1):38–45.

“Fanconi Anemia Mutation Database.” n.d. http://www2.rockefeller.edu/fanconi/.

Franz, Max, Christian T. Lopes, Gerardo Huck, Yue Dong, Onur Sumer, and Gary D.

Bader. 2016. “Cytoscape.js: A Graph Theory Library for Visualisation and Analysis.” *Bioinformatics* 32 (2):309–11.

Greene, Casey S., Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, et al. 2015. “Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.” *Nature Genetics* 47 (6):569–76.

Grivet, S., D. Auber, J. P. Domenger, and G. Melancon. 2006. “BUBBLE TREE DRAWING ALGORITHM.” In *Computer Vision and Graphics*, 633–41. Computational Imaging and Vision. Springer, Dordrecht.

GTEx Consortium. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6):580–85.

Hunter, Sarah, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, et al. 2009. “InterPro: The Integrative Protein Signature Database.” *Nucleic Acids Research* 37 (Database issue):D211–15.

Huttlin, Edward L., Raphael J. Bruckner, Joao A. Paulo, Joe R. Cannon, Lily Ting, Kurt Baltier, Greg Colby, et al. 2017. “Architecture of the Human Interactome Defines Protein Communities and Disease Networks.” *Nature* 545 (7655):505–9.

Kibbe, Warren A., Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, Christopher J. Mungall, et al. 2015. “Disease Ontology 2015 Update: An Expanded and Updated Database of Human Diseases for Linking Biomedical Knowledge through Disease Data.” *Nucleic Acids Research* 43 (Database issue):D1071–78.

Knies, Kerstin, Shojiro Inano, María J. Ramírez, Masamichi Ishiai, Jordi Surrallés, Minoru Takata, and Detlev Schindler. 2017. “Biallelic Mutations in the Ubiquitin Ligase RFWD3 Cause Fanconi Anemia.” *The Journal of Clinical Investigation* 127 (8):3013–27.

Kramer, Michael, Janusz Dutkowski, Michael Yu, Vineet Bafna, and Trey Ideker. 2014. “Inferring Gene Ontologies from Pairwise Similarity Data.” *Bioinformatics* 30 (12):i34–42.

Kramer, Michael H., Jean-Claude Farré, Koyel Mitra, Michael Ku Yu, Keiichiro Ono, Barry Demchak, Katherine Licon, et al. 2017. “Active Interaction Mapping Reveals the Hierarchical Organization of Autophagy.” *Molecular Cell* 65 (4):761–74.e5.

Lee, Insuk, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. 2011. “Prioritizing Candidate Disease Genes by Network-Based Boosting of Genome-Wide Association Data.” *Genome Research* 21 (7):1109–21.

Lin, Andy, Richard T. Wang, Sangtae Ahn, Christopher C. Park, and Desmond J. Smith. 2010. “A Genome-Wide Map of Human Genetic Interactions Inferred from Radiation Hybrid Genotypes.” *Genome Research* 20 (8):1122–32.

Li, Taibo, Rasmus Wernersson, Rasmus B. Hansen, Heiko Horn, Johnathan Mercer, Greg Slodkowicz, Christopher T. Workman, et al. 2017. “A Scored Human Protein-

Protein Interaction Network to Catalyze Genomic Interpretation." *Nature Methods* 14 (1):61–64.

Mungall, Christopher J., Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, et al. 2017. "The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species." *Nucleic Acids Research* 45 (D1):D712–22.

Park, Jisoo, Benjamin J. Hescott, and Donna K. Slonim. 2017. "Towards a More Molecular Taxonomy of Disease." *Journal of Biomedical Semantics* 8 (1):25.

Pratt, Dexter, Jing Chen, David Welker, Ricardo Rivas, Rudolf Pillich, Vladimir Rynkov, Keiichiro Ono, et al. 2015. "NDEx, the Network Data Exchange." *Cell Systems* 1 (4):302–5.

Resnik, Philip. 1999. "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language." *The Journal of Artificial Intelligence Research* 11:95–130.

Robinson, Peter N., Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. "The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease." *American Journal of Human Genetics* 83 (5):610–15.

Rolland, Thomas, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, et al. 2014. "A Proteome-Scale Map of the Human Interactome Network." *Cell* 159 (5):1212–26.

Saha, Ashis, Yungil Kim, Ariel D. H. Gewirtz, Brian Jo, Chuan Gao, Ian C. McDowell, GTEx Consortium, Barbara E. Engelhardt, and Alexis Battle. 2017. "Co-Expression Networks Reveal the Tissue-Specific Regulation of Transcription and Splicing." *Genome Research* 27 (11):1843–58.

Sefer, Emre, Michael Kleyman, and Ziv Bar-Joseph. 2016. "Tradeoffs between Dense and Replicate Sampling Strategies for High-Throughput Time Series Experiments." *Cell Systems* 3 (1):35–42.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11):2498–2504.

Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. "BioGRID: A General Repository for Interaction Datasets." *Nucleic Acids Research* 34 (Database issue):D535–39.

Sumpter, Rhea, Jr, Shyam Sirasanagandla, Álvaro F. Fernández, Yongjie Wei, Xiaonan Dong, Luis Franco, Zhongju Zou, et al. 2016. "Fanconi Anemia Proteins Function in Mitophagy and Immunity." *Cell* 165 (4):867–81.

Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, et al. 2017. “The STRING Database in 2017: Quality-Controlled Protein–protein Association Networks, Made Broadly Accessible.” *Nucleic Acids Research* 45 (D1). Oxford University Press:D362–68.

Wang, Sheng, Jianzhu Ma, Michael Ku Yu, Fan Zheng, Edward W. Huang, Jiawei Han, Jian Peng, and Trey Ideker. 2017. “Annotating Gene Sets by Mining Large Literature Collections with Protein Networks.” In Biocomputing 2018. https://doi.org/10.1142/9789813235533_0055.

Yu, Michael Ku, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason Kreisberg, Cherie T. Ng, Nevan Krogan, Roded Sharan, and Trey Ideker. 2016. “Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems.” *Cell Systems* 2 (2):77–88.

CHAPTER 2: TRANSLATION OF GENOTYPE TO PHENOTYPE BY A HIERARCHY OF CELL SUBSYSTEMS

2.1 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 impacting DNA repair or nuclear lumen. Ontotypes also generalize to larger knockout combinations, setting the stage for interpreting the complex genetics of disease.

2.2 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 associate 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 Hacohen, 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 (1nm scale) give rise to functional changes in proteins (1-10nm), which in turn affect protein complexes (10-100nm), cellular processes (100nm), organelles (1μm) and, ultimately, phenotypic behaviors of cells (1-10μm), tissues (100μm-100mm) and complex organisms (>1m). 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 (The Gene Ontology Consortium, 2014). Thus far, genotype-phenotype association methods have sometimes used prior knowledge in 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: 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 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 data sets, including molecular interactions and gene expression profiles; we call this assembly NeXO (Dutkowski et al. 2013; Kramer et al. 2014; Gligorijević, Janjić, and Pržulj 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 upwards through the hierarchy: in this way, genetic variants comprising genotype are converted to effects on the cell subsystems impacted by those variants. We call the set of all such effects '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 one of the largest genotype-phenotype compendia in existence. We integrate these data with 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 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.

2.3 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 2.1A**). 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.

Due to 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 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 2.1B**). We found that such hierarchical relationships were widespread in GO: approximately half of within-term enrichments could be factored into between-term enrichments among their descendants (**Figure 2.1C**). 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 2.1D**). Thus, GO integrates genetic interactions in an overarching hierarchy capturing multiple scales of cell biology. As one moves upwards 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. *bΔdΔ*, **Figure 2.2A**). 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, since the gene *KIP1* encodes a subunit of the kinesin complex (**Figure 2.1B**), its deletion in a *kip1Δ* strain propagates upwards in the ontology to affect the parent term 'kinesin complex' and continues to propagate upwards 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 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 2.2A**). 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/false tests to evaluate the states of several terms (e.g., T4, T5, and T7 in **Figure 2.2A**). 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. a*ΔdΔ* 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 Gene Ontology 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 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, including protein-protein interactions, gene expression profiles, and protein sequence properties but excluding any prior information about genetic interactions (datasets taken from YeastNet v3 study, Kim et al., 2014). NeXO (4,805 terms) was tuned so that the resulting ontology was approximately similar in size to 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 (**Figure 2.3A,B**). The correlation between predicted and measured scores was highly significant (**Figure 2.3C**, 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 (Baryshnikova et al., 2010), was r = 0.67. Progressively removing either small or large terms from the model degraded the correlation (**Figure 2.3D,E**), indicating that all scales in the hierarchy aid in prediction. F_{NeXO} achieved nearly the same correlation (**Figure 2.3C**, r = 0.32) and was also sensitive to randomization (r = 0.03).

Both functionalized ontologies compared favorably to 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 based on the phenotypes of their network neighbors (I. Lee et al. 2010); and the Multi-Network Multi-Classifier (MNMC), a 'black box' supervised learning system which uses many different lines of experimental evidence as features to predict genetic interactions (Pandey et al., 2010, **Experimental Procedures**). In comparison to all of these approaches, the functionalized ontologies achieved substantially greater correlation between predicted and measured interaction scores (**Figure 2.3C**) as well as better tradeoffs in precision versus recall (**Figure 2.3F**) in four-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

(Mackay 2014; Gillis and Pavlidis 2012) could be used to explain training data but not test data. In spite of this challenge, F_{GO} still outperformed predictions made with a randomized GO or with the non-hierarchical methods (**Supplemental Figure S2.1**).

We found that the accuracy of growth phenotype prediction depends significantly on the degree to which cellular systems have been characterized in the gene ontology. 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 (**Supplemental Figure S2.2**). Moreover, many genes that are poorly characterized in GO are better characterized in NeXO, such that genotypes involving these genes lead to better phenotypic predictions by F_{NeXO} than by F_{GO} (**Supplemental Figure S2.2A-C**). 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 equally good predictions could be made from 'flat' features derived from the same ontologies. 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 from the two ontologies resulted in a substantial degradation in prediction performance ($FLAT_{GO}$ and $FLAT_{NeXO}$, **Figure**

2.3C), 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 2.4A, Supplemental File S2.1**). A total of 41,605 genetic interactions were predicted. These predictions were concentrated within and between particular terms and term pairs (**Figures 2.4A,B**), 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 2.4C**), with negative interactions linking the sub-systems of electron and proton transport and positive interactions segregating entirely within electron transport. These distinct patterns of positive/negative segregation were observed broadly across F_{GO} (**Supplemental Figure S2.3**). Of particular interest were predicted interactions between 71 term pairs, as these terms were only distantly related in GO (**Table 2.1, Supplemental Table S2.1, 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 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.

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 (**Figure 2.5A, Supplemental Figure S2.4**). 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. 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, **Supplemental Figure S2.2G**, **Supplemental Table S2.2**). 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 (**Supplemental Table S2.3, Supplemental Experimental Procedures**). Of these, 1,345 mutants had also been scored previously by Costanzo et al. 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 2.5B**). 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 (Costanzo et al., 2010) but not by our new screens. This limited F_{GO} achieved a high sensitivity versus specificity (**Figure 2.5C**) and precision versus recall

(**Figure 2.5D**) 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/cYIXWJ, UUID: 01b46d52-c3a5-11e5-8fbc-06603eb7f303, Pratt et al., 2015) and as a Cytoscape file in **Supplemental File S2.2**.

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*, *HPC2*, *RTT106*), results in a synthetic growth defect (negative genetic interaction); however, the 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 2.6A**). Visual inspection of the model (**Figure 2.6B**) implicated decision logic based on 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*ΔHIRΔ), 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.

2.4 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 ontotype— give rise to a phenotype of interest. Together, these two steps constitute a paradigm by which 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 (**Figure 2.3C,F**), a notable finding given the simplicity of the ontotype 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 2.3C**) or truncated versions of the ontology (**Figure 2.3D,E**). 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 2.1B**). A

genotype may include some mutations that map to the same gene, others to the same protein complex; still others to different complexes but 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. 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.

The last factor that 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 ontotype requires no parameters and instead leverages the topology of the ontology. Logical rules for predicting phenotype are based on the ontotype, 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. While 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 2.4A**). 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 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, where it is often challenging to interpret the effects of individual genes due to many multi-gene interactions (Kim and Przytycka, 2012; Zuk et al., 2012). Towards 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 towards 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.

2.5 Experimental Procedures

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 × 3,885 arrays), were downloaded from http://drygin.ccbr.utoronto.ca/~costanzo2009/. Double gene deletion mutants impacting 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 Gene Ontology (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 GO relevant to yeast (**Supplemental Table S2.4**), following a previously described procedure (Dutkowski et al., 2013, http://mhk7.github.io/alignOntology/).

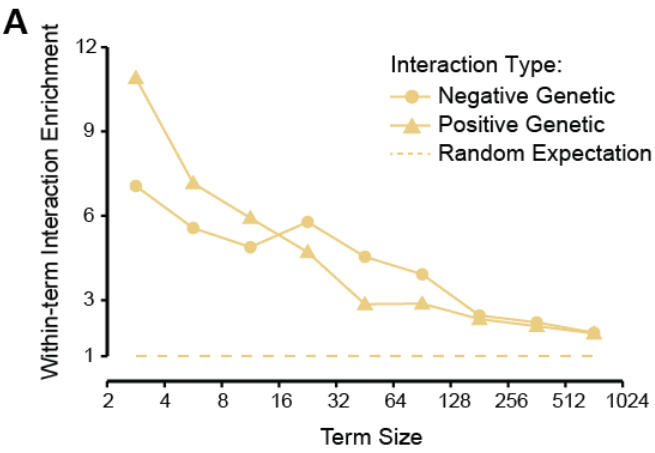
To construct NeXO (**Supplemental Table S2.5**), we integrated the YeastNet v3 networks (Kim et al., 2014), spanning 68 experimental studies across 8 data types excluding genetic interactions, into a single network, and then applied the method of Clique Extracted Ontology (CliXO) (Kramer et al., 2014, http://mhk7.github.io/clixo_0.3/). See also **Supplemental Experimental Procedures**.

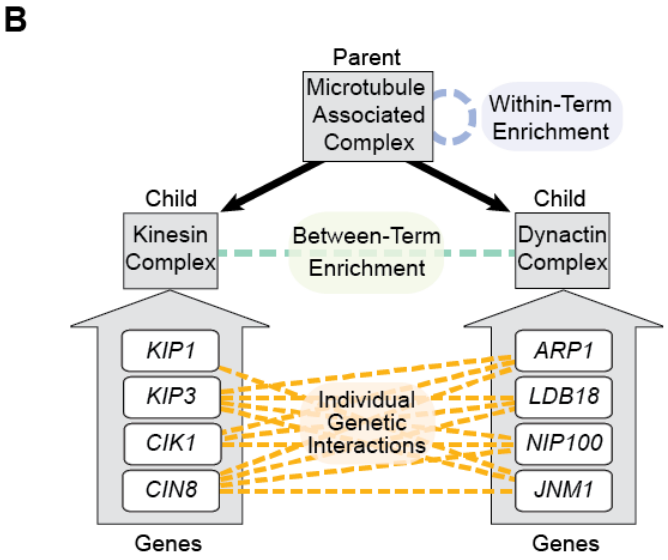
Random forests regression. Random forests (Breiman, 2001) were used to regress genetic interaction scores ϵ_{ab}, as described in the **Results**. Due to the very large size of the ontotype feature matrix, we optimized the random forest library from the Python scikit-learn package (Pedregosa et al., 2011); the modified code is available at https://github.com/michaelkyu/scikit-learn-fasterRF. While trees grown at approximately 29% (GO) or 37% (NeXO) of the maximal depth did improve performance slightly (<0.02 gain in correlation, **Supplemental Figure S2.5**), we chose to grow trees to maximal depth because it is unclear how significant this gain is and whether it would reproducible in different random partitions of the data for cross validation or in different genotype-phenotype datasets. See also **Supplemental Experimental Procedures**.

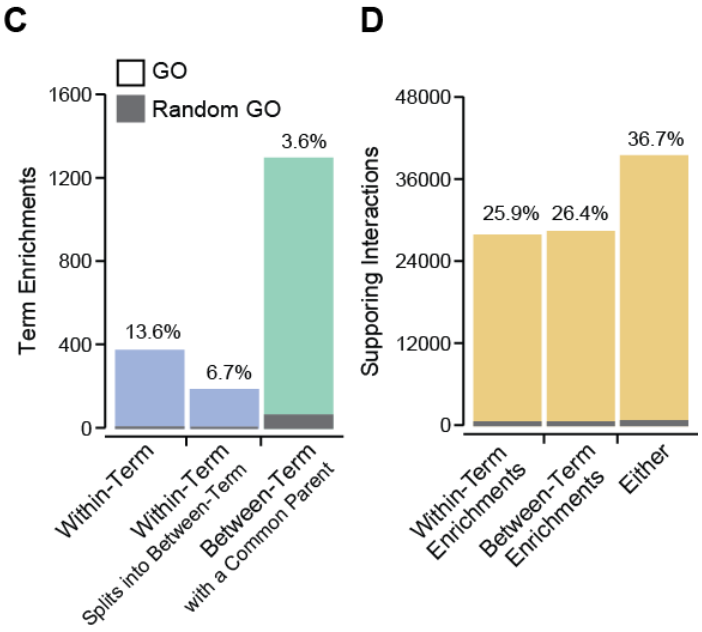
Comparison of methods for predicting genetic interactions. The MNMC method was updated from the original (Pandey et al., 2010), which was trained on a set of literature-curated synthetic lethal interactions that was much smaller in size than the set of genetic interactions considered in our study, and 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 basic 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 (**Supplemental Figure S2.6**); this performance difference may be due to the five basic features being collected more recently. See also **Supplemental Experimental Procedures**.

2.6 Figure and Tables

Figure 2.1: Patterns of genetic interaction reflect the hierarchical structure of the Gene Ontology. (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. 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).


A
Within-term Interaction Enrichment
12
9
6
3
1
2
4
8
16
32
64
128
256
512
1024
Term Size
Interaction Type:
Negative Genetic
Positive Genetic
Random Expectation


B
Parent
Microtubule Associated Complex
Within-Term Enrichment
Child
Kinesin Complex
Between-Term Enrichment
Child
Dynactin Complex
KIP1
KIP3
CIK1
CIN8
Individual Genetic Interactions
ARP1
LDB18
NIP100
JNM1
Genes
Genes


C
GO
Random GO
Term Enrichments
1600
1200
800
400
0
13.6%
6.7%
3.6%
Within-Term
Within-Term Splits into Between-Term
Between-Term with a Common Parent
D
Supporing Interactions
48000
36000
24000
12000
0
25.9%
26.4%
36.7%
Within-Term Enrichments
Between-Term Enrichments
Either

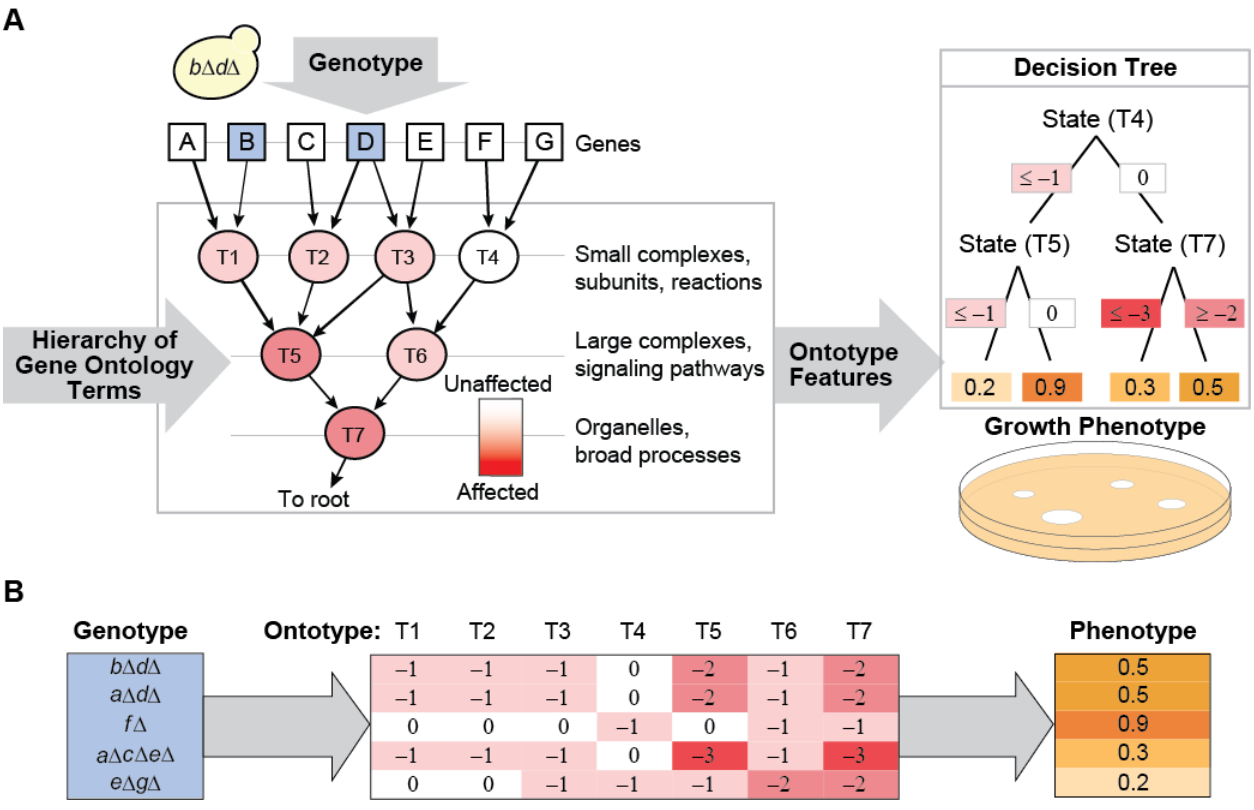


Figure 2.2: The ontotype method of translating genotype to phenotype. (A) The relationship between genotypic and phenotypic variation is modeled through an intermediate 'ontotype', defined as the profile of states corresponding to the effect of genotype on each cellular component, biological process, 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 / phenotype associations from the ontology in **(A)**. Different genotypes (e.g. *b*Δ*d*Δ and *a*Δ*d*Δ) give rise to similar or identical phenotypes by influencing similar or identical combinations of terms.

Figure 2.3: Genome-wide prediction of pairwise genetic interactions in yeast. (A) Measured genetic interaction scores versus those predicted from ontotypes constructed from GO using four-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. Comparison is made to ontotypes constructed from a randomized GO 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 GO. The ontotype correlations do not fluctuate greatly (<4%) whether computed over all gene pairs (shown) or the FBA gene pairs. See also **Supplemental Figure S1-2**. (**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 criteria. **(F)** Precision-recall curves for classification of negative genetic interactions.

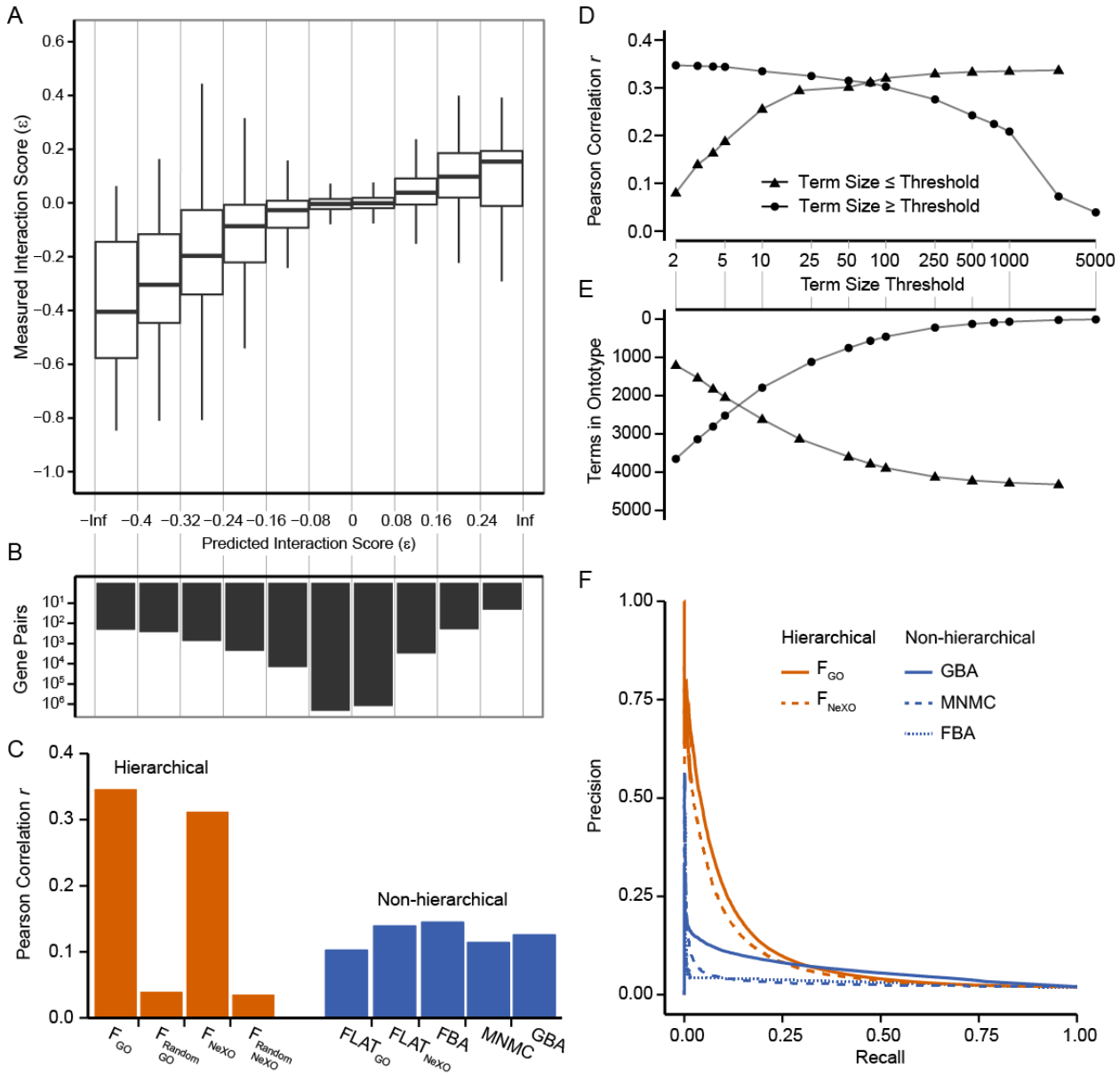

A
Measured Interaction Score (ε)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−Inf −0.4 −0.32 −0.24 −0.16 −0.08 0 0.08 0.16 0.24 Inf
Predicted Interaction Score (ε)
B
Gene Pairs
10¹
10²
10³
10⁴
10⁵
10⁶
C
Pearson Correlation r
Hierarchical
Non-hierarchical
F GO
F Random GO
F NeXO
F Random NeXO
FLAT GO
FLAT NeXO
FBA
MNMC
GBA
D
Pearson Correlation r
Term Size ≤ Threshold
Term Size ≥ Threshold
2 5 10 25 50 100 250 500 1000 5000
Term Size Threshold
E
Terms in Ontotype
0
1000
2000
3000
4000
5000
F
Precision
Recall
Hierarchical
F GO
F NeXO
Non-hierarchical
GBA
MNMC
FBA
0.00 0.25 0.50 0.75 1.00

Figure 2.4: The Functionalized Gene Ontology. (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 **Supplemental Figure S2.3**.

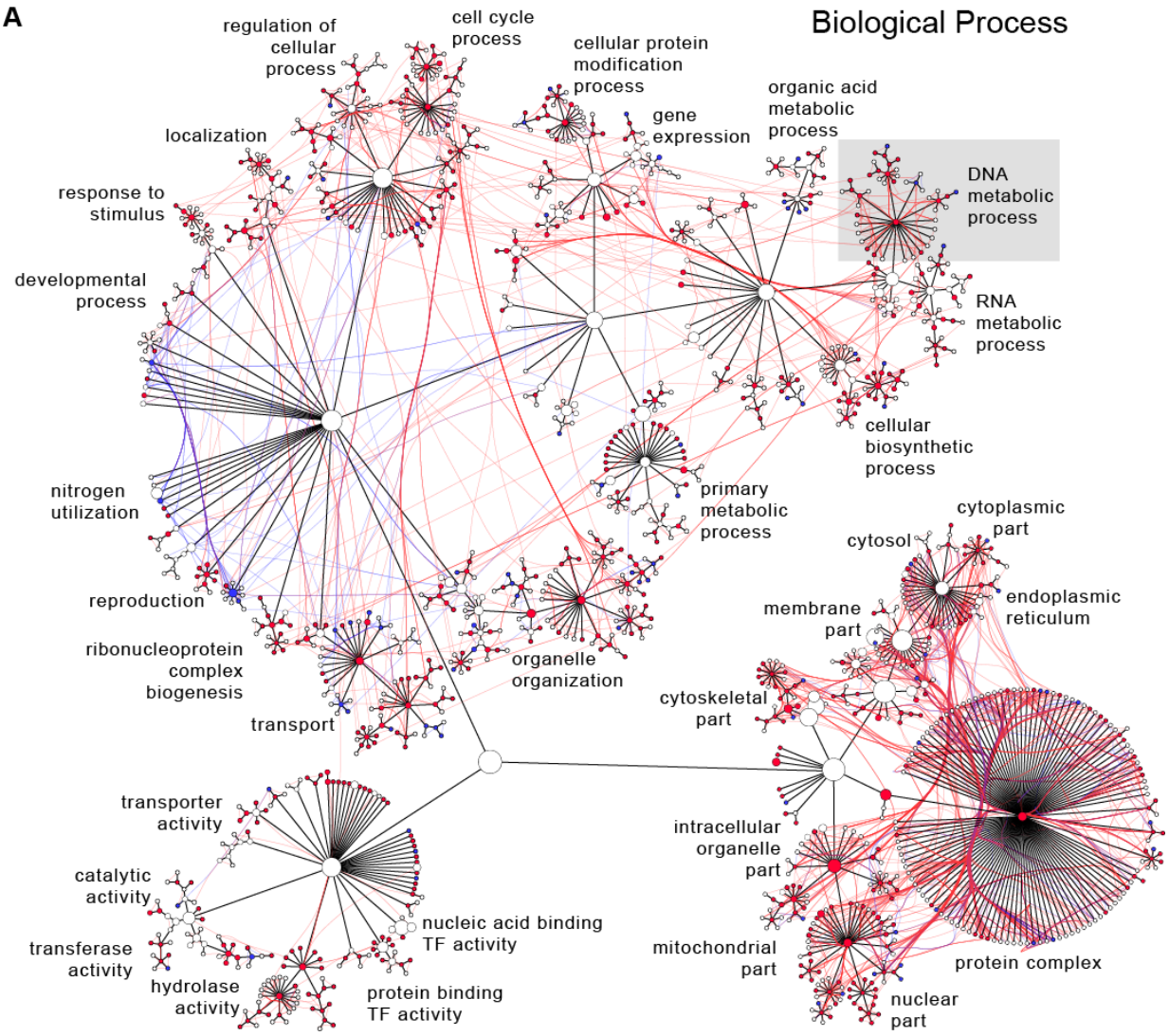
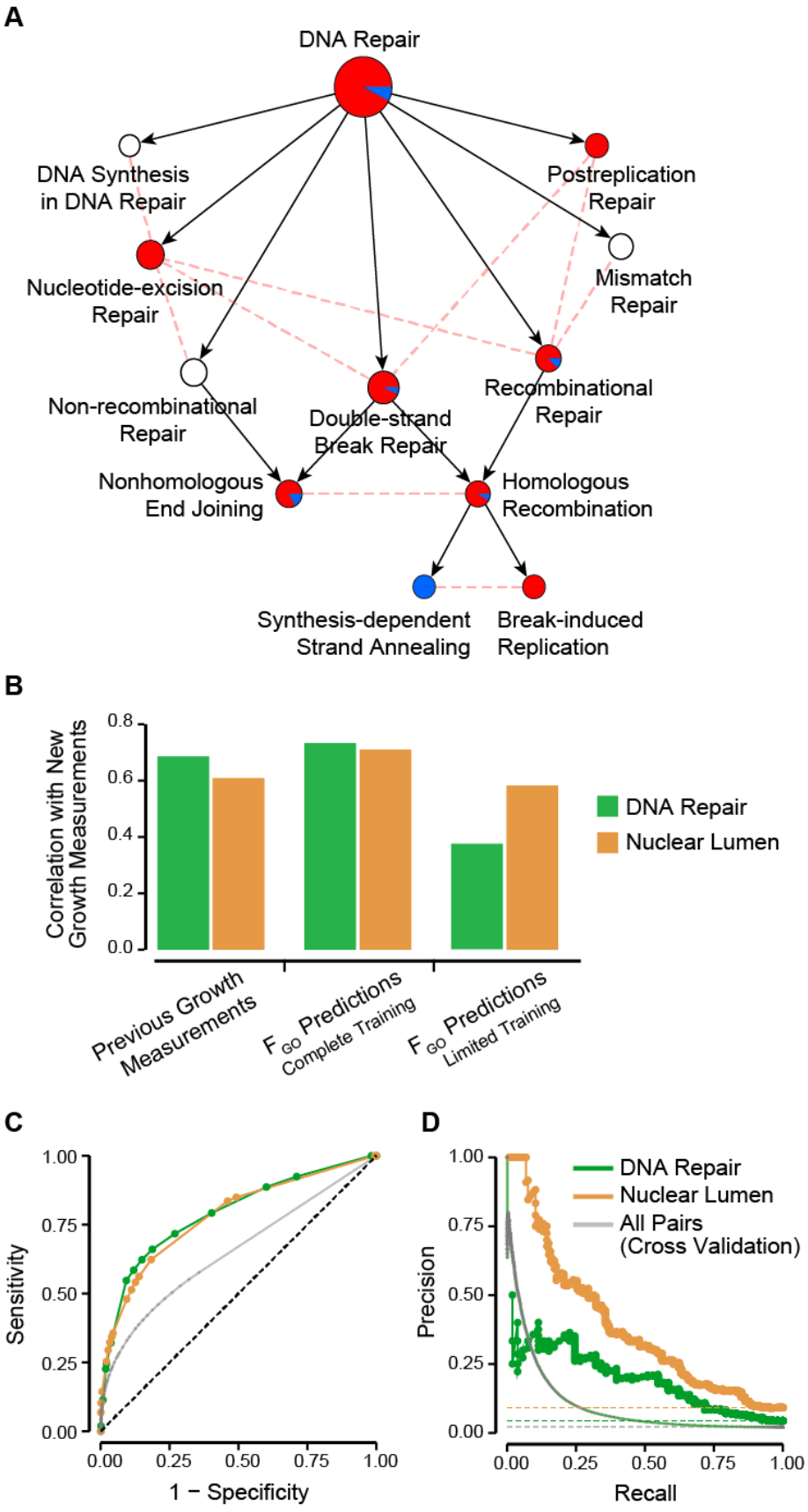

A
Biological Process
regulation of cellular process
cell cycle process
cellular protein modification process
organic acid metabolic process
gene expression
localization
response to stimulus
DNA metabolic process
developmental process
RNA metabolic process
cellular biosynthetic process
primary metabolic process
nitrogen utilization
cytoplasmic part
cytosol
endoplasmic reticulum
membrane part
reproduction
ribonucleoprotein complex biogenesis
organelle organization
cytoskeletal part
transport
transporter activity
intracellular organelle part
catalytic activity
nucleic acid binding TF activity
transferase activity
mitochondrial part
protein complex
hydrolase activity
protein binding TF activity
nuclear part
Molecular Function
Cellular Component
Parent-Child Relation
Interaction Enrichment: Positive Negative
Within-Term:
Between-Term:
Term Size
Small
Medium
Large
B
Within-Term
Between-Term
18
270
296
181
851
1,214
Measured
Predicted
C
Oxidative Phosphorylation
Mitochondrial ATP Synthesis Coupled Electron Transport
ATP Synthesis Coupled Proton Transport
Positive
Fraction of Interactions
Negative

Figure 2.5: 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 previous figures. See also **Supplemental Figure S2.4**. **(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 **Supplemental Table S2.2-2.3**. **(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. 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 2.3F**).


A
DNA Repair
DNA Synthesis in DNA Repair
Postreplication Repair
Nucleotide-excision Repair
Mismatch Repair
Non-recombinational Repair
Double-strand Break Repair
Recombinational Repair
Nonhomologous End Joining
Homologous Recombination
Synthesis-dependent Strand Annealing
Break-induced Replication
B
Correlation with New Growth Measurements
0.8
0.6
0.4
0.2
0.0
DNA Repair
Nuclear Lumen
Previous Growth Measurements
F GO Predictions Complete Training
F GO Predictions Limited Training
C
Sensitivity
1.00
0.75
0.50
0.25
0.00
1 − Specificity
D
Precision
DNA Repair
Nuclear Lumen
All Pairs (Cross Validation)
Recall

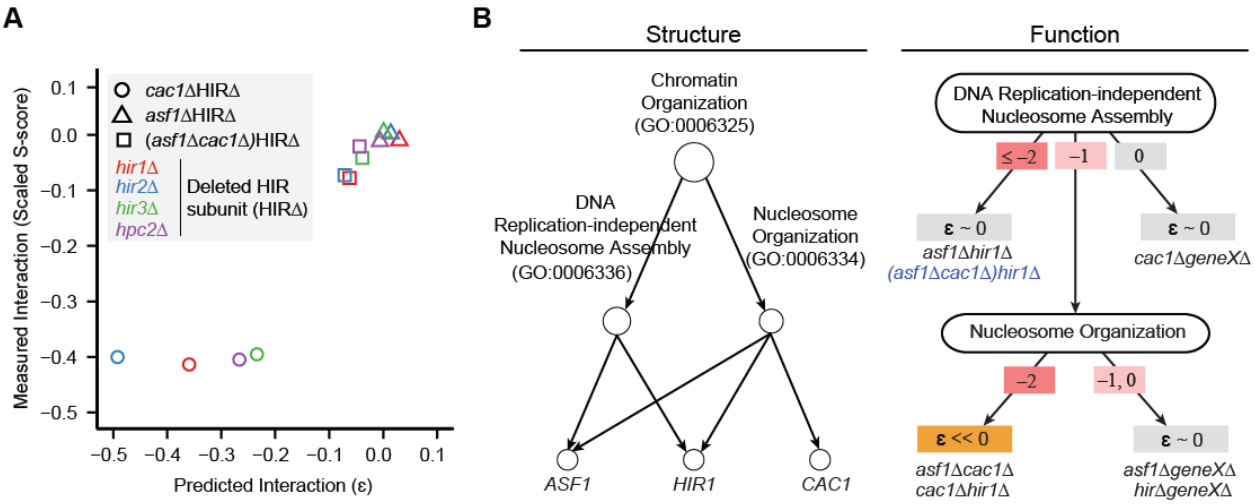


Figure 2.6: Prediction of triple mutants. (a) Measured versus predicted interaction scores for genotypes involving pairwise and three-way deletions involving *ASF1*, *CAC1*, and genes in the HIRA complex (*HIR1*, *HIR2*, *HIR3*, *HPC2*) (Haber et al., 2013). **(B)** Relevant GO structure (left) and corresponding 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 ontotype: numbers on arrows are term states; arrows point to predicted interaction scores (ε).

Table 2.1: Top new functional relationships in F_{GO}. See also Supplemental Table S2.1.

	Term A (# of Genes)	Term B (# of Genes)	Interactions / Total (%)	*p*-value[a]
Negative Interactions	intron homing (10)	phosphatidylinositol 3-kinase complex II (4)	40/40 (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%)	3.56E-55
	pre-mRNA binding (5)	RNA pol II transcription coactivator activity in preinitiation complex assembly (3)	15/15 (100.0%)	2.86E-32
	protein lipoylation (4)	carbon-oxygen lyase activity, acting on phosphates (3)	12/12 (100.0%)	1.23E-24
	Swr1 complex (8)	U6 snRNP (3)	22/24 (91.7%)	1.20E-47
	alpha-1,6-mannosyltransferase complex (6)	negative regulation of chromatin silencing involved in replicative cell aging (4)	21/24 (87.5%)	3.08E-44
	tubulin complex assembly (5)	maintenance of DNA trinucleotide repeats (3)	13/15 (86.7%)	3.67E-25
	inositol phosphate biosynthetic process (5)	minus-end-directed microtubule motor activity (3)	12/15 (80.0%)	5.56E-22
	regulation of ARF GTPase activity (6)	phosphatidylinositol-3,5-bisphosphate 5-phosphatase activity (4)	19/24 (79.2%)	7.92E-38
	regulation of histone H2B conserved C-terminal lysine ubiquitination (5)	HIR complex (4)	14/20 (70.0%)	3.82E-25
	negative regulation of chromatin silencing at silent mating-type cassette (8)	U6 snRNP (3)	19/24 (79.2%)	7.92E-38
Positive Interactions	tubulin complex assembly (5)	DNA-directed RNA polymerase I complex (4)	15/20 (75.0%)	4.37E-28
	RSC complex (8)	inactivation of MAPK activity (4)	19/32 (59.4%)	6.33E-34
	vacuolar proton-transporting V-type ATPase, V1 domain (8)	free ubiquitin chain polymerization (3)	14/24 (58.3%)	1.91E-23
	alpha-1,6-mannosyltransferase complex (6)	dynactin complex (5)	16/30 (53.3%)	1.14E-26
	vacuolar proton-transporting V-type ATPase, V0 domain (7)	AP-3 adaptor complex (4)	13/28 (46.4%)	1.26E-19
	SLIK (SAGA-like) complex (14)	positive regulation of stress-activated MAPK cascade (3)	14/42 (33.3%)	4.92E-19
	histone exchange (9)	minus-end-directed microtubule motor activity (3)	9/27 (33.3%)	2.38E-10
	histone methyltransferase activity (H3-K4 specific) (7)	snoRNA transcription from an RNA polymerase II promoter (3)	7/21 (33.3%)	7.33E-07
	glycerol transport (4)	transcription-coupled nucleotide-excision repair (4)	5/16 (31.3%)	3.42E-03

[a] Bonferroni corrected for family-wise error rate

2.7 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 × 3,885 arrays), were downloaded from http://drygin.ccbr.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 × array B versus query B × 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 S2.3**) 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 S2.2**). 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 2.3C,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 α=0.04 and β=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 2.1A**, **Figure 2.1C,D**, and **Figure 2.3C**, 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 ontotype feature matrix, e.g., 3,331,172 genotypes (rows) by 4,341 terms (columns) for GO, training a random forest on the ontotype 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 S2.5**), 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 (Kelly, Lamb, and Kelly 2001) and annotated to at least one non-root term in any of the GO branches (**Supplemental File S2.1**). 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 2.5B** 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 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 S2.6**); this performance difference may have been due to calculating more recent versions of the base features. For **Figure 2.3C,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 (I. 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 2.3C**) 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 2.3C,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 S2.1**, the set of 3,686 genes in the processed GO were partitioned into 8 approximately equal-sized subsets G_1,G_2…,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, Palsson, and Herrgård 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 2.1C**, 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 2.1, Supplemental Table S2.1**).

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 2.4A** (http://goo.gl/ViTztH, UUID: 2f0cd18a-c3a4-11e5-8fbc-06603eb7f303), **Figure 2.4C** (http://goo.gl/iWmvNk, UUID: 420e4e54-c3a2-11e5-8fbc-06603eb7f303), **Figure 2.5A** (http://goo.gl/mPkMVs, UUID: 555cd256-c3a3-11e5-8fbc-06603eb7f303), and **Supplemental Figure S2.4** (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 S2.2**.

2.8 Supplemental Figures

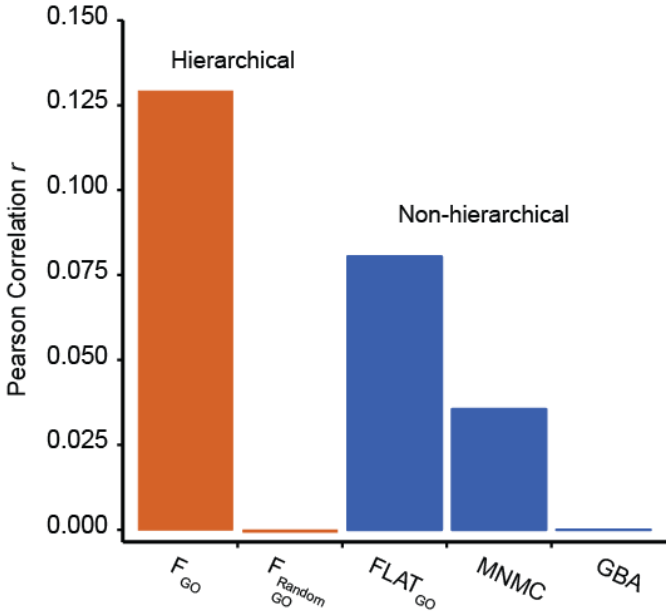
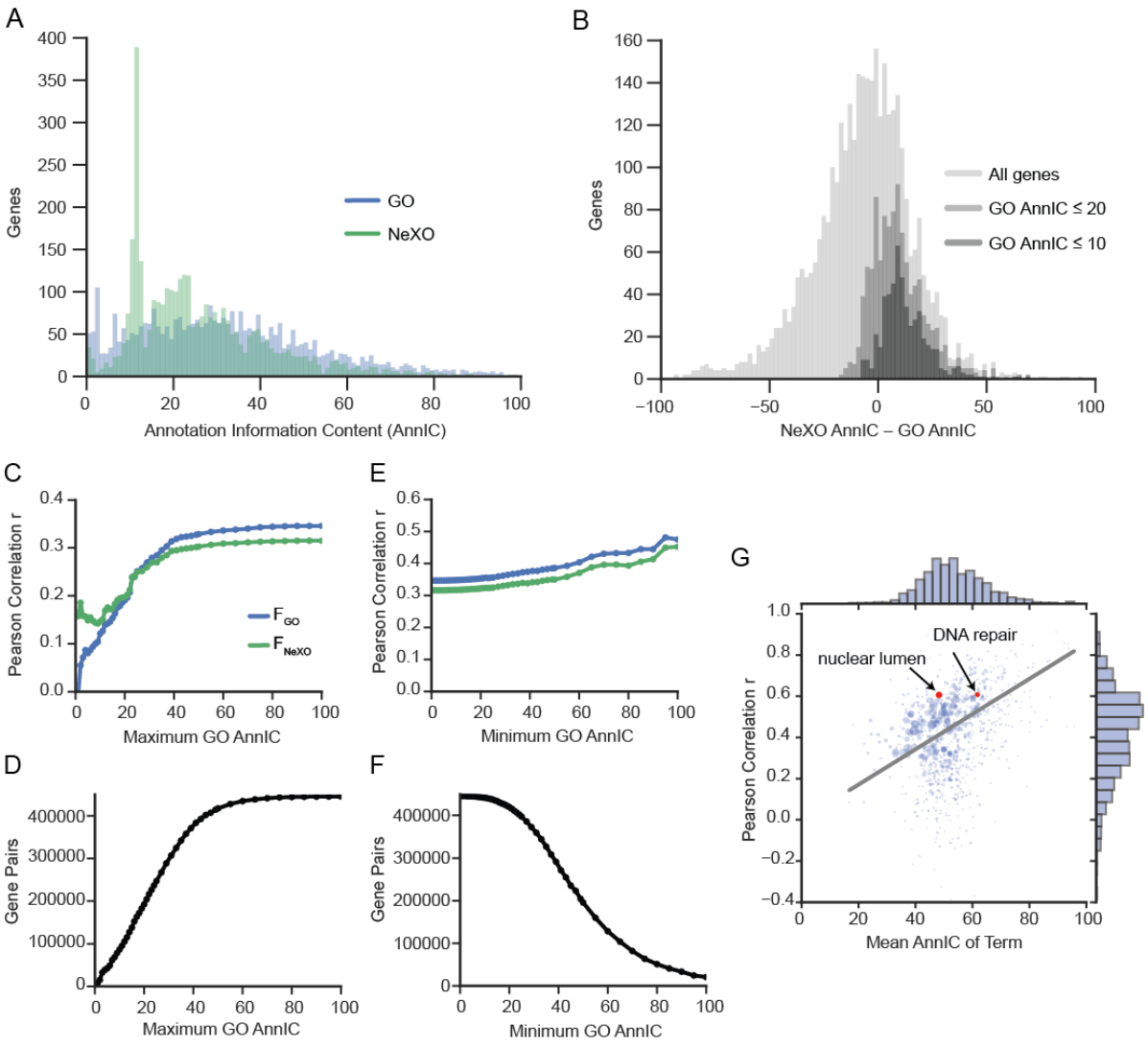


Figure S2.1. Prediction of pairwise genetic interactions under a stringent cross-validation setup, related to Figure 2.3. Gene pairs were partitioned such that genes represented in the training set are absent in the corresponding test set (**Supplemental Experimental Procedures**). Correlation was calculated across gene pairs that meet an interaction significance criterion of $p<0.05$ in Costanzo et al.

Figure S2.2. Fuller characterization of genes in an ontology increases prediction accuracy, related to Figure 2.3. 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.


A
Genes
Annotation Information Content (AnnIC)
GO
NeXO
B
Genes
NeXO AnnIC – GO AnnIC
All genes
GO AnnIC ≤ 20
GO AnnIC ≤ 10
C
Pearson Correlation r
Maximum GO AnnIC
F_{GO}
F_{NeXO}
D
Gene Pairs
Maximum GO AnnIC
E
Pearson Correlation r
Minimum GO AnnIC
F
Gene Pairs
Minimum GO AnnIC
G
Pearson Correlation r
Mean AnnIC of Term
DNA repair
nuclear lumen

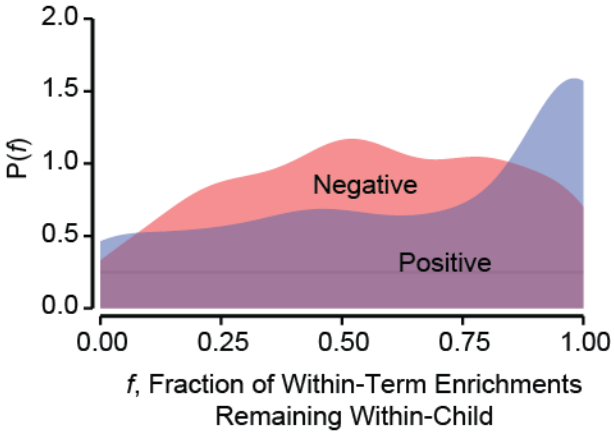


Figure S2.3: Segregation of positive and negative interactions across F_{GO}, related to Figure 2.4. For every term enriched for within-term genetic interactions, we calculate the fraction of positive versus negative interactions that continue to fall within one of its children. The distribution of this 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.

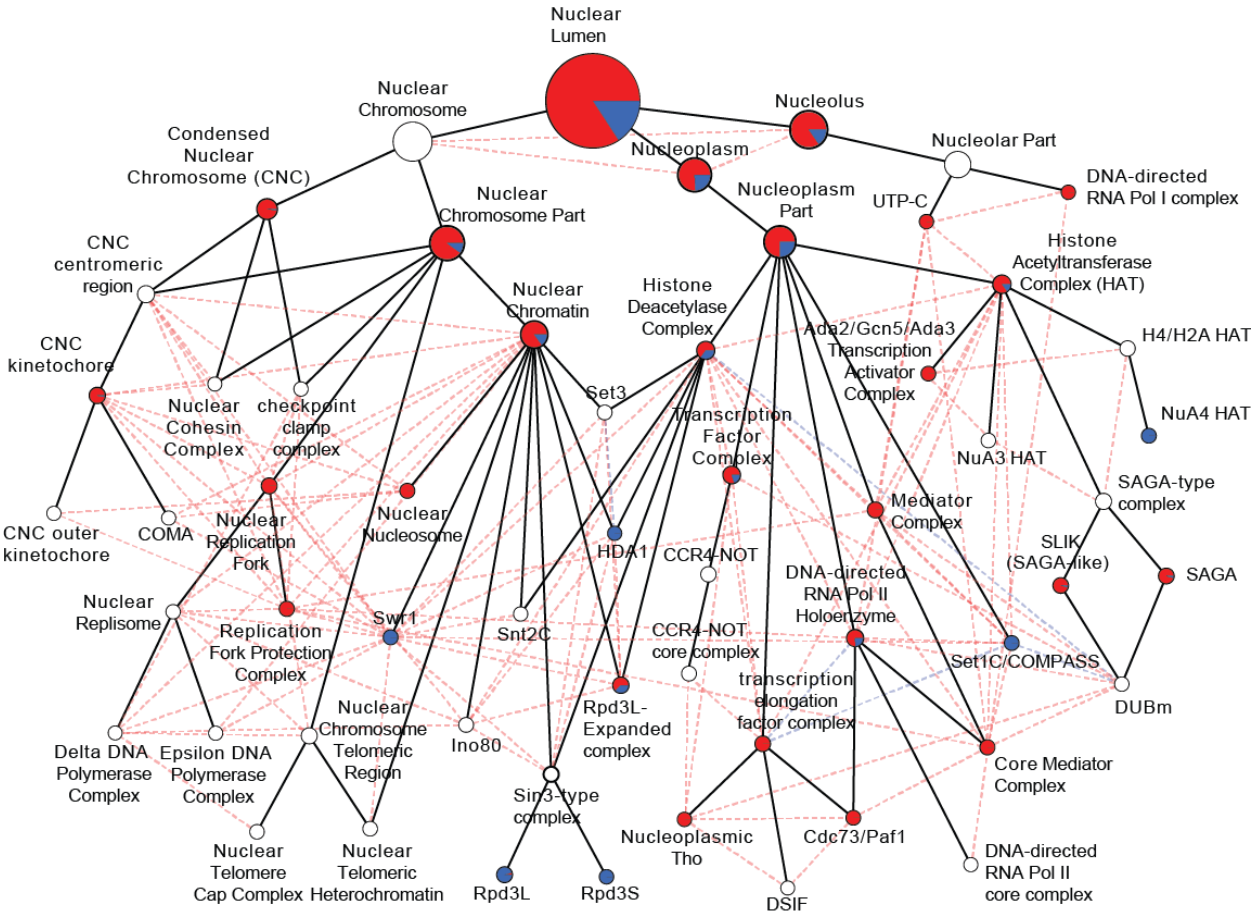


Figure S2.4: Genetic logic of the nuclear lumen, related to Figure 2.5. The nuclear lumen contains a rich hierarchy of cellular spaces, sub-components, and protein complexes interwoven by genetic interactions. Coloring and visual style of panels follow the convention of previous figures.

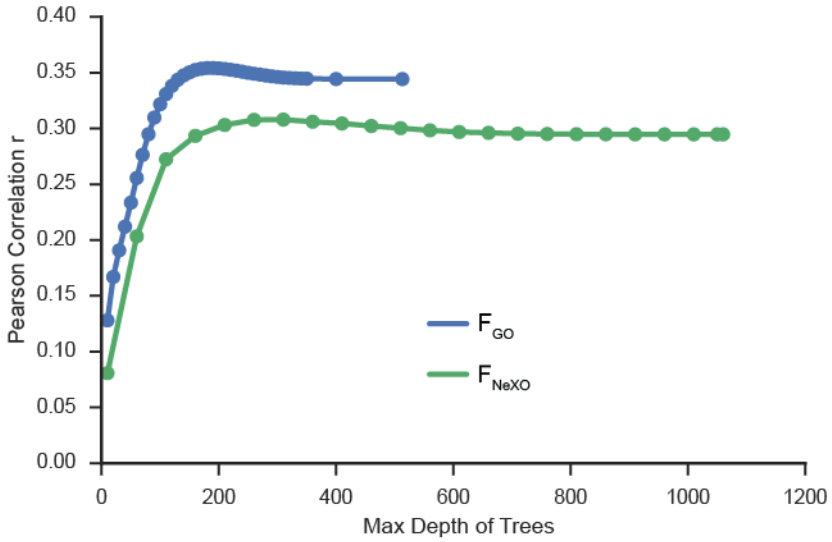


Figure S2.5: Prediction performance versus the depth of decision trees, related to Experimental Procedures. Decision trees in F_{GO} and F_{NeXO} were pruned to progressively shorter depths and evaluated across one of the cross-validation folds used in **Figure 2.3C**.

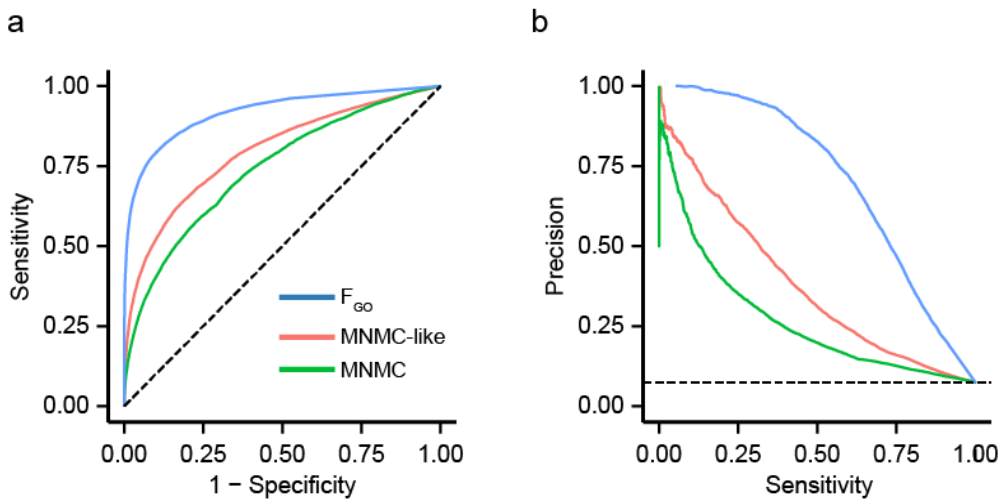


Figure S2.6: Prediction of synthetic lethal interactions curated in the Saccharomyces Genome Database, related to Experimental Procedures. 9,963 synthetic lethal and 124,312 non-synthetic lethal relations, as used in the original MNMC study (Pandey et al., 2010), were taken from the May 2008 version 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 2.3C,F**).

2.9 Author Contributions

MKY, JD, MK, R. Sharan, and TI designed the study and developed the conceptual ideas. MK constructed NeXO. MKY implemented all other computational methods and analysis. MKY and TI wrote the manuscript with input from the other authors. R. Srivas and KL performed the DNA repair and nuclear lumen interaction screen.

2.10 Acknowledgements

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). MY 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 no. 241/11). MK 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). The authors wish to declare no competing financial interests related to this work.

Chapter 2, in full, is a reformatted reprint of the material as it appears as "Translation of genotype to phenotype by a hierarchy of cell subsystems" in *Cell Systems*, 2017. Michael Ku Yu, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason F. Kreisberg, Cherie T. Ng, Nevan Krogan, Roded Sharan, and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

2.11 References

Apgar, Joshua F, David K Witmer, Forest M White, and Bruce Tidor. 2010. “Sloppy Models, Parameter Uncertainty, and the Role of Experimental Design.” *Molecular bioSystems* 6 (10):1890–1900. https://doi.org/10.1039/b918098b.

Ashyraliyev, Maksat, Yves Fomekong-Nanfack, Jaap a Kaandorp, and Joke G Blom. 2009. “Systems Biology: Parameter Estimation for Biochemical Models.” *The FEBS Journal* 276 (4):886–902. https://doi.org/10.1111/j.1742-4658.2008.06844.x.

Balakrishnan, Rama, Midori a Harris, Rachael Huntley, Kimberly Van Auken, and J Michael Cherry. 2013. “A Guide to Best Practices for Gene Ontology (GO) Manual Annotation.” *Database : The Journal of Biological Databases and Curation* 2013 (January). https://doi.org/10.1093/database/bat054.

Bandyopadhyay, Sourav, Ryan Kelley, Nevan J Krogan, and Trey Ideker. 2008. “Functional Maps of Protein Complexes from Quantitative Genetic Interaction Data.” *PLoS Computational Biology* 4 (4):e1000065. https://doi.org/10.1371/journal.pcbi.1000065.

Baryshnikova, Anastasia, Michael Costanzo, Yungil Kim, Huiming Ding, Judice Koh, Kiana Toufighi, Ji-young Youn, et al. 2010. “Quantitative Analysis of Fitness and Genetic Interactions in Yeast on a Genome Scale.” *Nature Methods* 7 (12):1017–24. https://doi.org/10.1038/nmeth.1534.

Bellay, Jeremy, Gowtham Atluri, Tina L Sing, Kiana Toufighi, Michael Costanzo, Philippe Souza Moraes Ribeiro, Gaurav Pandey, et al. 2011. “Putting Genetic Interactions in Context through a Global Modular Decomposition.” *Genome Research* 21 (8):1375–87. https://doi.org/10.1101/gr.117176.110.

Boucher, Benjamin, and Sarah Jenna. 2013. “Genetic Interaction Networks: Better Understand to Better Predict.” *Frontiers in Genetics* 4 (December):290. https://doi.org/10.3389/fgene.2013.00290.

Brachman, Ronald J, and Hector J Levesque. 2004. *Knowledge Representation and Reasoning*. The Morgan Kaufmann Series in Artificial Intelligence. San Francisco: Morgan Kaufmann. https://doi.org/http://dx.doi.org/10.1016/B978-155860932-7/50103-5.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1):5–32.

Bruggeman, Frank J, and Hans V Westerhoff. 2007. “The Nature of Systems Biology.” *Trends in Microbiology* 15 (1):45–50. https://doi.org/10.1016/j.tim.2006.11.003.

Cahan, Patrick, Hu Li, Samantha A. Morris, Edroaldo Lummertz da Rocha, George Q. Daley, and James J. Collins. 2014. “CellNet: Network Biology Applied to Stem Cell Engineering.” *Cell* 158 (4):903–15. https://doi.org/10.1016/j.cell.2014.07.020.

Califano, Andrea, Atul J Butte, Stephen Friend, Trey Ideker, and Eric Schadt. 2012. “Leveraging Models of Cell Regulation and GWAS Data in Integrative Network-Based Association Studies.” *Nature Genetics* 44 (8). Nature Publishing Group:841–47. https://doi.org/10.1038/ng.2355.

Carrera, Javier, Raissa Estrela, Jing Luo, Navneet Rai, and Athanasios Tsoukalas. 2014. “An Integrative, Multi-Scale, Genome-Wide Model Reveals the Phenotypic Landscape of Escherichia Coli.” *Molecular Systems Biology* 10 (735).

Carvunis, Anne-Ruxandra, and Trey Ideker. 2014. “Siri of the Cell: What Biology Could Learn from the iPhone.” *Cell* 157 (3). Elsevier Inc.:534–38. https://doi.org/10.1016/j.cell.2014.03.009.

Chen, William W, Mario Niepel, and Peter K Sorger. 2010. “Classic and Contemporary Approaches to Modeling Biochemical Reactions.” *Genes & Development* 24:1861–75. https://doi.org/10.1101/gad.1945410.Freely.

Cherry, J Michael, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, et al. 2012. “Saccharomyces Genome Database: The Genomics Resource of Budding Yeast.” *Nucleic Acids Research* 40 (Database issue):D700-5. https://doi.org/10.1093/nar/gkr1029.

Collins, Sean R, Assen Roguev, and Nevan J Krogan. 2010. “Quantitative Genetic Interaction Mapping Using the E-MAP Approach.” *Methods in Enzymology* 470 (January):205–31. https://doi.org/10.1016/S0076-6879(10)70009-4.

Collins, Sean R, Maya Schuldiner, Nevan J Krogan, and Jonathan S Weissman. 2006. “A Strategy for Extracting and Analyzing Large-Scale Quantitative Epistatic Interaction Data.” *Genome Biology* 7 (7):R63. https://doi.org/10.1186/gb-2006-7-7-r63.

Costanzo, Michael, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, et al. 2010. “The Genetic Landscape of a Cell.” *Science (New York, N.Y.)* 327 (5964):425–31. https://doi.org/10.1126/science.1180823.

Deisboeck, Thomas S, Zhihui Wang, Paul Macklin, and Vittorio Cristini. 2011. “Multiscale Cancer Modeling.” *Annual Review of Biomedical Engineering* 13 (August):127–55. https://doi.org/10.1146/annurev-bioeng-071910-124729.

Deutscher, David, Isaac Meilijson, Martin Kupiec, and Eytan Ruppin. 2006. “Multiple Knockout Analysis of Genetic Robustness in the Yeast Metabolic Network.” *Nature Genetics* 38 (9):993–98. https://doi.org/10.1038/ng1856.

Dowell, Robin D, Owen Ryan, An Jansen, Doris Cheung, Sudeep Agarwala, Timothy Danford, Douglas A Bernstein, et al. 2010. “Genotype to Phenotype : A Complex Problem.” *Science* 328:469.

Dutkowski, Janusz, Michael Kramer, Michal a Surma, Rama Balakrishnan, J Michael Cherry, Nevan J Krogan, and Trey Ideker. 2013. “A Gene Ontology Inferred from Molecular Networks.” *Nature Biotechnology* 31 (1):38–45. https://doi.org/10.1038/nbt.2463.

Eissing, Thomas, Lars Kuepfer, Corina Becker, Michael Block, Katrin Coboeken, Thomas Gaub, Linus Goerlitz, et al. 2011. “A Computational Systems Biology Software Platform for Multiscale Modeling and Simulation: Integrating Whole-Body Physiology, Disease Biology, and Molecular Reaction Networks.” *Frontiers in Physiology* 2 (January):1–10. https://doi.org/10.3389/fphys.2011.00004.

Formosa, Tim, Susan Ruone, Melissa D Adams, Aileen E Olsen, Peter Eriksson, Yaxin Yu, Alison R Rhoades, Paul D Kaufman, and David J Stillman. 2002. “On the Hir / Hpc Pathway : Polymerase Passage May Degrade Chromatin Structure.” *Genetics* 162:1557–71.

Gillis, Jesse, and Paul Pavlidis. 2012. “‘Guilt by Association’ is the Exception rather than the Rule in Gene Networks.” *PLoS Computational Biology* 8 (3):e1002444. https://doi.org/10.1371/journal.pcbi.1002444.

Gligorijević, Vladimir, Vuk Janjić, and Nataša Pržulj. 2014. “Integration of Molecular Network Data Reconstructs Gene Ontology.” *Bioinformatics (Oxford, England)* 30 (17):i594-600. https://doi.org/10.1093/bioinformatics/btu470.

Greene, Casey S, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene a Zelaya, Daniel S Himmelstein, Ran Zhang, et al. 2015. “Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.” *Nature Genetics* 47 (6). Nature Publishing Group:569–76. https://doi.org/10.1038/ng.3259.

Gutenkunst, Ryan N, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. 2007. “Universally Sloppy Parameter Sensitivities in Systems Biology Models.” *PLoS Computational Biology* 3 (10):1871–78. https://doi.org/10.1371/journal.pcbi.0030189.

Haber, James E., Hannes Braberg, Qiuqin Wu, Richard Alexander, Julian Haase, Colm Ryan, Zach Lipkin-Moore, et al. 2013. “Systematic Triple-Mutant Analysis Uncovers Functional Connectivity between Pathways Involved in Chromosome Regulation.” *Cell Reports* 3 (6). The Authors:2168–78. https://doi.org/10.1016/j.celrep.2013.05.007.

Hanahan, Douglas, and Robert a Weinberg. 2011. “Hallmarks of Cancer: The next Generation.” *Cell* 144 (5). Elsevier Inc.:646–74. https://doi.org/10.1016/j.cell.2011.02.013.

Hillenmeyer, Maureen E, Eula Fung, Jan Wildenhain, Sarah E Pierce, Shawn Hoon, William Lee, Michael Proctor, et al. 2008. “The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes.” *Science (New York, N.Y.)* 320 (5874):362–65. https://doi.org/10.1126/science.1150021.

Hofree, Matan, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. 2013. “Network-Based Stratification of Tumor Mutations.” *Nature Methods* 10 (11):1108–15. https://doi.org/10.1038/nmeth.2651.

Huntley, Rachael P, Tony Sawford, Maria J Martin, and Claire O’Donovan. 2014. “Understanding How and Why the Gene Ontology and Its Annotations Evolve: The GO within UniProt.” *GigaScience* 3 (1):4. https://doi.org/10.1186/2047-217X-3-4.

Ideker, Trey, and Nevan J Krogan. 2012. “Differential Network Biology.” *Molecular Systems Biology* 8 (565). Nature Publishing Group:565. https://doi.org/10.1038/msb.2011.99.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2014. “Data, Information, Knowledge and Principle: Back to Metabolism in KEGG.” *Nucleic Acids Research* 42 (Database issue):D199-205. https://doi.org/10.1093/nar/gkt1076.

Karr, Jonathan R., Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. 2012. “A Whole-Cell Computational Model Predicts Phenotype from Genotype.” *Cell* 150 (2):389–401. https://doi.org/10.1016/j.cell.2012.05.044.

Kelley, Ryan, and Trey Ideker. 2005. “Systematic Interpretation of Genetic Interactions Using Protein Networks.” *Nature Biotechnology* 23 (5):561–66. https://doi.org/10.1038/nbt1096.

Kelly, D E, D C Lamb, and S L Kelly. 2001. “Genome-Wide Generation of Yeast Gene Deletion Strains.” *Comparative and Functional Genomics* 2 (4):236–42. https://doi.org/10.1002/cfg.95.

Kim, Hanhae, Junha Shin, Eiru Kim, Hyojin Kim, Sohyun Hwang, Jung Eun Shim, and Insuk Lee. 2014. “YeastNet v3: A Public Database of Data-Specific and Integrated Functional Gene Networks for Saccharomyces Cerevisiae.” *Nucleic Acids Research* 42 (Database issue):D731-6. https://doi.org/10.1093/nar/gkt981.

Kim, Yoo-Ah, and Teresa M Przytycka. 2012. “Bridging the Gap between Genotype and Phenotype via Network Approaches.” *Frontiers in Genetics* 3 (May):227. https://doi.org/10.3389/fgene.2012.00227.

Kramer, Michael, Janusz Dutkowski, Michael Yu, Vineet Bafna, and Trey Ideker. 2014. “Inferring Gene Ontologies from Pairwise Similarity Data.” *Bioinformatics (Oxford, England)* 30 (12):i34-42. https://doi.org/10.1093/bioinformatics/btu282.

Lee, Anna Y, Robert P St Onge, Michael J Proctor, Iain M Wallace, Aaron H Nile, Paul A Spagnuolo, Yulia Jitkova, et al. 2014. “Mapping the Cellular Response to Small Molecules Using Chemogenomic Fitness Signatures.” *Science* 344:208–11.

Lee, Insuk, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte.

2011. “Prioritizing Candidate Disease Genes by Network-Based Boosting of Genome-Wide Association Data.” *Genome Research* 21 (7):1109–21. https://doi.org/10.1101/gr.118992.110.

Lee, Insuk, Ben Lehner, Tanya Vavouri, Junha Shin, Andrew G Fraser, and Edward M Marcotte. 2010. “Predicting Genetic Modifier Loci Using Functional Gene Networks,” 1143–53. https://doi.org/10.1101/gr.102749.109.et.

Lehner, Ben. 2013. “Genotype to Phenotype: Lessons from Model Organisms for Human Genetics.” *Nature Reviews. Genetics* 14 (3). Nature Publishing Group:168–78. https://doi.org/10.1038/nrg3404.

Leiserson, Mark D M, Diana Tatar, Lenore J Cowen, and Benjamin J Hescott. 2011. “Inferring Mechanisms of Compensation from E-MAP and SGA Data Using Local Search Algorithms for Max Cut.” *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 18 (11):1399–1409. https://doi.org/10.1089/cmb.2011.0191.

Leiserson, Mark D M, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, et al. 2014. “Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes.” *Nature Genetics* 47 (2). https://doi.org/10.1038/ng.3168.

Lerman, Joshua a, Daniel R Hyduke, Haythem Latif, Vasiliy a Portnoy, Nathan E Lewis, Jeffrey D Orth, Alexandra C Schrimpe-Rutledge, et al. 2012. “In Silico Method for Modelling Metabolism and Gene Product Expression at Genome Scale.” *Nature Communications* 3 (May). Nature Publishing Group:929. https://doi.org/10.1038/ncomms1928.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest” 2 (December):18–22.

Ma, Xiaotu, Aaron M Tarone, and Wenyuan Li. 2008. “Mapping Genetically Compensatory Pathways from Synthetic Lethal Interactions in Yeast.” *PloS One* 3 (4):e1922. https://doi.org/10.1371/journal.pone.0001922.

Machado, Daniel, Rafael S Costa, Miguel Rocha, Eugénio C Ferreira, Bruce Tidor, and Isabel Rocha. 2011. “Modeling Formalisms in Systems Biology.” *AMB Express* 1 (1). Springer Open Ltd:45. https://doi.org/10.1186/2191-0855-1-45.

Mackay, Trudy F C. 2014. “Epistasis and Quantitative Traits: Using Model Organisms to Study Gene-Gene Interactions.” *Nature Reviews. Genetics* 15 (1). Nature Publishing Group:22–33. https://doi.org/10.1038/nrg3627.

Mo, Monica L, Bernhard O Palsson, and Markus J Herrgård. 2009. “Connecting Extracellular Metabolomic Measurements to Intracellular Flux States in Yeast.” *BMC Systems Biology* 3 (January):37. https://doi.org/10.1186/1752-0509-3-37.

Nepusz, Gabor Csardi and Tamas. 2006. “The Igraph Software Package for Complex Network Research.” *InterJournal* Complex Sy:1695.

Ng, Sam, Eric a Collisson, Artem Sokolov, Theodore Goldstein, Abel Gonzalez-Perez, Nuria Lopez-Bigas, Christopher Benz, David Haussler, and Joshua M Stuart. 2012. “PARADIGM-SHIFT Predicts the Function of Mutations in Multiple Cancers Using Pathway Impact Analysis.” *Bioinformatics (Oxford, England)* 28 (18):i640–46. https://doi.org/10.1093/bioinformatics/bts402.

O’Brien, Edward J, Joshua a Lerman, Roger L Chang, Daniel R Hyduke, and Bernhard Ø Palsson. 2013. “Genome-Scale Models of Metabolism and Gene Expression Extend and Refine Growth Phenotype Prediction.” *Molecular Systems Biology* 9 (693):693. https://doi.org/10.1038/msb.2013.52.

Orth, Jeffrey D, Ines Thiele, and Bernhard Ø Palsson. 2010. “What Is Flux Balance Analysis?” *Nature Biotechnology* 28 (3). Nature Publishing Group:245–48. https://doi.org/10.1038/nbt.1614.

Pamblanco, Mercè, Paula Oliete-Calvo, Encar García-Oliver, M Luz Valero, Manuel M Sanchez del Pino, and Susana Rodríguez-Navarro. 2014. “Unveiling Novel Interactions of Histone Chaperone Asf1 Linked to TREX-2 Factors Sus1 and Thp1.” *Nucleus* 5 (3):247–59. https://doi.org/10.4161/nucl.29155.

Pandey, Gaurav, Bin Zhang, Aaron N Chang, Chad L Myers, Jun Zhu, Vipin Kumar, and Eric E Schadt. 2010. “An Integrative Multi-Network and Multi-Classifier Approach to Predict Genetic Interactions.” *PLoS Computational Biology* 6 (9). https://doi.org/10.1371/journal.pcbi.1000928.

Park, Yungki, and Edward M Marcotte. 2012. “Flaws in Evaluation Schemes for Pair-Input Computational Predictions.” *Nature Methods* 9 (12):1134–36. https://doi.org/10.1038/nmeth.2259.

Pe’er, Dana, and Nir Hacohen. 2011. “Principles and Strategies for Developing Network Models in Cancer.” *Cell* 144 (6). Elsevier Inc.:864–73. https://doi.org/10.1016/j.cell.2011.03.001.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn : Machine Learning in Python.” *Journal of Machine Learning Research* 12:2825–30.

Pesquita, Catia, Daniel Faria, André O Falcão, Phillip Lord, and Francisco M Couto. 2009. “Semantic Similarity in Biomedical Ontologies.” *PLoS Computational Biology* 5 (7):e1000443. https://doi.org/10.1371/journal.pcbi.1000443.

Pratt, Dexter, Jing Chen, David Welker, Ricardo Rivas, Rudolf Pillich, Vladimir Rynkov, Keiichiro Ono, et al. 2015. “NDEx, the Network Data Exchange.” *Cell Systems* 1 (4). Elsevier Inc.:302–5. https://doi.org/10.1016/j.cels.2015.10.001.

Qi, Yan, Yasir Suhail, Yu-yi Lin, Jef D Boeke, and Joel S Bader. 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 Research* 18:1991–2004. https://doi.org/10.1101/gr.077693.108.

Ramanan, Vijay K, Li Shen, Jason H Moore, and Andrew J Saykin. 2012. “Pathway Analysis of Genomic Data: Concepts, Methods, and Prospects for Future Development.” *Trends in Genetics : TIG* 28 (7). Elsevier Ltd:323–32. https://doi.org/10.1016/j.tig.2012.03.004.

Resnik, Philip. 1995. “Using Information Content to Evaluate Semantic Similarity in a Taxonomy.” In *Joint Conference on Artificial Intelligence*, 1:448–53.

Saito, Rintaro, Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Samad Lotia, Alexander R Pico, Gary D Bader, and Trey Ideker. 2012. “A Travel Guide to Cytoscape Plugins.” *Nature Methods* 9 (11):1069–76. https://doi.org/10.1038/nmeth.2212.

Schwabish, Marc a, and Kevin Struhl. 2006. “Asf1 Mediates Histone Eviction and Deposition during Elongation by RNA Polymerase II.” *Molecular Cell* 22 (3):415–22. https://doi.org/10.1016/j.molcel.2006.03.014.

Segrè, Daniel, Alexander Deluna, George M Church, and Roy Kishony. 2005. “Modular Epistasis in Yeast Metabolism.” *Nature Genetics* 37 (1):77–83. https://doi.org/10.1038/ng1489.

Skafidas, E, R Testa, D Zantomio, G Chana, I P Everall, and C Pantelis. 2014. “Predicting the Diagnosis of Autism Spectrum Disorder Using Gene Pathway Analysis.” *Molecular Psychiatry* 19 (4). Nature Publishing Group:504–10. https://doi.org/10.1038/mp.2012.126.

Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. “BioGRID: A General Repository for Interaction Datasets.” *Nucleic Acids Research* 34 (Database issue):D535-9. https://doi.org/10.1093/nar/gkj109.

Sullivan, Patrick F. 2012. “Puzzling over Schizophrenia: Schizophrenia as a Pathway Disease.” *Nature Medicine* 18 (2). Nature Publishing Group:210–11. https://doi.org/10.1038/nm.2670.

Szappanos, Balázs, Károly Kovács, Béla Szamecz, Frantisek Honti, Michael Costanzo, Anastasia Baryshnikova, Gabriel Gelius-Dietrich, et al. 2011. “An Integrated Approach to Characterize Genetic Interaction Networks in Yeast Metabolism.” *Nature Genetics* 43 (7). Nature Publishing Group:656–62. https://doi.org/10.1038/ng.846.

Szczurek, Ewa, Irit Gat-Viks, Jerzy Tiuryn, and Martin Vingron. 2009. “Elucidating

Regulatory Mechanisms Downstream of a Signaling Pathway Using Informative Experiments." *Molecular Systems Biology* 5 (287). Nature Publishing Group:287. https://doi.org/10.1038/msb.2009.45.

Takahashi, K., N. Ishikawa, Y. Sadamoto, H. Sasamoto, S. Ohta, a. Shiozawa, F. Miyoshi, Y. Naito, Y. Nakayama, and M. Tomita. 2003. "E-Cell 2: Multi-Platform E-Cell Simulation System." *Bioinformatics* 19 (13):1727–29. https://doi.org/10.1093/bioinformatics/btg221.

Tao, Ying, Lee Sam, Jianrong Li, Carol Friedman, and Yves a Lussier. 2007. "Information Theory Applied to the Sparse Gene Ontology Annotation Network to Predict Novel Gene Function." *Bioinformatics (Oxford, England)* 23 (13):i529-38. https://doi.org/10.1093/bioinformatics/btm195.

The Gene Ontology Consortium. 2014. "Gene Ontology Consortium: Going Forward." *Nucleic Acids Research* 43 (November 2014):1049–56. https://doi.org/10.1093/nar/gku1179.

Tomita, Masaru, Kenta Hashimoto, Kouichi Takahashi, Thomas Simon Shimizu, Yuri Matsuzaki, Fumihiko Miyoshi, Kanako Saito, et al. 1999. "E-CELL : Software Environment for Whole-Cell Simulation." *Bioinformatics* 15 (1):72–84.

Tong, Amy Hin Yan, and Charles Boone. 2006. "Synthetic Genetic Array Analysis in Saccharomyces Cerevisiae." Edited by W Xiao. *Methods in Molecular Bio* 313. Humana Press:171–91.

Ulitsky, Igor, Tomer Shlomi, Martin Kupiec, and Ron Shamir. 2008. "From E-MAPs to Module Maps: Dissecting Quantitative Genetic Interactions Using Physical Interactions." *Molecular Systems Biology* 4 (209):209. https://doi.org/10.1038/msb.2008.42.

Waddington, C H. 1942. "Canalization of Development and the Inheritance of Acquired Characters." *Nature* 150 (3811):563–65.

Walpole, Joseph, Jason a Papin, and Shayn M Peirce. 2013. "Multiscale Computational Models of Complex Biological Systems." *Annual Review of Biomedical Engineering* 15 (January):137–54. https://doi.org/10.1146/annurev-bioeng-071811-150104.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. "Analysing Biological Pathways in Genome-Wide Association Studies." *Nature Reviews. Genetics* 11 (12). Nature Publishing Group:843–54. https://doi.org/10.1038/nrg2884.

Willsey, A Jeremy, Stephan J Sanders, Mingfeng Li, Shan Dong, Andrew T Tebbenkamp, Rebecca a Muhle, Steven K Reilly, et al. 2013. "Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism." *Cell* 155 (5). Elsevier Inc.:997–1007. https://doi.org/10.1016/j.cell.2013.10.020.

Zuk, Or, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. 2012. "The Mystery of

Missing Heritability: Genetic Interactions Create Phantom Heritability.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (4):1193–98. https://doi.org/10.1073/pnas.1119675109.

Chapter 3: USING DEEP LEARNING TO MODEL THE HIERARCHICAL STRUCTURE AND FUNCTION OF A CELL

3.1 Abstract

Although artificial neural networks capture a variety of human functions, their internal structures are hard to interpret. In the life sciences, extensive knowledge of cell biology provides an opportunity to design 'visible' neural networks (VNNs) which couple the model's inner workings to those of real systems. Here we develop DeepCell, a VNN embedded in the hierarchical structure of 2526 subsystems comprising a eukaryotic cell (http://deep-cell.ucsd.edu/). Trained on 12 million genotypes, DeepCell simulates cellular growth nearly as accurately as laboratory observations. During simulation, genotypes induce patterns of subsystem activities, enabling *in-silico* investigations of the molecular mechanisms underlying each genotype-phenotype association. These mechanisms can be validated and many are unexpected; some are governed by Boolean logic. Cumulatively, 80% of the importance for growth prediction is captured by 484 subsystems (21%), reflecting the emergence of a complex phenotype. DeepCell provides a foundation for decoding the genetics of disease, drug resistance, and synthetic life.

3.2 Introduction

Deep learning techniques have revolutionized the field of artificial intelligence by enabling machines to perform human activities like seeing, listening and speaking (Farabet et al. 2013; Mikolov et al. 2011; Hinton et al. 2012; Sainath et al. 2013; Collobert et al. 2011; LeCun, Bengio, and Hinton 2015). Such systems are constructed from many-layered, 'deep', artificial neural networks (ANNs), inspired by actual neural networks in the brain and how they process patterns. The function of the ANN is created during a training phase, in which the model learns to capture as accurately as possible the correct answer, or output, that should be returned for each example input pattern. In this way, machine vision learns to recognize objects like dogs, people, and faces, and machine players learn to distinguish good from bad moves in games like chess and Go (Silver et al. 2016).

In modern ANN architectures, the connections between neurons as well as their strengths are subject to extensive mathematical optimization, leading to densely entangled network structures that are neither tied to an actual physical system nor based on human reasoning. Consequently, it is typically difficult to grasp how any particular set of neurons relates to system function. For instance, AlphaGo beats top human players (Silver et al. 2016), but examination of its underlying network yields little insight into the rules behind its moves or how these are encoded by neurons. These are so-called 'black boxes' (Brosin 1958), in which the input/output function accurately models an actual system but the internal structure does not (**Figure 3.1a**). Such models, while often useful, are insufficient in cases where simulation is needed not only of system function but also of system structure. In particular, many applications in biology and medicine seek to

model both functional outcome (e.g., cellular phenotypes or states of patient health and disease) and the mechanisms leading to that outcome so that these can be understood and manipulated through drugs, genes or environment.

Here we report DeepCell, an interpretable or 'visible' neural network (VNN) simulating a basic eukaryotic cell. The structure of this model is formulated from extensive prior knowledge of the cell's hierarchy of subsystems documented for the budding yeast *Saccharomyces cerevisiae*, drawn from either of two sources: the Gene Ontology (GO), a literature-curated reference database from which we extracted 2526 intracellular components, processes, and functions (The Gene Ontology Consortium 2016); or CliXO, an alternative ontology of similar size inferred from large-scale molecular datasets rather than literature curation (Dutkowski et al. 2013; Kramer et al. 2014). While CliXO and GO overlap in approximately 37% of subsystems, some in CliXO are apparent in large-scale datasets but not yet characterized in literature, whereas some in GO are documented in the literature but difficult to identify in big data. Subsystems in these ontologies are interrelated through hierarchical parent-child relationships of membership or containment. Such hierarchies form a natural bridge from variations in genotype, at the scale of nucleotides and genes, to variations in phenotype, at the scale of cells and organisms (Carvunis and Ideker 2014; Yu et al. 2016).

The function of DeepCell is learned during a training phase, in which perturbations to genes propagate through the hierarchy to impact parent subsystems that contain them, giving rise to functional changes in protein complexes, biological processes, organelles and, ultimately, a predicted response at the level of cell growth phenotype (**Figure 3.1b**). Previously, we saw that hierarchical groups of genes in an ontology could be used to

formulate very effective input features for such phenotypic predictions (Carvunis and Ideker 2014; Yu et al. 2016). However, the features were provided to standard black-box machine learning models for prediction which, while accurate, could not be interpreted biologically. Here, we use the biological hierarchy to directly embed the structure of a deep neural network, enabling more accurate prediction and transparent biological interpretation.

3.3 Results

DeepCell Design. In DeepCell, the functional state of each subsystem is represented by a bank of neurons (**Figure 3.1c**). Connectivity of these neurons is set to mirror the biological hierarchy, so that they take input only from neurons of child subsystems and send output only to neurons of parent (super)systems, with weights determined during training. The use of multiple neurons (ranging from 20 to 1,075 per system, see **Methods**) acknowledges that cellular components can be multifunctional, with distinct states able to adopt a range of values along multiple dimensions (Copley 2012). For example, interactions between subcomponents of the proteasome are thought to give rise to many pseudo-independent functional states (Hochstrasser 2016), while the mediator complex plays distinct functional roles related to transcriptional initiation, pausing, and elongation (Allen and Taatjes 2015). The input layer of the hierarchy comprises nodes representing genes, the most atomic subsystems in the model. The output layer, or root, is occupied by a single neuron representing the phenotype of the whole cell. By this design, the neural network model embedded in GO includes 43,721 neurons while the corresponding network for CliXO includes 22,167 neurons. The depth of both networks is 12 layers, on par with deep neural networks in other fields (Silver et al. 2016).

Training and Performance in Genotype-Phenotype Translation. Given this architecture, we next taught DeepCell to predict phenotypes related to cellular fitness, a model genotype-to-phenotype translation task. Extensive training was made possible by a recently published compendium of yeast growth phenotypes measured for nearly every

single and double gene deletion genotype, comprising over 12 million genotype-phenotype training examples (Costanzo et al. 2016). Each genotype was used to set the state of gene inputs to the model, indicating which had been genetically disrupted; each phenotype was encoded as the state of the root node. Two related phenotypes were considered: (i) capacity for growth measured by yeast colony size relative to wild-type cells for gene deletions; (ii) For double gene deletions, genetic interaction score measured as the difference in growth from that expected from independent measurements of the single gene deletions. Predicting genetic interaction represents a harder task than predicting absolute growth, as it requires that the model learn non-linear effects beyond the superposition of more elemental genotypes. Based on the training examples, the weights of input connections to each neuron were optimized by stochastic gradient descent computed by backpropagation (**Methods**). The DeepCell VNN can be executed and inspected through an interactive website we have created at http://deep-cell.ucsd.edu/ (**Figure 3.1d, Methods**).

We found that DeepCell was able to make accurate phenotypic predictions across the entire observed range (**Figures 3.2a,b**). The correlation between predicted and measured growth scores was very high for both structural hierarchies (Pearson's r = 0.94 by either GO or CliXO in four-fold cross validation; **Figure 3.2a**). The correlation between predicted and measured genetic interactions was also favorable (r = 0.51) given the greater experimental variability associated with this data type (r = 0.67 for replicate genetic interaction scores, versus r = 0.89 for replicate growth scores (Baryshnikova et al. 2010)). DeepCell outperformed previous predictors of genetic interactions, including those based on metabolic models (Szappanos et al. 2011) and protein-protein interaction

networks (I. Lee et al. 2010; Pandey et al. 2010) (average $r = 0.15$), as well as a previous hierarchical method not related to deep learning (**Figure 3.2c, Supplemental Figure S3.1**; $r = 0.35$) (Yu et al. 2016).

We also compared performance to 'black box' neural network configurations, to isolate the improved performance of deep learning methodology from contributions due to prior knowledge of biological structure. First, we constructed a series of ANNs with matching structure to DeepCell but permuting the assignment of genes to subsystems. The predictive performance of these matched structures decreased substantially (**Figure 3.2c**), becoming similar to the original only after the number of neurons was increased by an order of magnitude (**Figure 3.2d**). Thus, the biological hierarchy provides significant information not found in randomized versions. Next, we constructed a fully connected artificial neural network with the same number of layers and neurons as DeepCell but unlimited connectivity of neurons between adjacent layers. Despite these extra parameters, the performance of this fully connected model was not significantly better (**Figure 3.2c**).

Unlike standard ANNs, however, DeepCell's simulations were tied to an extensive hierarchy of internal biological subsystems with states that could be queried. This 'visible' aspect raised the possibility that these simulations might be useful for *in-silico* studies of biological mechanism. Thus, we next turned to DeepCell's ability to enable or accelerate four types of studies of central importance in the biological sciences:

1) Explaining a genotype-phenotype relationship
2) Prioritizing all important mechanisms in determination of phenotype overall
3) Study and classification of the genetic logic implemented by a process

4) Discovery of previously unknown biological processes and states

In the remainder of this work, we explore applications of DeepCell in each of these areas.

Explaining a Genotype-Phenotype Relationship. A fundamental challenge in genetics is to understand the molecular mechanisms underlying an observed relationship between genotype and phenotype. We reasoned that DeepCell should be able to generate such mechanistic explanations automatically, by attributing each genotype-phenotype relationship to a corresponding set of changes in cellular subsystems. For this purpose, a 'genotype-phenotype association' was defined as a specific change in genotype, relative to wild type, that leads to a significant alteration in growth phenotype. Reducing the state of each subsystem (complete set of neuron outputs) to the top two principal components, we then identified subsystems for which this reduced state was significantly altered during DeepCell simulation. Such subsystems were proposed as candidate explanations in translation of genotype to phenotype, whereas those without state changes – typically the vast majority – were excluded from further consideration. In this way, DeepCell could be used to test alternative hypotheses for the subsystems driving a genotype-phenotype relationship, out of potentially many that might be proposed based on current knowledge.

As a case study of this capability, we examined the severe growth defect caused by genotype *pmt1*Δ*ire1*Δ, disrupting the genes *PMT1* and *IRE1*. We observed a total of 243 subsystems positioned "above" *PMT1* or *IRE1* in the DeepCell hierarchy (ancestors of one or both genes), providing candidate explanations for the genotype-phenotype relationship. For instance, *PMT1* is an enzyme that facilitates O-mannosylation, a reaction

that terminates aberrantly folded proteins in the Endoplasmic Reticulum Unfolded Protein Response (ER-UPR) (Xu et al. 2013) and, separately, promotes the creation of mature glycoproteins in the cell wall (Free 2013). *IRE1* is an ER transmembrane sensor that detects ER accumulation of unfolded proteins (Walter and Ron 2011) and is also required for cell wall organization by maintaining cell wall integrity (Scrimale et al. 2009). Therefore, we sought to use the DeepCell simulation to determine which of the candidate subsystems were most responsible for the observed growth defect.

Examining the states of the 243 candidate subsystems in DeepCell, we found that the state of ER-UPR was substantially decreased in the *pmt1*Δ*ire1*Δ genotype compared to wild type, whereas the states of cell wall organization and other subsystems were less affected (**Figure 3.3a**). To validate this predicted decrease in ER-UPR activity, we examined a dataset monitoring this process by Green Fluorescent Protein (GFP) driven by a promoter responsive to Hac1, a key transcriptional activator of ER-UPR, over numerous pairwise gene disruptions (Jonikas et al. 2009). Hac1 activity was significantly lowered in the *pmt1*Δ*ire1*Δ genotype compared to wild type, consistent with model simulations (**Figure 3.3b**). Model inspection suggested that the difference between ER-UPR and other subsystems had been learned by DeepCell during training, in which genotypes disrupting ER-UPR had more than twice the genetic interaction score on average (−0.34) than genotypes disrupting, e.g., cell wall organization (−0.16). Beyond this single genotype, we noted that Hac1 activity had also been measured for 16 other genotypes disrupting genes in ER-UPR. Over all of these, we found that the simulated state of the subsystem was well correlated with the experimental Hac1 activity (**Figure 3.3b**). To address the concern that Hac1 activity might associate non-specifically with

state changes in many diverse subsystems, not just those related to ER, we examined the correlation between Hac1 activity and the simulated states of every subsystem in DeepCell addressed by this dataset (subsystems containing genes disrupted in at least ten genotypes with Hac1-GFP measurements). High correlation was observed only for ER-UPR and its super-systems (**Figure 3.3c**) demonstrating specific validation. These results show how DeepCell can be used to generate and test specific hypotheses about the cellular mechanisms leading from genotype to phenotype.

In explaining the above genotype-phenotype relationship, a key requirement was that the state of a subsystem simulated *in silico* recapitulate its true biological state observed *in vivo*. To explore this capability further, we also examined the subsystem of DNA repair (**Figure 3.3d**) which, like ER-UPR, had been previously interrogated over many double gene deletion genotypes (Srivas et al. 2013). In particular, DNA repair status had been characterized by cellular resistance to ultraviolet radiation (UV), which causes DNA damage by pyrimidine dimers and double stranded breaks (Cadet, Sage, and Douki 2005). Once again we saw good agreement between experiments and simulations: the overall pattern of UV resistance among genotypes impacting DNA repair genes significantly tracked model predictions (**Figure 3.3e**), in a manner that was highly specific to DNA repair and its supersystems (**Figure 3.3f**). These results provided a second demonstration that the simulated states of a subsystem can reflect true biological states in a manner that can be used to generate mechanistic hypotheses.

Prioritizing all important systems in determination of phenotype overall. Beyond individual explanations, a critical question was whether a complex phenotype such as

growth depends on equal contributions from many subsystems or is dominated by a few. To address this question, we set out to systematically characterize the importance of every subsystem in determining growth phenotype. We reasoned that the overall importance of a biological system in genotype-phenotype translation is related to the degree to which its state is more predictive of growth phenotype than the states of its children. Based on this reasoning, we developed a score called Relative Local Improvement in Predictive Power (RLIPP score, **Methods**) which we calculated for each subsystem over all genotypes that disrupt it. Positive RLIPP scores correspond to an increase in predictive power for that subsystem relative to its children, while negative scores correspond to a decrease in power (e.g. RLIPP = 1 corresponds to a 100% increase).

We observed that RLIPP approximately followed a Pareto (power-law) distribution, in which a few subsystems were highly important for the predictive power of the model, with a long tail of weakly important systems (**Figure 3.4a**). In particular, 80% of the cumulative importance was captured by 21% of subsystems (the Pareto 80/20 rule (Pareto and Page 1971)), while >88% of subsystems retained at least some improvement in phenotypic prediction over their children (RLIPP > 0). The subsystem of greatest individual importance was ‘Negative regulation of cellular macromolecule biosynthesis’, which organizes cellular circuits that inhibit biosynthesis and, as evidenced by DeepCell simulations, can lead to strong increases in growth when disrupted. Other subsystems important for growth related to the proper function of organelles, biomolecular transport, stress response, and protein modification and assembly of complexes, all of which are consistent with a general understanding of cell biology (**Figure 3.4b-j**). We also examined

the importance of subsystems in the DeepCell model trained on the alternative hierarchy inferred directly from data (CliXO). While several important subsystems were the same as GO's (e.g., transport, stress response, **Supplemental Figure S3.2c-j**), others were newly discovered subsystems worthy of future examination (see **Supplemental Figure S3.2b** and below).

Study and classification of the functional genetic logic implemented by a process. Another type of model interpretation is to study the mathematical functions by which subsystems functionally integrate information from their inputs. We investigated whether these functions can be reduced to simple forms, such as Boolean logic gates, which are easily interpreted (**Methods**). Indeed, this analysis found 1119 subsystems were at least partly governed by Boolean logic (44% of GO, **Supplemental Table S3.2**). For instance, the principal activity of Mitochondrial Respiratory Chain (**Figure 3.5a**), while relatively high in wild-type cells, was driven low by disruptions in any of its several enzymatic complexes involved in electron transport, such as complexes III or IV (**Figure 3.5b**). With regard to these components, the state of Mitochondrial Respiratory Chain resembles a logical AND gate (**Figure 3.5c**). We also observed many cases of OR, XOR, and (A not B), although the AND configuration arose most frequently (**Supplemental Table S3.2**). Logical AND gates capture epistasis between components, in which both are necessary for function, for example two subunits of a protein complex or sequential steps of a signaling cascade. OR gates capture functional complementarity or redundancy, in which one component sufficient for function can be buffered or substituted by another. The

remaining 66% of GO subsystems did not map clearly to Boolean logic functions, suggesting molecular machinery that is more complex than a simple on/off switch.

Discovery of previously unknown biological processes and states. Finally, since DeepCell's hierarchy could be structured from systematic datasets as an alternative to literature (CliXO versus GO), we investigated the extent to which model simulations relied on entirely new cellular subsystems not previously appreciated in biology. As one such example we focused on CliXO:10651, a previously undocumented process which ranked among the top ten systems important for growth prediction according to RLIPP score (**Supplemental Table S3.1**). We found that CliXO had inferred this system based on the elevated density of protein-protein interactions observed among its 154 genes (**Figure 3.5d**, 9-fold enrichment, $p<10^{-200}$). Many of these interactions were found to interconnect two subsystems which were much better understood: Actin filament-based process (GO:0030029) and Cellular monovalent inorganic cation homeostasis (GO:0030004) (5-fold enrichment between subsystems, p=0.00029). Further functional analysis indicated that the state of CliXO:10651 is governed approximately by a Boolean AND of the states of its two subsystems (**Figure 3.5d**), i.e. both are required to maintain a wild-type status. These findings were supported by previous reports that the homeostasis of ions, such as iron, regulates the level of oxidative stress, which in turn disrupts actin cytoskeletal organization (Farrugia and Balzan 2012; Pujol-Carrion and de la Torre-Ruiz 2010). Hence, we assigned the provisional name "Maintenance of actin cytoskeletal organization by ion homeostasis".

A second undocumented system we considered was CliXO:10582, consisting of 71 genes (**Figure 3.6a**). Although the majority of these genes had known functional roles

in DNA repair, nothing like this distinct grouping had been previously recognized. Examination of the surrounding model structure revealed that CliXO:10582 integrates components of three known DNA repair subsystems -- postreplication repair, mismatch repair, and non-recombinational repair -- along with a collection of previously unrelated genes. Such organization was inferred by CliXO based on the extreme density of protein-protein interactions falling within each of the three subsystems (**Figure 3.6a**, 41 to 90-fold enrichment, $p<10^{-200}$) and, unexpectedly, lower but still very significant densities of protein-protein interactions spanning each pair of subsystems (average 15-fold enrichment, $p<10^{-200}$). Revisiting the experimental data on resistance to UV-induced DNA damage (Srivas et al. 2013) (**Figures 3.3e,f**), we saw that the simulated state of CliXO:10582 was strongly correlated with UV resistance over a large number of genotypes (**Figure 3.6b**). This association was stronger than for the state of any single child and, in fact, for any other CliXO or GO subsystem interrogated by the experimental data (**Figure 3.6c**). Mathematically, the state of CLiXO:10582 (along with the UV resistance phenotype) was well-approximated by a weighted linear summation of the activities of the three child processes, with postreplication repair having the greatest single contribution (**Figure 3.6d**). CliXO:10582 thus presents as a physically interacting collection of subcomponents which specifically coordinate the response to UV damage, leading us to assign the provisional name “UV damage coordination supercomplex (UVCS)”. For the eight genes in this subsystem newly assigned roles in specific DNA repair pathways (e.g. *LGE1*, *BRE1*, and *TEL2*, green nodes in **Figure 3.6a**), the evidence summarized by the model -- that these gene products physically interact in a cluster of known DNA repair factors, and that they also functionally manifest with the same UV

sensitivity phenotype when disrupted -- creates a compelling case for further biological studies.

Beyond these vignettes, we found that a total of 236 subsystems in the CliXO hierarchy, previously undocumented in GO, were assigned high importance scores for genotype-phenotype translation (RLIPP scores, **Supplemental Table S3.1**). Of these, we noted 28 subsystems with states governed by Boolean logic (see above analysis), which may facilitate their further study. All of these systems correspond to potentially new cellular components or processes with a definitive role in determination of growth.

3.4 Discussion

A direct route to interpretable neural networks is to encode not only function but form. Here, we have explored such visible learning in the context of cell biology, by incorporating an unprecedented collection of knowledge (Dutkowski et al. 2013; Kramer et al. 2014; Gene Ontology Consortium 2015) and data (Kim et al. 2014; Costanzo et al. 2016, 2010) to simultaneously simulate the hierarchical structure and function of cells. DeepCell captured nearly all of the phenotypic variation observed in cellular growth, a classic complex phenotype, including much of the less-understood non-additive portion due to genetic interactions (**Figures 3.2a-c**). Armed with this explanatory power, the model was able to simulate the intermediate functional states of thousands of cellular subsystems. Knowledge of these states enabled a wide variety of *in-silico* studies of biological mechanism, including the dissection of cellular subsystems important in determining the growth phenotype, the identification of new subsystems, and the reduction of subsystem functions, where possible, to simple mathematical relations such as Boolean logic (**Figures 3.3-3.5**).

Conceptually, the approach is to work towards a synthesis of statistical genetics and systems biology. State-of-the-art methods in statistical genetics (Yang et al. 2015, 2014) are based on linear regression of phenotype against the independent effects of genetic polymorphisms, without modeling the underlying biological mechanisms that give rise to nonlinearity and genetic interaction. Separately, studies in systems biology capture biological mechanisms using a spectrum of mathematical models (e.g. differential equations (Chen, Niepel, and Sorger 2010), flux-balance analysis (Szappanos et al. 2011), guilt-by-association (I. Lee et al. 2010)), but such models typically do not have the

breadth for large-scale genetic dissection of phenotype. DeepCell bridges these two avenues. Its neural network encodes a complex nonlinear regression, an extension of statistical genetics in which the additional complexity is supported and made feasible by a hierarchical model of biological mechanism, an extension of systems biology.

In comparison to some other mechanistic models that have attempted large-scale genotype-phenotype prediction (Yu et al. 2016; Karr et al. 2012), the framework of hierarchical neural networks is very general and expressive, such that a very large class of biological structures and functions can be learned and represented. For example, our earlier approach (Yu et al. 2016) used hierarchical knowledge of subsystems to create new features based on the number of gene perturbations in each subsystem. These subsystem-level features were then used as input to a black-box machine learning model for prediction of phenotype. This approach, however, had limited expressivity for two reasons. First, these features are predetermined before modeling and thus cannot learn the true state of each subsystem. Second, as with other black boxes, it is challenging to interpret the internal logic combining these features to predict phenotype. We found that the high expressivity of DeepCell improves phenotype prediction (**Figure 3.2**) while revealing biological functions and states (**Figures 3.3-3.5**) whose breadth falls outside the realm of previous models.

It is also instructive to view DeepCell in context of previous research in interpretable machine learning, in which the notion of interpretability has been defined in different ways (Lipton 2016). One direction has been to perform a post-hoc examination of a machine learning model that has already been trained, typically involving visualization of internal (hidden) neurons and creation of plausible explanations for their decisions. For

example, a neural network trained to identify images of dogs might, upon visualization of the internal wiring, be seen to have neurons that capture general properties of dogs, like “tail” or “furry” (Mahendran and Vedaldi 2015; Vondrick et al. 2013; Weinzaepfel, Jégou, and Pérez 2011).

A fundamental limitation of post-hoc interpretation is that the training process is independent of the interpretation process, leaving no guarantees as to what level of human-understandable explanation can be achieved (Chakraborty et al. 2017). For this reason, a second direction has been to construct the machine learning model in a manner that enables direct human interpretation. In attention-based neural networks (Bahdanau, Cho, and Bengio 2014; Lei, Barzilay, and Jaakkola 2016), a separate module is designed to select key input features important for prediction, prior to the main black-box learning machinery. This prior feature selection step can provide a type of interpretation: For example, in building a model to predict the emotional attitude of a blog author (positive or negative, angry or calm, and so on), the “interpretation” might be a key feature within a page of text (LOL, I’m so upset) that justifies the reason for the prediction.

DeepCell extends this direct approach. To enable human interpretation, the structure of the network model is designed to reflect the inner structure of the prediction task. In contrast to the previous attention-based approaches, however, the hierarchical structure allows us to identify feature clusters at multiple scales, pushing the interpretation from the input feature level to internal features representing biological subsystems.

In two case studies, involving genotypes impacting ER-UPR and DNA repair subsystems, the subsystem states learned by DeepCell could be directly confirmed by molecular measurements. Notably, such learning took place without exposure to any

information about subsystem states during training. Rather, these states emerged while learning to translate genotypes (model inputs) to growth phenotypes (model outputs) under the structural constraints of the subsystem hierarchy; together, the input/output data and hierarchical structure were sufficient to guide the neurons of each subsystem to learn a biologically correct function. In future, one might supervise a VNN to learn (potentially multiple) subsystem states and/or complex phenotypes directly, in which case training data could be provided at any level: genotype, phenotype, or points in between.

In supervised machine learning, it is often possible to identify many alternative model structures, all of which have very good predictive performance. In biology, however, a key question is which of the many predictive structures is the one actually used by the biological system, optimized by evolution and natural selection. With these principles in mind, the explorations here provide proof-of-concept for potential extensions to many disciplines as enough relevant structural and functional measurements become available. Such models are of immediate interest in genome-wide association studies of human disease (Visscher et al. 2012), in which different patient genotypes can influence disease outcomes by diverse and complex mechanisms currently hidden from black-box statistical approaches. Once trained on sufficient patient data, these models might have application in personalized therapy by analyzing a patient's genotype in combination with potential points of intervention targeted by drugs. We also see compelling uses in design of synthetic organisms, in which candidate genotypes can be efficiently evaluated *in silico* prior to validation of these designs *in vivo*. Finally, beyond the structure/function architecture of the cell, biological systems at other scales may benefit from this type of constrained learning, including modeling of neural connections in the brain.

3.5 Methods

Preparation of Ontologies. We guided the deep neural network structure using a biological ontology, consisting of terms representing cellular subsystems, child-parent relations representing containment of one term by another, and gene-to-term annotations. The first ontology considered was the Gene Ontology (GO), in which all three branches of GO (biological process, cellular component, and molecular function) were joined under a single root. We used the following criteria to filter (remove) terms from GO:

1. Terms with the evidence code ‘‘inferred by genetic interaction’’ (IGI), to avoid potential circularity in predicting genetic interactions in the genotype-phenotype samples.
2. Terms containing fewer than six yeast genes disrupted in the available genotypes (with “containment” defined as all genes annotated to that term or its descendants).
3. Terms that are redundant with respect to their children terms in the ontology.

When a term was removed, all children were connected directly to all parent terms to maintain the hierarchical structure. The remaining 2526 terms were used to define the hierarchy of DeepCell subsystems.

To complement the GO structure, we also constructed a data-driven gene ontology using the method of Clique Extracted Ontologies (CliXO) as previously described (Kramer et al. 2014). Briefly, data on gene pairs were sourced from YeastNet v3 (Kim et al. 2014), which lists 68 experimental studies of 8 data types, excluding genetic interactions to avoid circularity similar to criterion 1 above. All features were integrated to create a single gene-gene similarity network following a previously described procedure (Dutkowski et al. 2013), in which each gene-gene pair is assigned a weighted similarity based on a

combination of the YeastNet data. This network was subsequently analyzed with the CliXO algorithm, which identifies nested cliques as the threshold gene-gene similarity becomes progressively less stringent. This process yields a hierarchy (directed acyclic graph) of parent-child relations among cliques at different similarity thresholds.

DeepCell Architecture and Training Algorithm. DeepCell trains a deep neural network to predict phenotype from genotype, with architecture that exactly mirrors the hierarchical structure of an ontology of cellular subsystems. Each cellular subsystem is represented by a group of hidden variables (neurons) in the neural network, and each parent-child relation is represented by a set of edges that fully connect these groups of hidden variables. The depth of this architecture (12 layers) presents two challenges for training: 1) There is no guarantee that each subsystem will learn new patterns instead of copying those of its child subsystems; 2) Gradients tend to vanish lower in the hierarchy. To tackle these challenges, we borrow ideas from two previous systems, GoogLeNet (Szegedy et al., n.d.) and Deeply-Supervised Net (C.-Y. Lee et al. 2015), which improve the transparency and discriminative power of hidden variables and reduce the effect of vanishing gradients.

We denote our input training dataset as $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\}$, where N is the number of samples. For each sample i, $X_i \in R^M$ denotes the genotype, represented as a binary vector of states on M genes (1 = disrupted; 0 = wild type), and $y_i \in R$ denotes the observed phenotype, which can be either relative growth rate or genetic interaction value. The multi-dimensional state of each subsystem t, denoted by

the output vector $O_i^{(t)}$, is defined by a nonlinear function of the states of all of its child subsystems and annotated genes, concatenated in the input vector $I_i^{(t)}$:

$$O_i^{(t)} = BatchNorm\,(\,Tanh\,(\,Linear\,(\,I_i^{(t)}\,)\,)\,) \quad (1)$$

$Linear(\,I_i^{(t)}\,)$ is a linear transformation of $I_i^{(t)}$ defined as $W^{(t)}I_i^{(t)} + b^{(t)}$. Let $L_O^{(t)}$ denote the length of $O_i^{(t)}$, representing the number of values in the state of t and determined by:

$$L_O^{(t)} = max\,(\,20, \lceil 0.3 * number\ of\ genes\ contained\ by\ t \rceil\,) \quad (2)$$

Intuitively, larger subsystems have larger state vectors to capture potentially more complex biological responses. Similarly, let $L_I^{(t)}$denote the length of $I_i^{(t)}$. In Eqn. (1), $W^{(t)}$ is a weight matrix with dimensions $L_O^{(t)} \times L_I^{(t)}$ and $b^{(t)}$ is a column vector with size $L_O^{(t)}$. $W^{(t)}$and $b^{(t)}$ provide the parameters to be learned for subsystem t. $Tanh$ is the nonlinear transforming hyperbolic tangent function. $BatchNorm$ (Ioffe and Szegedy 2015) is a normalizing function that reduces the impact of internal covariate shift caused by different scales of weights in $W^{(t)}$. Batch normalization can be viewed as a type of regularization of model weights and reduces the need for the traditional dropout step in deep learning. We perform the training process by minimizing the objective function:

$$\frac{1}{N}\sum_{i=1}^{N}(Loss(Linear(O_i^{(r)}), y_i) + \alpha \sum_{t \neq r} Loss(Linear(O_i^{(t)}), y_i)) + \lambda||W||_2 \quad (3)$$

Here, $Loss$ is the squared error loss function, and r is the root of the hierarchy. Note that we compare y_i with not only the root's output, $O_i^{(r)}$, but also the outputs of all other subsystems, $O_i^{(t)}$. $Linear$ in (3) denotes linear functions transforming multi-dimensional vector $O_i^{(t)}$ into a scalar. In this way, every subsystem is optimized to serve its parents as features and to predict the phenotype itself, as used previously by GoogLeNet (Szegedy

et al., n.d.); the parameter α (=0.3) balances these two contributions. λ is a L2-norm regularization factor determined by four-fold cross validation. To train the DeepCell model, we initialize all weights uniformly at random between −0.001 and 0.001. We optimize the objective function using ADAM (Kingma and Ba 2014), a popular stochastic gradient descent algorithm, with mini-batch size of 15,000. Gradients with respect to model parameters are computed by standard back-propagation (Rumelhart, Hinton, and Williams 1988). Note that while other hyperparameters might influence the overall predictive performance, they are unrelated to our focus on biological interpretation as long as the same settings are applied to both DeepCell and the black-box models we use as controls (**Figure 3.2d**). We implemented DeepCell using the Torch7 library (https://github.com/torch/torch7) on Tesla K20 GPUs.

Alternative Genotype-Phenotype Translation Methods. We compared DeepCell to three state-of-the-art non-hierarchical approaches for predicting genetic interactions: flux balance analysis (FBA) (Szappanos et al. 2011), multi-network multi-classifier (MNMC) (Pandey et al. 2010), and guilt-by-association (GBA) (I. Lee et al. 2010). FBA uses a model of metabolism to assess the impact on cell growth of gene deletions in metabolic pathways. MNMC is an ensemble supervised learning system that uses many different datasets as features to predict genetic interactions. GBA predicts the genetic interaction score of pairwise gene deletions based on the phenotypes of their network neighbors. We also compared against our previous prediction method (Ontotype) (Yu et al. 2016) which applies prior knowledge from a hierarchy like GO or CLiXO but does not use deep learning nor simulate the internal states of subsystems. Ontotype counts the number of

genes knocked in every GO term and uses these counts as features in a random forest regression. To compare against these alternative models, we trained on a common set of ~3 million genetic interactions (Costanzo et al. 2010), as evaluated in the Ontotype study (Yu et al. 2016), and determined the neural network hyper-parameters by 4-fold cross validation.

Relative Local Improvement in Predictive Power (RLIPP). The RLIPP score was used to quantify and compare the importance of DeepCell's internal subsystems in prediction of phenotype. To calculate the RLIPP score of a subsystem, we compared two different linear models for phenotypic prediction. In the first model, the subsystem's neurons were used as features in a L2-norm penalized linear regression (**Supplemental Figure S3.3a**). In the second model, the neurons of the subsystem's children were used as the features instead. Each model was trained separately, with the optimal hyper-parameter associated with the L2-norm penalty determined in five-fold cross validation. The performance of each of these two models was calculated as the Spearman correlation between the predicted and measured phenotype, here taken as genetic interaction scores (**Supplemental Figure S3.3b,c**). The RLIPP score was defined as the performance of the parent model relative to that of the children (**Supplemental Figure S3.3d**). A positive RLIPP score indicates that the state of the parent subsystem is more predictive of phenotype than the states of its children. This situation can occur when the parent learns complex (nonlinear) patterns from the children, as opposed to merely copying or adding their values. The intuition behind the RLIPP score is similar to a related

'linear probe' technique developed in a previous study to characterize the utility of each layer of a deep neural network (Alain and Bengio 2016).

Identification of subsystems that mimic Boolean logic gates. As one means to interpret the mechanisms by which DeepCell translates genotype to phenotype, we evaluated each subsystem for the extent to which it approximates Boolean logic. In particular, we considered all trios of subsystems, each consisting of a parent subsystem and two of its children, and tested whether their binary states (S, C_1, C_2) were well-approximated by non-trivial Boolean logic (**Supplemental Table S3.2**). For each genotype, the binary state of each child subsystem was defined as either 'Wild Type' (True) or 'Disrupted' (False), by comparing PC1 to the wild-type state. The binary state of each parent subsystem was defined as either '≤Wild Type' (True) or '>Wild Type' (False), by comparing PC1 to the wild-type state. For each combinatorial state (C_1, C_2) of two child subsystems, the parent state S implied by DeepCell was determined based on the majority parent states of genotypes annotated to (C_1, C_2). For instance, suppose that for all the genotypes that induce (C_1=True, C_2=False) in the two children, DeepCell transforms 80% to parent state S=True and 20% to state S=False. We conclude the underlying logic for the parent subsystem to translate the signal from children subsystems is (True, False)→True. By checking the parent states for all four possible (C_1, C_2) combinations, we can decide whether this trio of subsystems exhibits Boolean logic (**Supplemental Table S3.2**). A trio belongs to none of the logic functions if > 50% of all the genotypes or <4 genotypes are annotated to any (C_1, C_2) combinatorial state; or none of the annotated genotypes yield significant genetic interactions ($|\varepsilon|$≤0.08). For those

subsystems exhibiting Boolean logic, we excluded ‘trivial’ functions in which the parent is always True, always False, or follows one of the children without dependence on the other.

DeepCell server construction. The DeepCell server (http://www.deep-cell.ucsd.edu/) comprises several interconnected components working in unison to collect user input, run simulations, and transcode results to the web interface. On the backend, the DeepCell neural network model runs on the Torch library (Collobert, Kavukcuoglu, and Farabet 2011) on a dedicated multi-GPU machine. On the front end, the web interface is built on cytoscape.js (Franz et al. 2016) to display the full biological hierarchy, an in-house D3 (Bostock, Ogievetsky, and Heer 2011) graph visualizer to display a subgraph of the hierarchy, and React (Stefanov 2016) for agile DOM (Document Object Model) editing (Marini 2002). To respond to user input, including searching and viewing details of model subsystems, a low-latency proxy service translates between plain text fetched from the front end and binary data used by the backend. An Elasticsearch cluster (Gormley and Tong 2015) caches and indexes data for fast lookup and predictions. All web services run on a Kubernetes-based cloud infrastructure (“Kubernetes” 2017) that auto-scales to heavy workloads. The result of these efforts allows easy visualization and interactivity of the model.

3.6 Figures

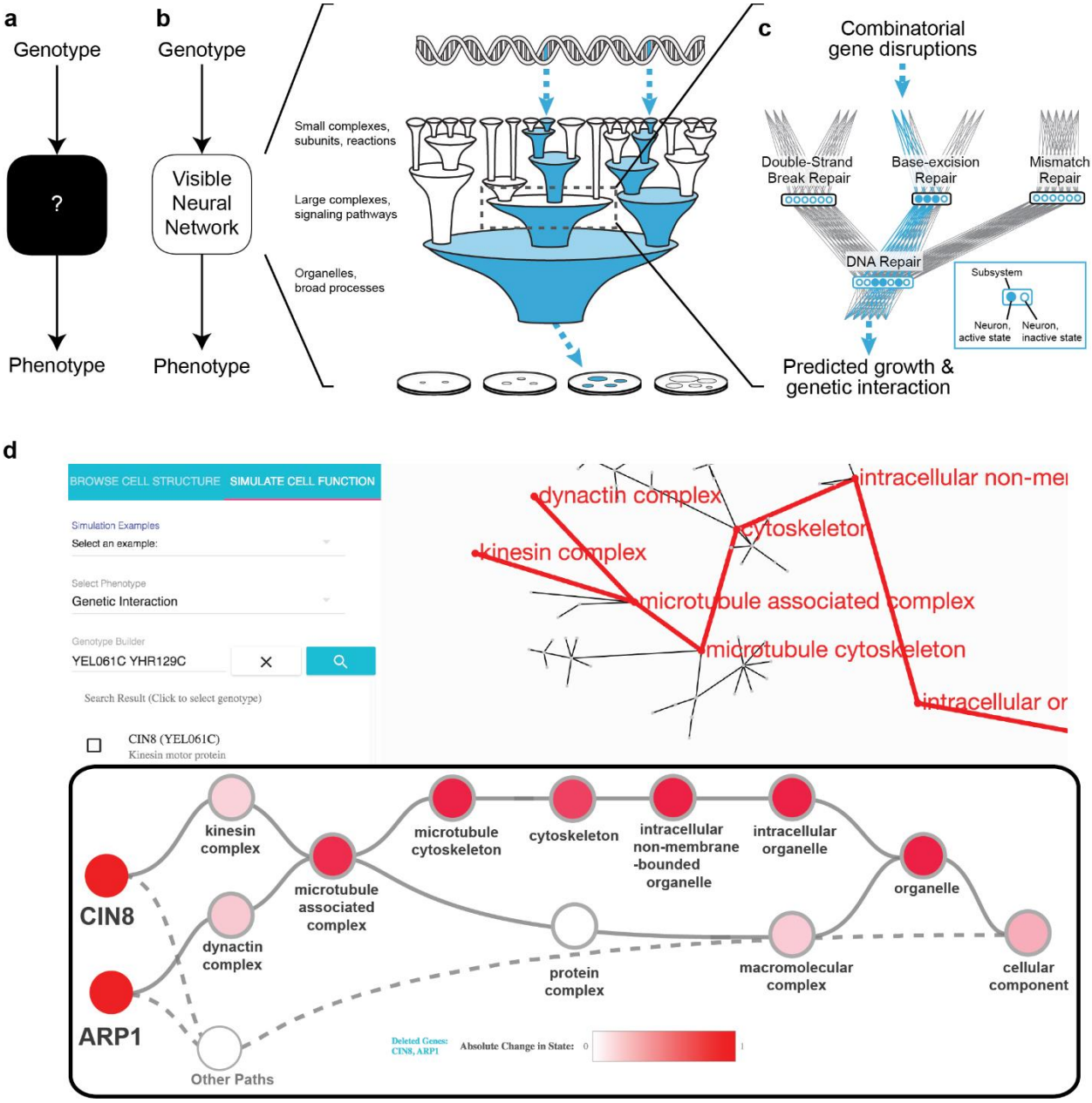
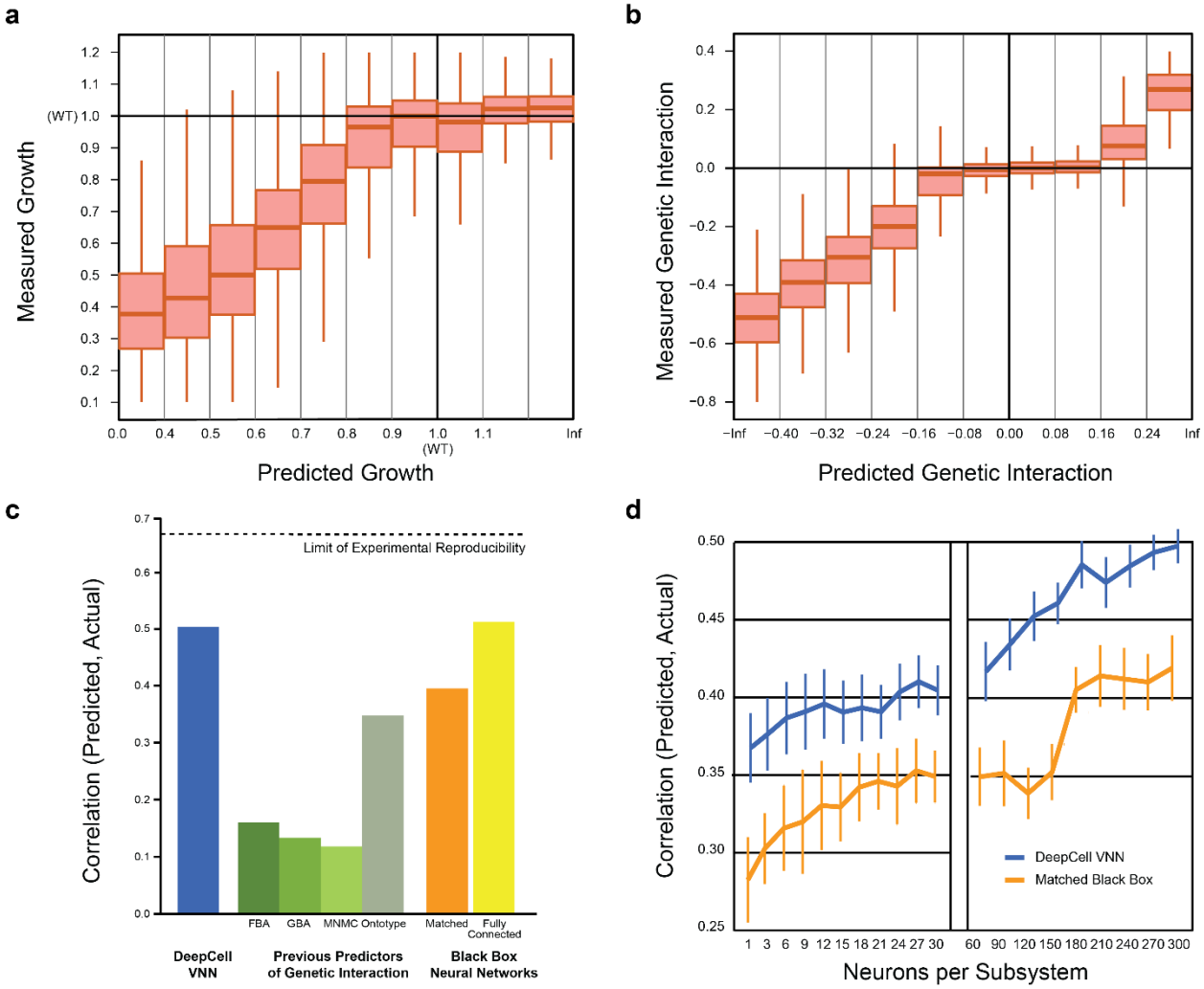


Figure 3.1. Modeling system structure and function with visible learning. a, A conventional neural network translates input to output as a black box without knowledge of system structure. **b**, In a visible neural network, input/output translation is based on prior structural knowledge. In DeepCell, gene disruption genotypes (top) are translated to cell growth predictions (bottom) through a hierarchy of cell subsystems (middle). **c**, A neural network is embedded in the prior structure using multiple neurons per subsystem. **d**, Screen capture of DeepCell online service, microtubule subsystems.

Figure 3.2. Prediction of cell viability and genetic interaction phenotypes. a, Measured versus predicted cell viability relative to wild type (WT = 1) on the Costanzo et al (Costanzo et al. 2010) dataset. **b**, Measured versus predicted genetic interaction scores for each double gene disruption genotype; genetic interactions between the disrupted genes can be positive (epistasis), zero (non-interaction), or negative (synthetic sickness or lethality). **c**, Model performance expressed as the correlation between measured and predicted genetic interaction scores. Performance of DeepCell (first bar, blue) is compared to previous methods for predicting genetic interactions (second four bars, green): FBA, Flux Balance Analysis (Szappanos et al. 2011); GBA, Guilt By Association (I. Lee et al. 2010); MNMC, Multi-Network Multi-Classifier (Pandey et al. 2010); Ontotype refers to our previous hierarchical approach not based on deep learning (Yu et al. 2016). Performance is also shown for matched structures in which gene-to-subsystem mappings are randomly permuted (orange bar, average of 10 randomizations) or for fully-connected neural networks with the same number of layers as DeepCell (final bar, yellow). Correlations were calculated across gene pairs that meet at an interaction significance criterion of $p<0.05$ in the dataset (Costanzo et al. 2010). **d**, Predictive performance as the number of neurons per subsystem increases from 1 to 300. The performance measure and the two structural hierarchies (DeepCell, blue; Matched Black Box, orange) are as in (c).


a
Measured Growth
Predicted Growth
(WT)
b
Measured Genetic Interaction
Predicted Genetic Interaction
c
Correlation (Predicted, Actual)
Limit of Experimental Reproducibility
FBA
GBA
MNMC Ontotype
Matched
Fully Connected
DeepCell VNN
Previous Predictors of Genetic Interaction
Black Box Neural Networks
d
Correlation (Predicted, Actual)
DeepCell VNN
Matched Black Box
Neurons per Subsystem

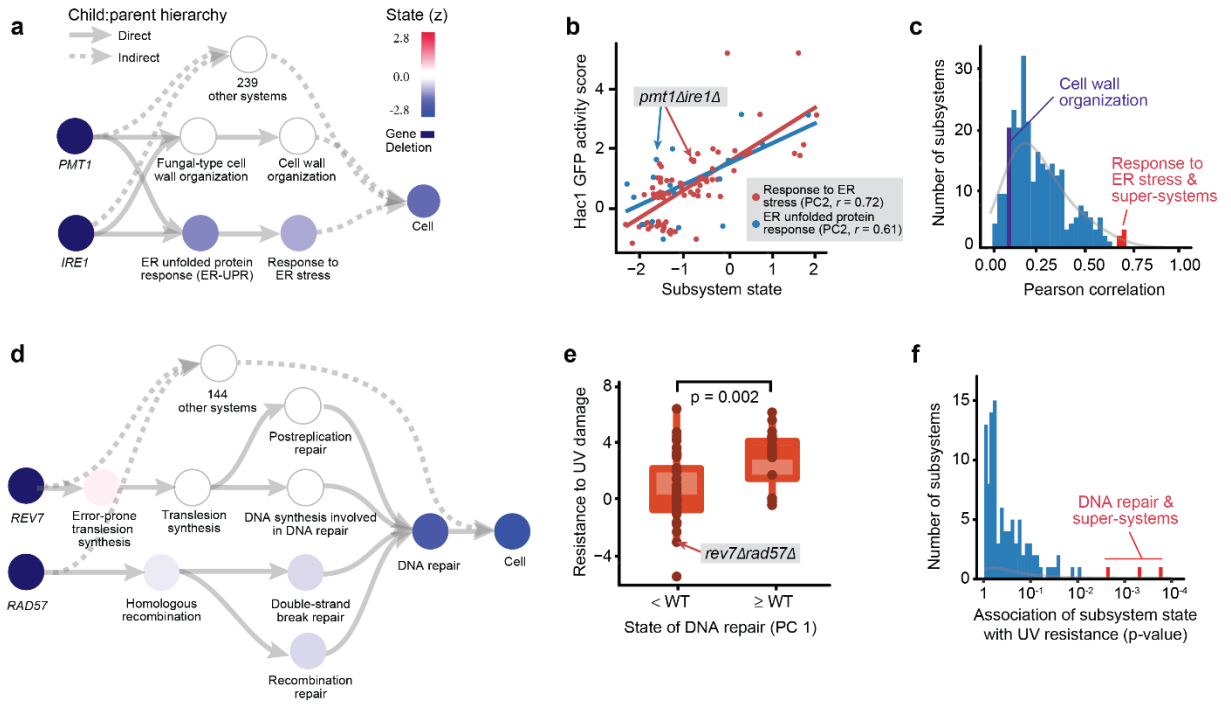


Figure 3.3. Interpretation of genotype-phenotype associations. a, Hierarchy of candidate subsystems that can explain the association of the *pmt1Δire1Δ* genotype with a negative genetic interaction phenotype (synthetic lethality). **b**, Correspondence between Hac1 GFP activity and the functional states of ER-UPR (blue) or its parent subsystem, Response to ER stress (red). Points represent genotypes, with the *pmt1Δire1Δ* genotype indicated. **c**, Distribution of correlations between Hac1 GFP activity and the states of every other subsystem containing at least 10 genotypes with measured GFP activity. **d**, DNA repair appears as a highly altered subsystem that explains the slow growth phenotype of *rev7Δrad57Δ*. **e**, Experimental resistance to UV damage plotted against the state of the DNA repair subsystem in DeepCell, separated into two classes: above or below the wild-type value. Significance measured by Mann–Whitney *U* test. **f**, Distribution of the associations between UV damage resistance and the states of other subsystems containing at least 10 genotypes with measured UV damage resistance value. Note panels a-c use genetic interaction prediction as the phenotypic readout; panels d-f use growth. All panels implement GO as the structural hierarchy for the DeepCell model.

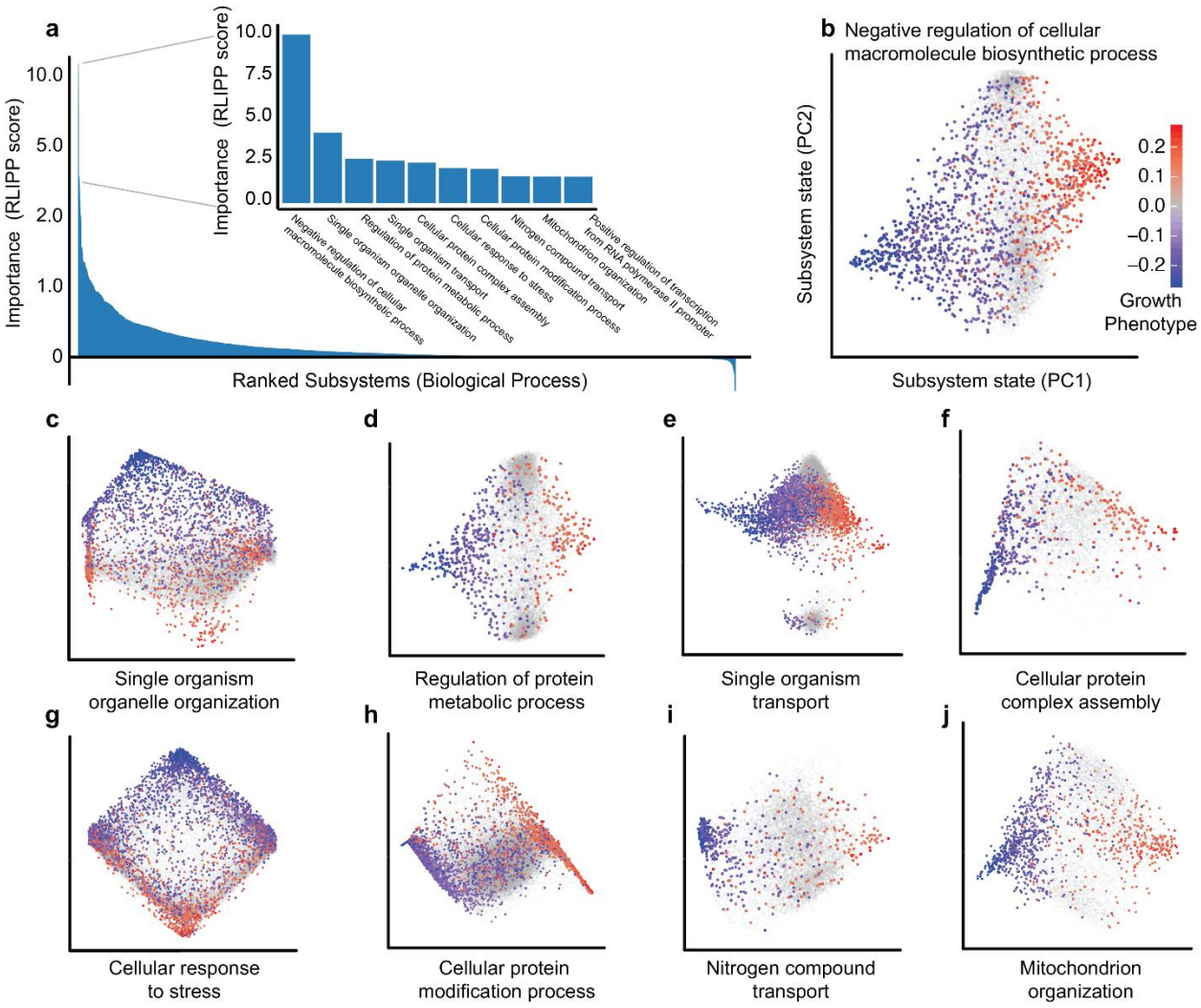


Figure 3.4. Top subsystem states for translation of genotype to growth. a, Ranking of all cellular subsystems in GO by their importance in determining genetic interactions (RLIPP score, **Methods**). Inset: ten highest-scoring subsystems. **b-j**, Two-dimensional state maps of informative subsystems (PC2 versus PC1). **Supplemental Figure S3.2** provides equivalent information for the CliXO hierarchy.

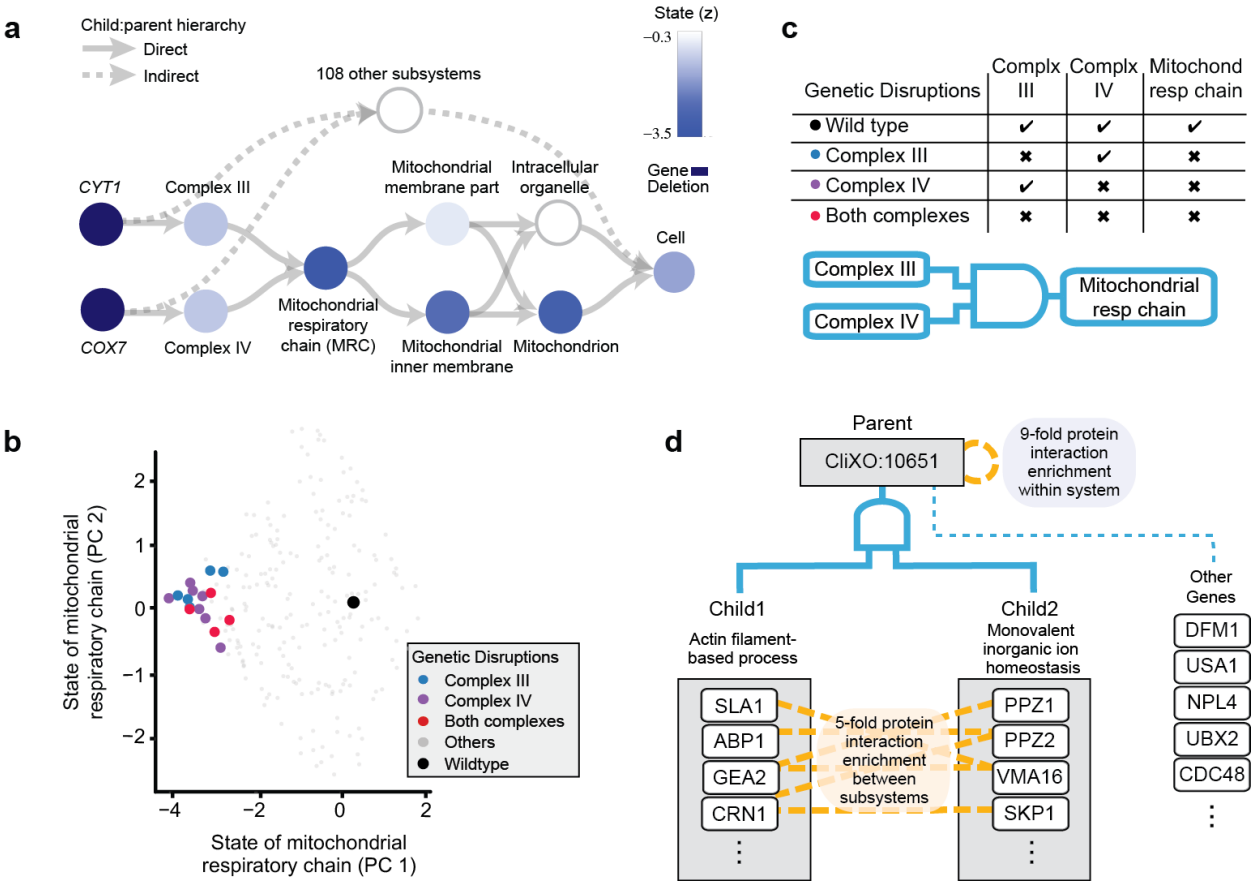


Figure 3.5. Analysis of subsystem functional logic. a, Causal hierarchy of important subsystems involving changes to Mitochondrial Respiratory Chain (MRC), displayed as in **Figure 3.3a**. **b**, 2D state map of MRC plotted as in **Figure 3.3b**. Genotypes disrupting MRC complex III only, MRC complex IV only, or both complexes are demarcated by colors. **c**, Truth table relating the state of MRC to the states of its children. The logic resembles an AND gate, pictured. **d,** A cellular subsystem newly identified by the CliXO hierarchical clustering algorithm (see text), which implements an AND function integrating the states of two known children: Actin filament-based process and Monovalent inorganic ion homeostasis. Identification of this novel subsystem was based on strong enrichments for protein-protein interactions between the two child subsystems and within its entire set of proteins as a whole.

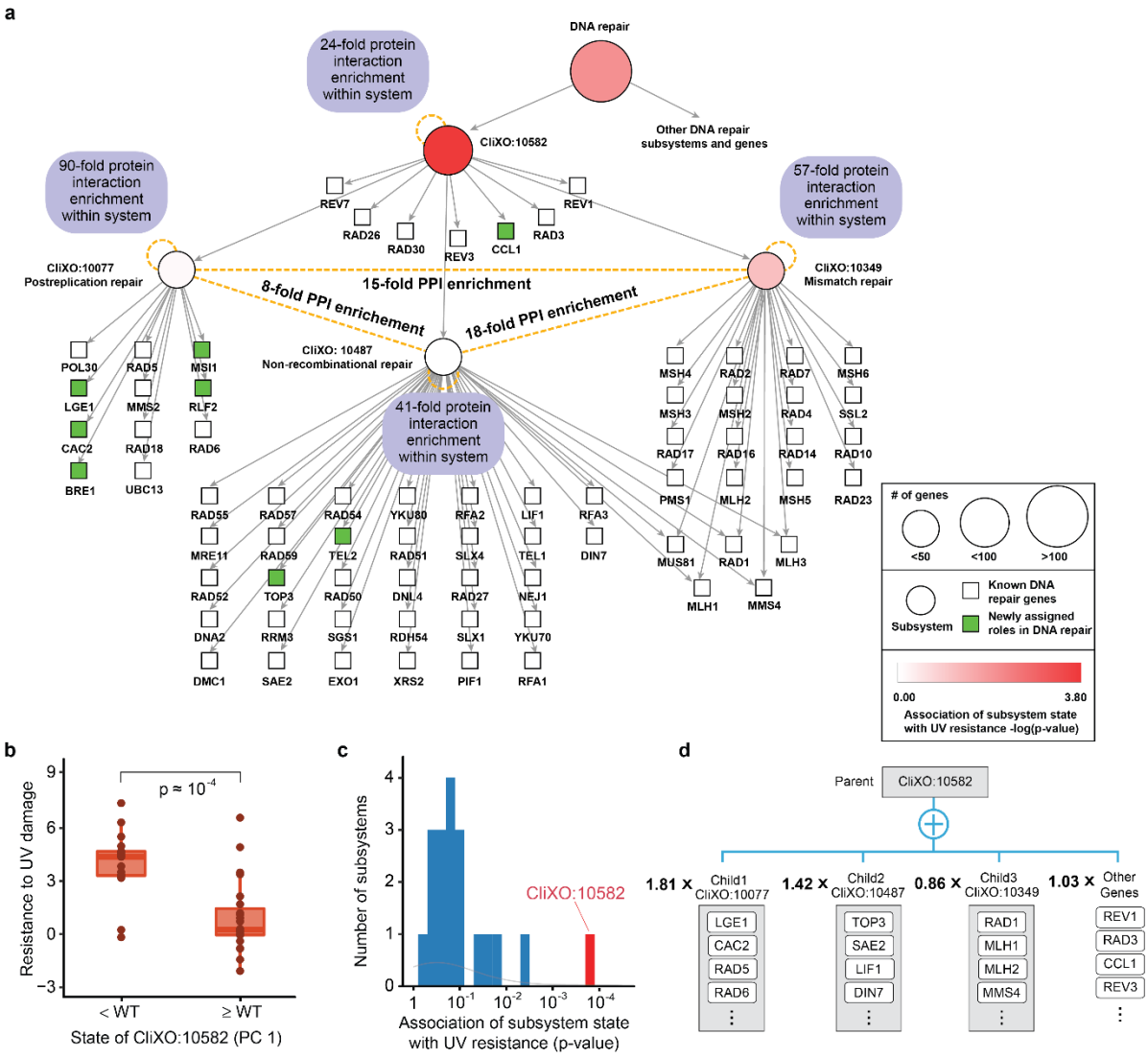


Figure 3.6. Analysis of a new DNA repair subsystem. a, Hierarchical organization of DNA repair including subsystems newly identified by CliXO. **b**, Experimental resistance to UV damage plotted against the state of CliXO:10582, separated into two classes: above or below its wild-type value. Significance measured by Mann–Whitney *U* test. **c**, Distribution of the associations between UV damage resistance and the states of all the subsystems identified by CliXO with at least 10 genotypes with measured UV damage resistance value. **d**, Approximate mathematical function implemented by CliXO:10582 neurons.

3.7 Supplemental Figures

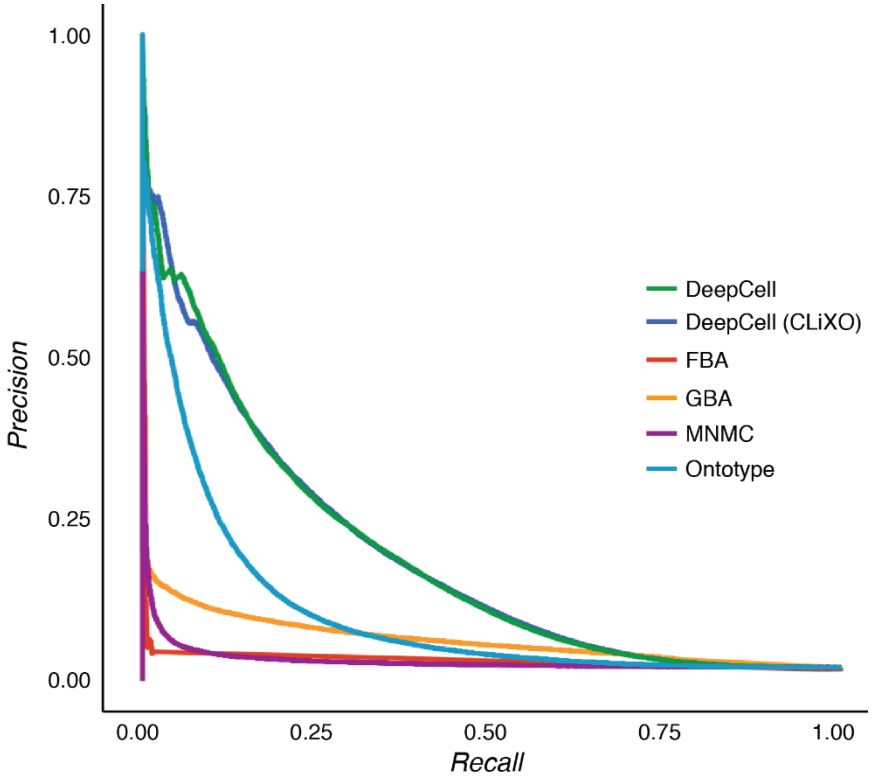


Figure S3.1. Precision-recall curves for classification of negative genetic interactions. Performance of DeepCell is compared to the same methods as in **Figure 3.2c**. Genetic interactions with scores ≤ -0.08 are labeled as negative.

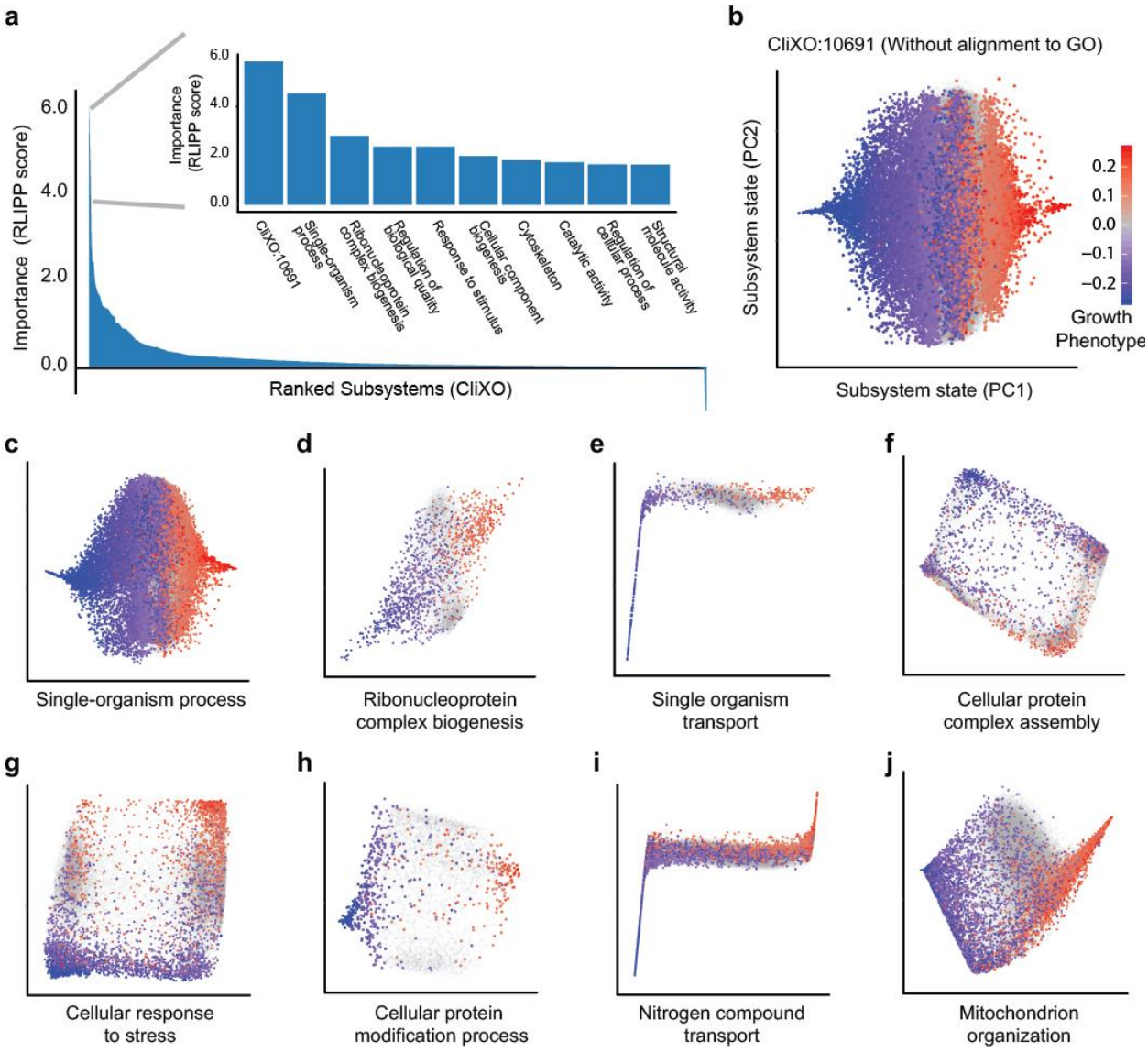


Figure S3.2. CliXO top subsystem states for translation of genotype to growth. a, Ranking of all CliXO subsystems by their importance in determining genetic interactions (RLIPP score, see **Methods**). Inset: ten highest-scoring subsystems. **b-j**, Two-dimensional state maps of informative subsystems from (a), in which each subsystem's set of neuron states is reduced to the first two Principal Components (PCs). Each point represents the subsystem state induced by a genotype, with point color indicating the corresponding growth phenotype (genetic interaction score).

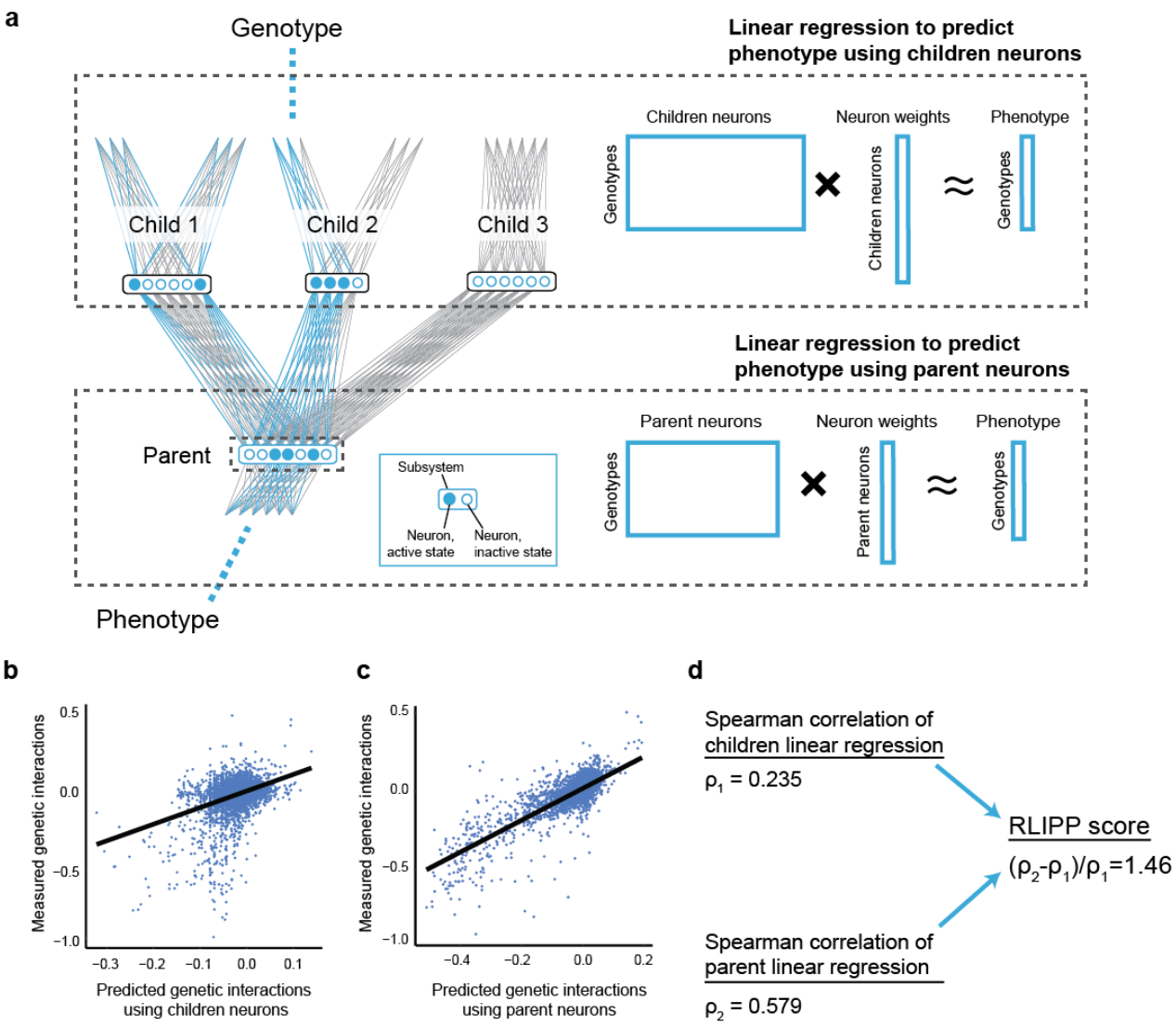


Figure S3.3. Calculating Relative Local Improvement in Predictive Power (RLIPP). **a,** Two L2-regularized linear regression models are fit to predict phenotype using either the neurons of a parent subsystem (bottom) or the neurons of that subsystem's children (top). **b-c**, Predicted versus measured phenotype (genetic interactions) for the children-based model **(b)** or the parent-based model **(c)**. The example values are for the "DNA repair" subsystem. **d**, The RLIPP score is calculated from the Spearman correlation of both models.

3.8 Acknowledgments

Chapter 3, in full, has been submitted for publication of the material as it may appear in *Nature Methods*, 2017. “Using deep learning to model the hierarchical structure and function of a cell”. Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

3.9 References

Alain, Guillaume, and Yoshua Bengio. 2016. “Understanding Intermediate Layers Using Linear Classifier Probes.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1610.01644.

Allen, Benjamin L., and Dylan J. Taatjes. 2015. “The Mediator Complex: A Central Integrator of Transcription.” *Nature Reviews. Molecular Cell Biology* 16 (3): 155–66.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. “Neural Machine Translation by Jointly Learning to Align and Translate.” *arXiv Preprint arXiv:1409. 0473*. https://arxiv.org/abs/1409.0473.

Baryshnikova, Anastasia, Michael Costanzo, Yungil Kim, Huiming Ding, Judice Koh, Kiana Toufighi, Ji-Young Youn, et al. 2010. “Quantitative Analysis of Fitness and Genetic Interactions in Yeast on a Genome Scale.” *Nature Methods* 7 (12): 1017–24.

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. 2011. “D^3: Data-Driven Documents.” *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2301–9.

Brosin, Henry W. 1958. “An Introduction to Cybernetics.” *The British Journal of Psychiatry: The Journal of Mental Science* 104 (435). The Royal College of Psychiatrists: 590–92.

Cadet, Jean, Evelyne Sage, and Thierry Douki. 2005. “Ultraviolet Radiation-Mediated Damage to Cellular DNA.” *Mutation Research* 571 (1-2): 3–17.

Carvunis, Anne-Ruxandra, and Trey Ideker. 2014. “Siri of the Cell: What Biology Could Learn from the iPhone.” *Cell* 157 (3): 534–38.

Chakraborty, Supriyo, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, et al. 2017. “Interpretability of Deep Learning Models: A Survey of Results.” In . DAIS. https://pdfs.semanticscholar.org/8358/f92fc5b8002ab11321d74d82a6bcb82b3f09.pdf?cm_mc_uid=93027777065415065530138&cm_mc_sid_50200000=1506392084.

Chen, William W., Mario Niepel, and Peter K. Sorger. 2010. “Classic and Contemporary Approaches to Modeling Biochemical Reactions.” *Genes & Development* 24 (17): 1861–75.

Collobert, Ronan, Koray Kavukcuoglu, and Clément Farabet. 2011. “Torch7: A Matlab-like Environment for Machine Learning.” In *BigLearn, NIPS Workshop*. https://infoscience.epfl.ch/record/192376/files/Collobert_NIPSWORKSHOP_2011.pdf.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. “Natural Language Processing (Almost) from Scratch.” *Journal*

of Machine Learning Research: JMLR 12 (Aug): 2493–2537.

Copley, Shelley D. 2012. “Moonlighting Is Mainstream: Paradigm Adjustment Required.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 34 (7): 578–88.

Costanzo, Michael, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D. Spear, Carolyn S. Sevier, Huiming Ding, et al. 2010. “The Genetic Landscape of a Cell.” *Science* 327 (5964): 425–31.

Costanzo, Michael, Benjamin VanderSluis, Elizabeth N. Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, et al. 2016. “A Global Genetic Interaction Network Maps a Wiring Diagram of Cellular Function.” *Science* 353 (6306). doi:10.1126/science.aaf1420.

Dutkowski, Janusz, Michael Kramer, Michal A. Surma, Rama Balakrishnan, J. Michael Cherry, Nevan J. Krogan, and Trey Ideker. 2013. “A Gene Ontology Inferred from Molecular Networks.” *Nature Biotechnology* 31 (1): 38–45.

Farabet, Clément, Camille Couprie, Laurent Najman, and Yann Lecun. 2013. “Learning Hierarchical Features for Scene Labeling.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8): 1915–29.

Farrugia, Gianluca, and Rena Balzan. 2012. “Oxidative Stress and Programmed Cell Death in Yeast.” *Frontiers in Oncology* 2. doi:10.3389/fonc.2012.00064.

Franz, Max, Christian T. Lopes, Gerardo Huck, Yue Dong, Onur Sumer, and Gary D. Bader. 2016. “Cytoscape.js: A Graph Theory Library for Visualisation and Analysis.” *Bioinformatics* 32 (2): 309–11.

Free, Stephen J. 2013. “Fungal Cell Wall Organization and Biosynthesis.” In *Advances in Genetics*, 33–82.

Gene Ontology Consortium. 2015. “Gene Ontology Consortium: Going Forward.” *Nucleic Acids Research* 43 (Database issue): D1049–56.

Gormley, Clinton, and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. 1 edition. O’Reilly Media.

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, et al. 2012. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups.” *IEEE Signal Processing Magazine* 29 (6): 82–97.

Hochstrasser, Mark. 2016. “Gyre and Gimble in the Proteasome.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (46): 12896–98.

Ioffe, Sergey, and Christian Szegedy. 2015. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” *arXiv [cs.LG]*. arXiv.

http://arxiv.org/abs/1502.03167.

Jonikas, Martin C., Sean R. Collins, Vladimir Denic, Eugene Oh, Erin M. Quan, Volker Schmid, Jimena Weibezahn, et al. 2009. “Comprehensive Characterization of Genes Required for Protein Folding in the Endoplasmic Reticulum.” *Science* 323 (5922): 1693–97.

Karr, Jonathan R., Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival Jr, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. 2012. “A Whole-Cell Computational Model Predicts Phenotype from Genotype.” *Cell* 150 (2): 389–401.

Kim, Hanhae, Junha Shin, Eiru Kim, Hyojin Kim, Sohyun Hwang, Jung Eun Shim, and Insuk Lee. 2014. “YeastNet v3: A Public Database of Data-Specific and Integrated Functional Gene Networks for Saccharomyces Cerevisiae.” *Nucleic Acids Research* 42 (Database issue): D731–36.

Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980.

Kramer, Michael, Janusz Dutkowski, Michael Yu, Vineet Bafna, and Trey Ideker. 2014. “Inferring Gene Ontologies from Pairwise Similarity Data.” *Bioinformatics* 30 (12): i34–42.

“Kubernetes.” 2017. *Kubernetes*. Accessed March 2. http://kubernetes.io/.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” *Nature* 521 (7553): 436–44.

Lee, Chen-Yu, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang, and Zhuowen Tu. 2015. “Deeply-Supervised Nets.” In *AISTATS*, 2:5.

Lee, Insuk, Ben Lehner, Tanya Vavouri, Junha Shin, Andrew G. Fraser, and Edward M. Marcotte. 2010. “Predicting Genetic Modifier Loci Using Functional Gene Networks.” *Genome Research* 20 (8): 1143–53.

Lei, Tao, Regina Barzilay, and Tommi Jaakkola. 2016. “Rationalizing Neural Predictions.” *arXiv Preprint arXiv:1606. 04155*. https://arxiv.org/abs/1606.04155.

Lipton, Zachary C. 2016. “The Mythos of Model Interpretability.” *arXiv Preprint arXiv:1606. 03490*. https://arxiv.org/abs/1606.03490.

Mahendran, Aravindh, and Andrea Vedaldi. 2015. “Understanding Deep Image Representations by Inverting Them.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5188–96.

Marini, Joe. 2002. *Document Object Model*. 1st ed. New York, NY, USA: McGraw-Hill, Inc.

Mikolov, T., A. Deoras, D. Povey, L. Burget, and J. Černocký. 2011. “Strategies for Training Large Scale Neural Network Language Models.” In *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, 196–201.

Pandey, Gaurav, Bin Zhang, Aaron N. Chang, Chad L. Myers, Jun Zhu, Vipin Kumar, and Eric E. Schadt. 2010. “An Integrative Multi-Network and Multi-Classifier Approach to Predict Genetic Interactions.” *PLoS Computational Biology* 6 (9). doi:10.1371/journal.pcbi.1000928.

Pareto, Vilfredo, and Alfred N. Page. 1971. “Translation of Manuale Di Economia Politica (‘Manual of Political Economy’).” *AM Kelley*.

Pujol-Carrion, Nuria, and Maria Angeles de la Torre-Ruiz. 2010. “Glutaredoxins Grx4 and Grx3 of Saccharomyces Cerevisiae Play a Role in Actin Dynamics through Their Trx Domains, Which Contributes to Oxidative Stress Resistance.” *Applied and Environmental Microbiology* 76 (23): 7826–35.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1988. “Learning Representations by Back-Propagating Errors.” *Cognitive Modeling* 5 (3): 1.

Sainath, T. N., A. r. Mohamed, B. Kingsbury, and B. Ramabhadran. 2013. “Deep Convolutional Neural Networks for LVCSR.” In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8614–18.

Scrimale, Thomas, Louis Didone, Karen L. de Mesy Bentley, and Damian J. Krysan. 2009. “The Unfolded Protein Response Is Induced by the Cell Wall Integrity Mitogen-Activated Protein Kinase Signaling Cascade and Is Required for Cell Wall Integrity in Saccharomyces Cerevisiae.” *Molecular Biology of the Cell* 20 (1): 164–75.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. “Mastering the Game of Go with Deep Neural Networks and Tree Search.” *Nature* 529 (7587): 484–89.

Srivas, Rohith, Thomas Costelloe, Anne-Ruxandra Carvunis, Sovan Sarkar, Erik Malta, Su Ming Sun, Marijke Pool, et al. 2013. “A UV-Induced Genetic Network Links the RSC Complex to Nucleotide Excision Repair and Shows Dose-Dependent Rewiring.” *Cell Reports* 5 (6): 1714–24.

Stefanov, Stoyan. 2016. *React: Up & Running: Building Web Applications*. 1 edition. O’Reilly Media.

Szappanos, Balázs, Károly Kovács, Béla Szamecz, Frantisek Honti, Michael Costanzo, Anastasia Baryshnikova, Gabriel Gelius-Dietrich, et al. 2011. “An Integrated Approach to Characterize Genetic Interaction Networks in Yeast Metabolism.” *Nature Genetics* 43 (7): 656–62.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. n.d. “Going

Deeper with Convolutions." In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. IEEE.

The Gene Ontology Consortium. 2016. "Expansion of the Gene Ontology Knowledgebase and Resources." *Nucleic Acids Research*, November. doi:10.1093/nar/gkw1108.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of GWAS Discovery." *American Journal of Human Genetics* 90 (1): 7–24.

Vondrick, Carl, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. 2013. "Hoggles: Visualizing Object Detection Features." In *Proceedings of the IEEE International Conference on Computer Vision*, 1–8.

Walter, P., and D. Ron. 2011. "The Unfolded Protein Response: From Stress Pathway to Homeostatic Regulation." *Science* 334 (6059): 1081–86.

Weinzaepfel, P., H. Jégou, and P. Pérez. 2011. "Reconstructing an Image from Its Local Descriptors." In *CVPR 2011*, 337–44.

Xu, Chengchao, Songyu Wang, Guillaume Thibault, and Davis T. W. Ng. 2013. "Futile Protein Folding Cycles in the ER Are Terminated by the Unfolded Protein O-Mannosylation Pathway." *Science* 340 (6135): 978–81.

Yang, Jian, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna A. E. Vinkhuyzen, Sang Hong Lee, Matthew R. Robinson, et al. 2015. "Genetic Variance Estimation with Imputed Variants Finds Negligible Missing Heritability for Human Height and Body Mass Index." *Nature Genetics* 47 (10): 1114–20.

Yang, Jian, Noah A. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. 2014. "Advantages and Pitfalls in the Application of Mixed-Model Association Methods." *Nature Genetics* 46 (2): 100–106.

Yu, Michael Ku, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason Kreisberg, Cherie T. Ng, Nevan Krogan, Roded Sharan, and Trey Ideker. 2016. "Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems." *Cell Systems* 2 (2): 77–88.