Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cell Syst. Author manuscript; available in PMC 2017 February 24.

Published in final edited form as:

Cell Syst. 2016 February 24; 2(2): 77–88. doi:10.1016/j.cels.2016.02.003.

Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems

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

[1]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla CA 92093, USA

[2]Department of Medicine, University of California San Diego, La Jolla CA 92093, USA

[3]Biomedical Sciences Program, University of California San Diego, La Jolla CA 92093, USA

[4]Data4Cure, La Jolla, CA 92037, USA

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

[6]aTyr Pharmaceuticals, San Diego, CA 92121, USA

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

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

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

This manuscript version is made available under the CC BY-NC-ND 4.0 license.

†Correspondence: tideker@ucsd.edu.

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.

The authors wish to declare no competing financial interests related to this work.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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; Gligorijević et al., 2014; Kramer et al., 2014). This ‘data-driven’ ontology closely resembles, and in some cases greatly revises and expands, the literature-curated GO.

Here we report a general approach for using deep hierarchical knowledge of the cell, represented by an ontology, to translate genotype to phenotype. This approach recursively aggregates the effects of genetic variation 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.

Results

Association between genetic interactions and hierarchical relations among cellular systems

As preparation for modeling, we identified patterns by which genetic interactions are associated with, and thus biologically explained by, the structure of gene ontologies. We observed that sets of genes assigned to the same GO term tended to be highly enriched for genetic interactions ($p < 10^{-5}$), for both positive genetic interactions (double gene disruptions with better-than-expected growth, e.g. epistasis) and negative genetic interactions (double gene disruptions with worse-than-expected growth, e.g. synthetic lethality) (Figure 1A). Such interaction enrichment within GO terms occurred over a wide

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 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 1C). Occurrences of interactions within or between biological pathways have been previously investigated as separate biological interpretations (Bandyopadhyay et al., 2008; Bellay et al., 2011; Collins et al., 2010; Kelley and Ideker, 2005; Leiserson et al., 2011; Ma et al., 2008; Qi et al., 2008; Ulitsky et al., 2008). Here, both types of explanations can be applied to the same interaction, as they are related hierarchically within the unified structure of the cell. Overall, approximately 40,000 interactions were involved in 1,661 within- or between-term enrichments, representing a 24:1 compression of information (Figure 1D). Thus, 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. *b*Δ*d*Δ, Figure 2A). These disruptions are propagated recursively up the ontology, such that every term is assigned the disrupted genes annotated to that term plus all of those assigned to its children. For example, since the gene *KIP1* encodes a subunit of the kinesin complex (Figure 1B), its deletion in a *kip1*Δ 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

signaling pathways (Figure 2A). In its most general definition, ontotype encompasses both genotype and phenotype, with genes and observable traits positioned at lower and higher levels of the hierarchy of objects encoding life.

A functionalized gene ontology integrating cell structure and functional prediction

Third, once genotypes are transformed to ontotypes, a supervised learning approach based on the technique of random forests regression (Breiman, 2001) is used to learn rules by which term states predict phenotypes. Rules are organized as a collection, or 'forest', of decision trees (**Experimental Procedures**), with a typical decision tree describing a series of logical true/false tests to evaluate the states of several terms (e.g., T4, T5, and T7 in Figure 2A). Making decisions on the states of terms rather than nucleotide variants or genes enables machine learning across a range of scales, so that different genotypes converging on similar ontotypes (e.g. a$\Delta d\Delta$ and $b\Delta d\Delta$ 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 3A,B). The correlation between predicted and measured scores was highly significant (Figure 3C, Pearson's $r = 0.35$, $p < 2.2\times10^{-16}$) 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 3D,E), indicating that all scales in the hierarchy aid in prediction. F_{NeXO} achieved nearly the same correlation (Figure 3C, $r = 0.32$) and was also sensitive to randomization ($r = 0.03$).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 (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 3C) as well as better tradeoffs in precision versus recall (Figure 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 (Gillis and Pavlidis, 2012; Mackay, 2014) 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 S1).

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). 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 S2A–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 3C), even though the same random forests regression procedure was used as for the functionalized ontologies.

Simulating growth phenotypes for 'new' genotypes not yet observed or examined

We next used F_{GO} to simulate growth for all 12,512,503 pairwise deletions of non-essential yeast genes, 73% of which had not yet been tested in the laboratory (Figure 4A, Supplemental File S1). A total of 41,605 genetic interactions were predicted. These

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

predictions were concentrated within and between particular terms and term pairs (Figure 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 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 S3). Of particular interest were predicted interactions between 71 term pairs, as these terms were only distantly related in GO (Table 1, Supplemental Table S1, Supplemental Experimental Procedures). For example, all ten genes in 'intron homing' had negative interactions with all four genes in the 'Phosphatidylinositol-3-kinase complex', although neither these terms nor their parents shared any genes, and these terms were in entirely separate branches of 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 5A, Supplemental Figure S4). Genetic perturbations within each term led to particularly accurate growth phenotypes in cross-validation, as the correlation between predicted interactions and those measured by Costanzo et al. 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 S2G, Supplemental Table S2). To test whether this performance generalized to new data, we experimentally measured growth phenotypes for 1,218 pairwise deletions of DNA repair genes and 1,600 pairwise deletions of nuclear lumen genes and scored these mutants for genetic interactions (Supplemental Table S3, 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 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 5C) and precision versus recall (Figure 5D) in predicting the new interactions measured for DNA repair and nuclear lumen genes. Given this validation, we combined the genetic interaction scores from both new screens with previous data (Costanzo et al., 2010) and re-trained the ontotype decision logic on this more complete dataset. The structure of this improved F_{GO}, with the accompanying ontotype-phenotype logic, is available online on the Network Data Exchange (http://goo.gl/cYIXWJ,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

UUID: 01b46d52-c3a5-11e5-8fbc-06603eb7f303, Pratt et al., 2015) and as a Cytoscape file in Supplemental File S2.

Toward more complex genotypes

Although the ontotype had been trained using double deletion genotypes, we hypothesized that, once trained, it might be capable of predictions for genotypes involving mutations to larger numbers of genes. Although few studies have examined three-way or higher-order genetic interactions, a recent study (Haber et al., 2013) showed proof-of-principle for a three-way gene deletion methodology, representing one of the few systematic screens for triple mutants to-date. This work reported that deletion of *CAC1* in combination with any gene in the HIR complex (*HIR1*, *HIR2*, *HIR3*, *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 6A). Visual inspection of the model (Figure 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.

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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Functionalized ontologies substantially outperformed previous phenotypic predictors (Figure 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 3C) or truncated versions of the ontology (Figure 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 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Extending Functional Ontologies Beyond Current Limits

F_{GO} based its predictions principally on 1,367 terms, spread across various biological processes, cellular components and molecular functions (Figure 4A). Although this coverage of cell biology is substantial (27% of the yeast GO), one might wonder whether it should be more complete. First, some term logic is likely missed because those terms are not frequently disrupted in the current set of genotypes. For example, genes annotated to 783 GO terms were never disrupted in any genotype tested (Costanzo et al., 2010). Second, some biological processes are likely not required for the phenotype tested – growth of cells in rich media – but instead may drive a wide variety of other phenotypes (Dowell et al., 2010; Hillenmeyer et al., 2008; Ideker and Krogan, 2012; Lee et al., 2014). Third, important processes or components may not yet have been curated in 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.

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),

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 S4), following a previously described procedure (Dutkowski et al., 2013, http://mhk7.github.io/alignOntology/).

To construct NeXO (Supplemental Table S5), 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 S5), 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 S6); this performance difference may be due to the five basic features being collected more recently. See also Supplemental Experimental Procedures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Acknowledgments

We gratefully acknowledge helpful discussion and comments from Hannes Braberg, Anne-Ruxandra Carvunis, Manolis Kellis, Benjamin Kellman, Jianzhu Ma, Jenhan Tao, Alex Thomas, members of the Ideker laboratory, and the anonymous referees. This work was funded by the National Institute of General Medical Sciences (P41-GM103504, P50-GM085764) and the National Institute of Environmental Health Sciences (R01-ES014811). 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).

References

Apgar JF, Witmer DK, White FM, Tidor B. Sloppy models, parameter uncertainty, and the role of experimental design. Mol Biosyst. 2010; 6:1890–1900. [PubMed: 20556289]

Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp Ja, Blom JG. Systems biology: parameter estimation for biochemical models. FEBS J. 2009; 276:886–902. [PubMed: 19215296]

Balakrishnan R, Harris Ma, Huntley R, Van Auken K, Cherry JM. A guide to best practices for Gene Ontology (GO) manual annotation. Database (Oxford). 2013

Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. Functional maps of protein complexes from quantitative genetic interaction data. PLoS Comput Biol. 2008; 4:e1000065. [PubMed: 18421374]

Baryshnikova A, Costanzo M, Kim Y, Ding H, Koh J, Toufighi K, Youn J, Ou J, Luis BS, Bandyopadhyay S, et al. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat Methods. 2010; 7:1017–1024. [PubMed: 21076421]

Bellay J, Atluri G, Sing TL, Toufighi K, Costanzo M, Ribeiro PSM, Pandey G, Baller J, VanderSluis B, Michaut M, et al. Putting genetic interactions in context through a global modular decomposition. Genome Res. 2011; 21:1375–1387. [PubMed: 21715556]

Boucher B, Jenna S. Genetic interaction networks: better understand to better predict. Front Genet. 2013; 4:290. [PubMed: 24381582]

Brachman, RJ.; Levesque, HJ. Knowledge Representation and Reasoning. San Francisco: Morgan Kaufmann; 2004.

Breiman L. Random Forests. Mach Learn. 2001; 45:5–32.

Bruggeman FJ, Westerhoff HV. The nature of systems biology. Trends Microbiol. 2007; 15:45–50. [PubMed: 17113776]

Cahan P, Li H, Morris SA, Lummertz da Rocha E, Daley GQ, Collins JJ. CellNet: Network Biology Applied to Stem Cell Engineering. Cell. 2014; 158:903–915. [PubMed: 25126793]

Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. 2012; 44:841–847. [PubMed: 22836096]

Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. Mol Syst Biol. 2014; 10

Carvunis AR, Ideker T. Siri of the cell: what biology could learn from the iPhone. Cell. 2014; 157:534–538. [PubMed: 24766803]

Chen WW, Niepel M, Sorger PK. Classic and contemporary approaches to modeling biochemical reactions. Genes Dev. 2010; 24:1861–1875. [PubMed: 20810646]

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 2012; 40:D700–5. [PubMed: 22110037]

Collins SR, Roguev A, Krogan NJ. Quantitative genetic interaction mapping using the E-MAP approach. Methods Enzymol. 2010; 470:205–231. [PubMed: 20946812]

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, et al. The genetic landscape of a cell. Science. 2010; 327:425–431. [PubMed: 20093466]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Deisboeck TS, Wang Z, Macklin P, Cristini V. Multiscale cancer modeling. Annu Rev Biomed Eng. 2011; 13:127–155. [PubMed: 21529163]

Deutscher D, Meilijson I, Kupiec M, Ruppin E. Multiple knockout analysis of genetic robustness in the yeast metabolic network. Nat Genet. 2006; 38:993–998. [PubMed: 16941010]

Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA, Heisler LE, Chin B, et al. Genotype to Phenotype: A Complex Problem. Science (80-). 2010; 328:469.

Dutkowski J, Kramer M, Surma Ma, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T. A gene ontology inferred from molecular networks. Nat Biotechnol. 2013; 31:38–45. [PubMed: 23242164]

Eissing T, Kuepfer L, Becker C, Block M, Coboeken K, Gaub T, Goerlitz L, Jaeger J, Loosen R, Ludewig B, et al. A computational systems biology software platform for multiscale modeling and simulation: integrating whole-body physiology, disease biology, and molecular reaction networks. Front Physiol. 2011; 2:1–10. [PubMed: 21423411]

Formosa T, Ruone S, Adams MD, Olsen AE, Eriksson P, Yu Y, Rhoades AR, Kaufman PD, Stillman DJ. on the Hir/Hpc Pathway: Polymerase Passage May Degrade Chromatin Structure. Genetics. 2002; 162:1557–1571. [PubMed: 12524332]

Gillis J, Pavlidis P. “Guilt by association” is the exception rather than the rule in gene networks. PLoS Comput Biol. 2012; 8:e1002444. [PubMed: 22479173]

Gligorijević V, Janjić V, Pržulj N. Integration of molecular network data reconstructs Gene Ontology. Bioinformatics. 2014; 30:i594–600. [PubMed: 25161252]

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya Ra, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015; 47:569–576. [PubMed: 25915600]

Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol. 2007; 3:1871–1878. [PubMed: 17922568]

Haber JE, Braberg H, Wu Q, Alexander R, Haase J, Ryan C, Lipkin-Moore Z, Franks-Skiba KE, Johnson T, Shales M, et al. Systematic Triple-Mutant Analysis Uncovers Functional Connectivity between Pathways Involved in Chromosome Regulation. Cell Rep. 2013; 3:2168–2178. [PubMed: 23746449]

Hanahan D, Weinberg Ra. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674. [PubMed: 21376230]

Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science. 2008; 320:362–365. [PubMed: 18420932]

Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013; 10:1108–1115. [PubMed: 24037242]

Huntley RP, Sawford T, Martin MJ, O’Donovan C. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. Gigascience. 2014; 3:4. [PubMed: 24641996]

Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012; 8:565. [PubMed: 22252388]

Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW. A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell. 2012; 150:389–401. [PubMed: 22817898]

Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol. 2005; 23:561–566. [PubMed: 15877074]

Kim Y-A, Przytycka TM. Bridging the Gap between Genotype and Phenotype via Network Approaches. Front Genet. 2012; 3:227. [PubMed: 23755063]

Kim H, Shin J, Kim E, Kim H, Hwang S, Shim JE, Lee I. YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. Nucleic Acids Res. 2014; 42:D731–6. [PubMed: 24165882]

Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T. Inferring gene ontologies from pairwise similarity data. Bioinformatics. 2014; 30:i34–42. [PubMed: 24932003]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Lee AY, Onge RPS, Proctor MJ, Wallace IM, Nile AH, Spagnuolo PA, Jitkova Y, Gronda M, Wu Y, Kim MK, et al. Mapping the Cellular Response to Small Molecules Using Chemogenomic Fitness Signatures. Science (80-). 2014; 344:208–211.

Lee I, Lehner B, Vavouri T, Shin J, Fraser AG, Marcotte EM. Predicting genetic modifier loci using functional gene networks. 2010:1143–1153.

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011; 21:1109–1121. [PubMed: 21536720]

Lehner B. Genotype to phenotype: lessons from model organisms for human genetics. Nat Rev Genet. 2013; 14:168–178. [PubMed: 23358379]

Leiserson MDM, Tatar D, Cowen LJ, Hescott BJ. Inferring mechanisms of compensation from E-MAP and SGA data using local search algorithms for max cut. J Comput Biol. 2011; 18:1399–1409. [PubMed: 21882903]

Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2014; 47

Lerman, Ja; Hyduke, DR.; Latif, H.; Portnoy, Va; Lewis, NE.; Orth, JD.; Schrimpe-Rutledge, AC.; Smith, RD.; Adkins, JN.; Zengler, K., et al. In silico method for modelling metabolism and gene product expression at genome scale. Nat Commun. 2012; 3:929. [PubMed: 22760628]

Ma X, Tarone AM, Li W. Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. PLoS One. 2008; 3:e1922. [PubMed: 18398455]

Machado D, Costa RS, Rocha M, Ferreira EC, Tidor B, Rocha I. Modeling formalisms in Systems Biology. AMB Express. 2011; 1:45. [PubMed: 22141422]

Mackay TFC. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet. 2014; 15:22–33. [PubMed: 24296533]

Ng S, Collisson Ea, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, Benz C, Haussler D, Stuart JM. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics. 2012; 28:i640–i646. [PubMed: 22962493]

O'Brien EJ, Lerman Ja, Chang RL, Hyduke DR, Palsson BØ. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol. 2013; 9:693. [PubMed: 24084808]

Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat. Biotechnol. 2010; 28:245–248.

Pamblanco M, Oliete-Calvo P, García-Oliver E, Luz Valero M, Sanchez del Pino MM, Rodríguez-Navarro S. Unveiling novel interactions of histone chaperone Asf1 linked to TREX-2 factors Sus1 and Thp1. Nucleus. 2014; 5:247–259. [PubMed: 24824343]

Pandey G, Zhang B, Chang AN, Myers CL, Zhu J, Kumar V, Schadt EE. An integrative multi-network and multi-classifier approach to predict genetic interactions. PLoS Comput Biol. 2010; 6

Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. Nat Methods. 2012; 9:1134–1136. [PubMed: 23223166]

Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. Cell. 2011; 144:864–873. [PubMed: 21414479]

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; 12:2825–2830.

Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009; 5:e1000443. [PubMed: 19649320]

Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, et al. NDEx, the Network Data Exchange. Cell Syst. 2015; 1:302–305. [PubMed: 26594663]

Qi Y, Suhail Y, Lin Y, Boeke JD, Bader JS. Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. Genome Res. 2008; 18:1991–2004. [PubMed: 18832443]

Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet. 2012; 28:323–332. [PubMed: 22480918]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Joint Conference on Artificial Intelligence; 1995. p. 448-453.

Schwabish, Ma; Struhl, K. Asf1 mediates histone eviction and deposition during elongation by RNA polymerase II. Mol Cell. 2006; 22:415–422. [PubMed: 16678113]

Segrè D, Deluna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. Nat Genet. 2005; 37:77–83. [PubMed: 15592468]

Skafidas E, Testa R, Zantomio D, Chana G, Everall IP, Pantelis C. Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. Mol Psychiatry. 2014; 19:504–510. [PubMed: 22965006]

Sullivan PF. Puzzling over schizophrenia: schizophrenia as a pathway disease. Nat Med. 2012; 18:210–211. [PubMed: 22310687]

Szappanos B, Kovács K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, Gelius-Dietrich G, Lercher MJ, Jelasity M, Myers CL, et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet. 2011; 43:656–662. [PubMed: 21623372]

Szczurek E, Gat-Viks I, Tiuryn J, Vingron M. Elucidating regulatory mechanisms downstream of a signaling pathway using informative experiments. Mol Syst Biol. 2009; 5:287. [PubMed: 19584836]

Takahashi K, Ishikawa N, Sadamoto Y, Sasamoto H, Ohta S, Shiozawa a, Miyoshi F, Naito Y, Nakayama Y, Tomita M. E-Cell 2: Multi-platform E-Cell simulation system. Bioinformatics. 2003; 19:1727–1729. [PubMed: 15593410]

The Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2014; 43:1049–1056.

Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC, et al. E-CELL: software environment for whole-cell simulation. Bioinformatics. 1999; 15:72–84. [PubMed: 10068694]

Tong AHY, Boone C. Synthetic Genetic Array Analysis in Saccharomyces cerevisiae. Methods Mol Bio. 2006; 313:171–191. [PubMed: 16118434]

Ulitsky I, Shlomi T, Kupiec M, Shamir R. From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. Mol Syst Biol. 2008; 4:209. [PubMed: 18628749]

Waddington CH. Canalization of Development and the Inheritance of Acquired Characters. Nature. 1942; 150:563–565.

Walpole J, Papin Ja, Peirce SM. Multiscale computational models of complex biological systems. Annu Rev Biomed Eng. 2013; 15:137–154. [PubMed: 23642247]

Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010; 11:843–854. [PubMed: 21085203]

Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle Ra, Reilly SK, Lin L, Fertuzinhos S, Miller Ja, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell. 2013; 155:997–1007. [PubMed: 24267886]

Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012; 109:1193–1198. [PubMed: 22223662]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights

- Strategy for genotype-phenotype prediction using a deep hierarchy of cell systems
- Structure of model is seeded from the GO hierarchy or directly from data
- Convergence of genetic variation up the hierarchy enables functional prediction
- Striking ability to translate combinatorial yeast genotypes to growth rate

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

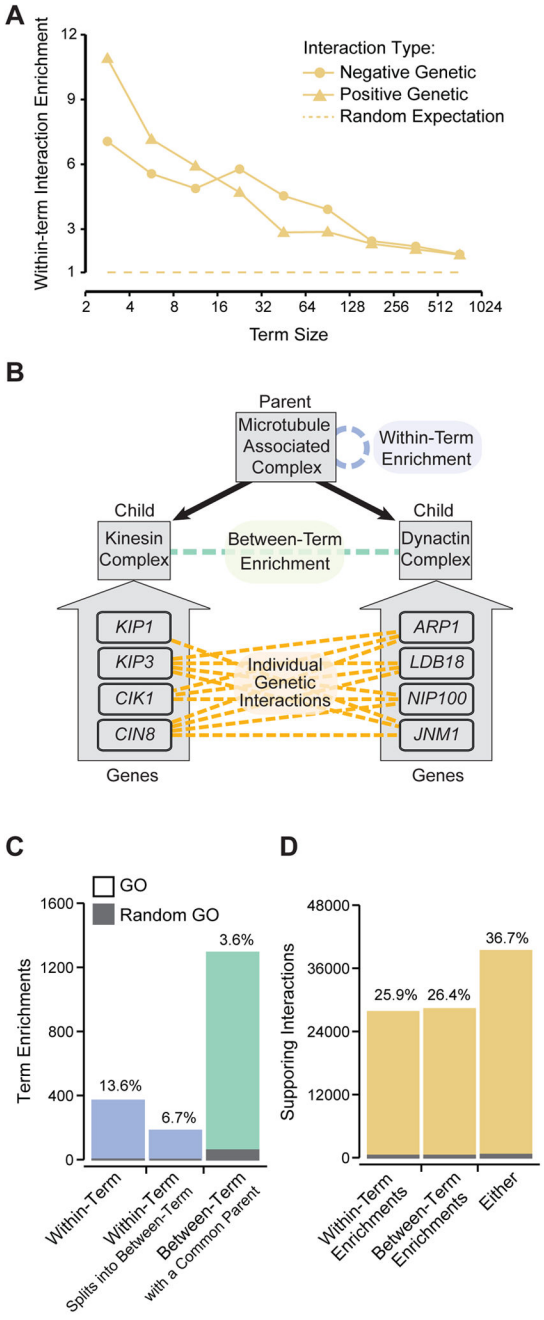

Figure 1. Patterns of genetic interaction reflect the hierarchical structure of the Gene Ontology

(**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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

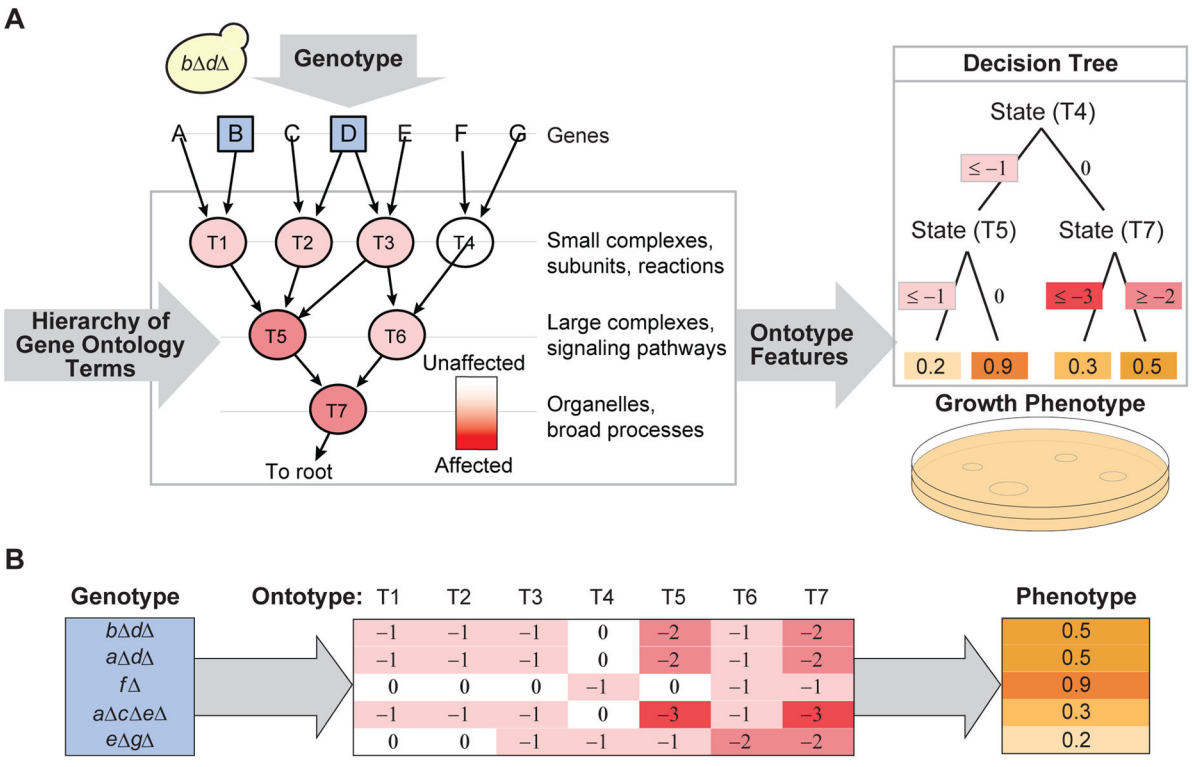

Figure 2. The ontotype method of translating genotype to phenotype

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

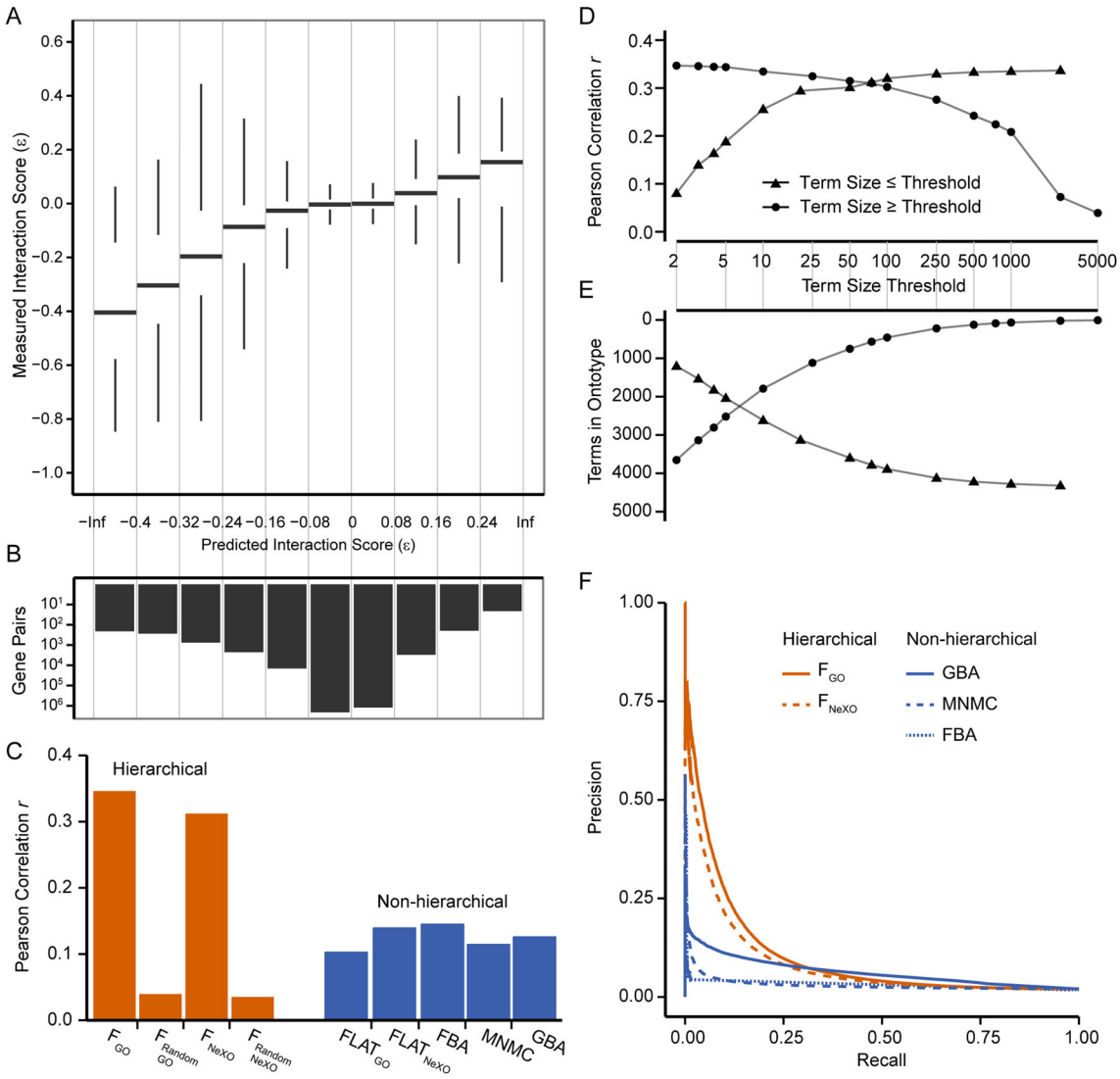

Figure 3. Genome-wide prediction of pairwise genetic interactions in yeast

(A) Measured genetic interaction scores versus those predicted from ontotypes constructed from GO using 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

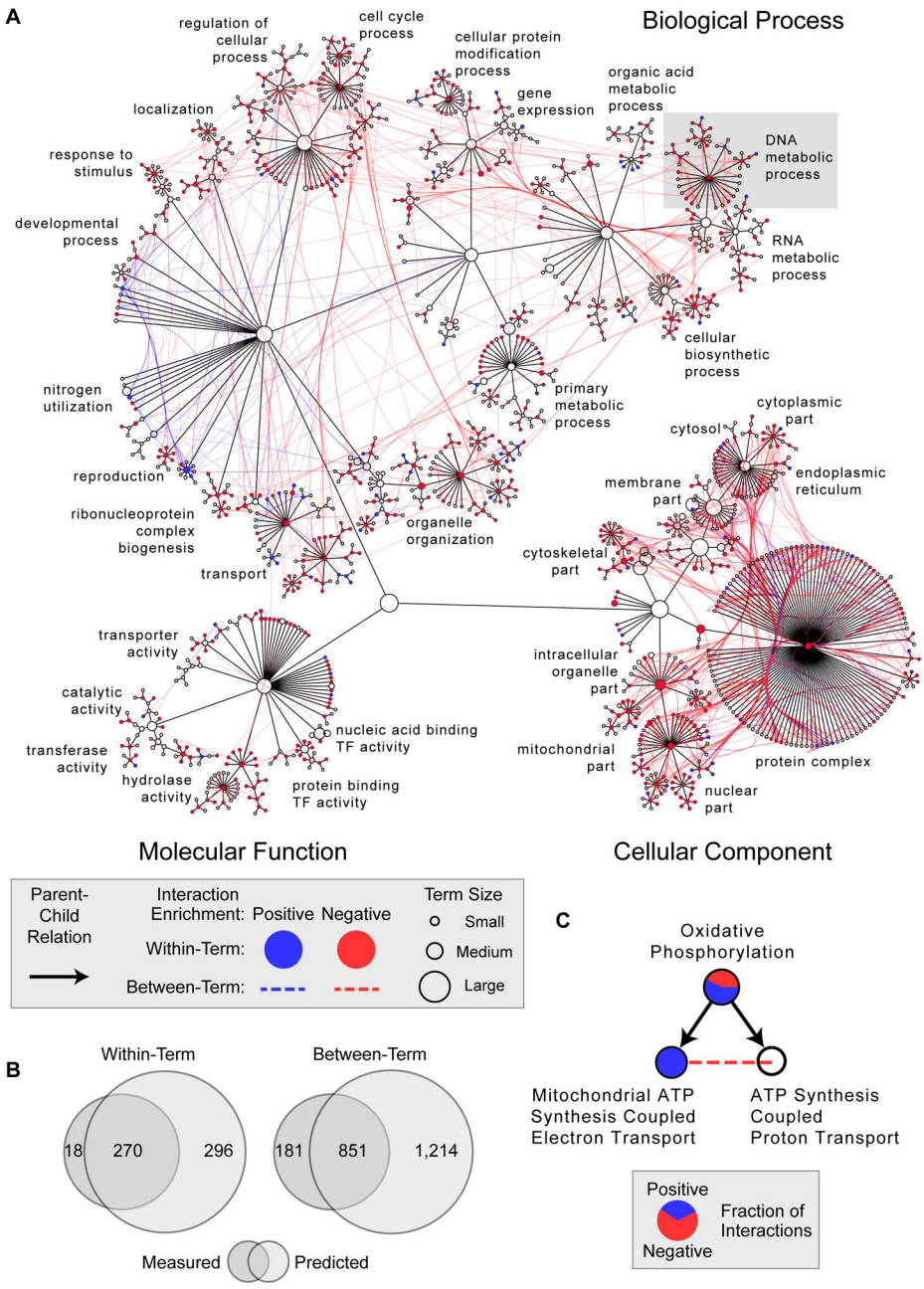

Figure 4. The Functionalized Gene Ontology

(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 S3.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

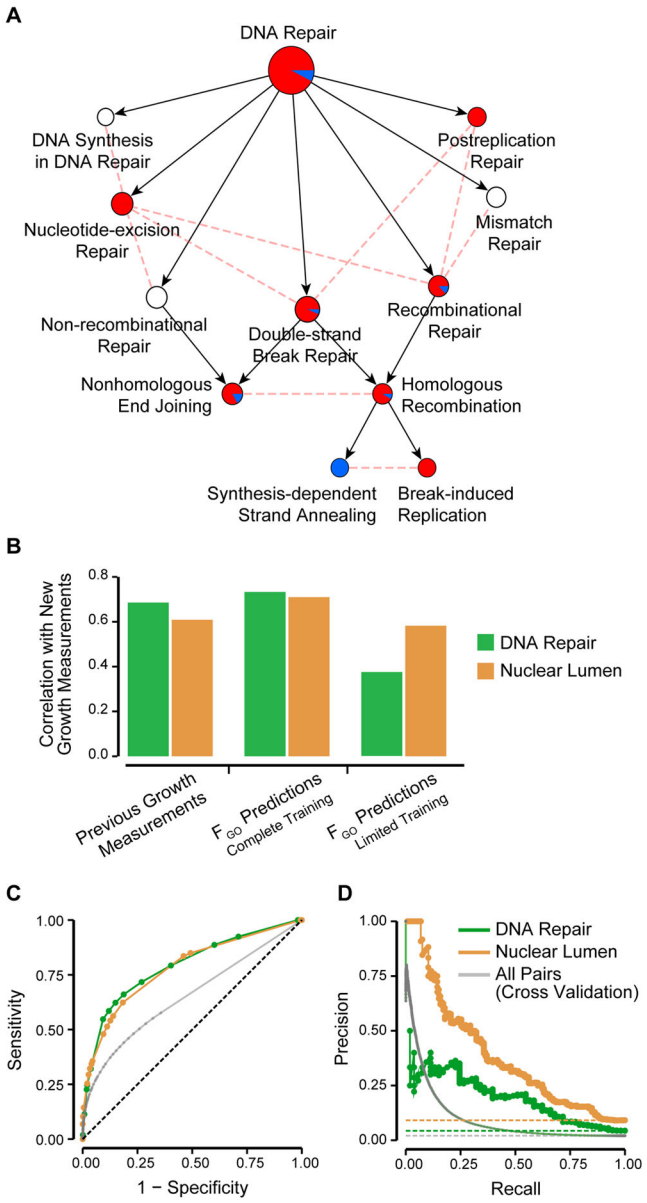

Figure 5. Elucidating the genetic logic of DNA repair and the nuclear lumen

(A) DNA repair has a rich structure of predicted genetic interactions among specific repair processes. Coloring and visual style of panels follow the convention of previous figures. See also Supplemental Figure S4. **(B–D)** Yeast growth was experimentally measured for double gene deletion strains in which both genes are involved in either DNA repair (green) or nuclear lumen (orange). See also Supplemental Table S2–3. **(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 3F).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

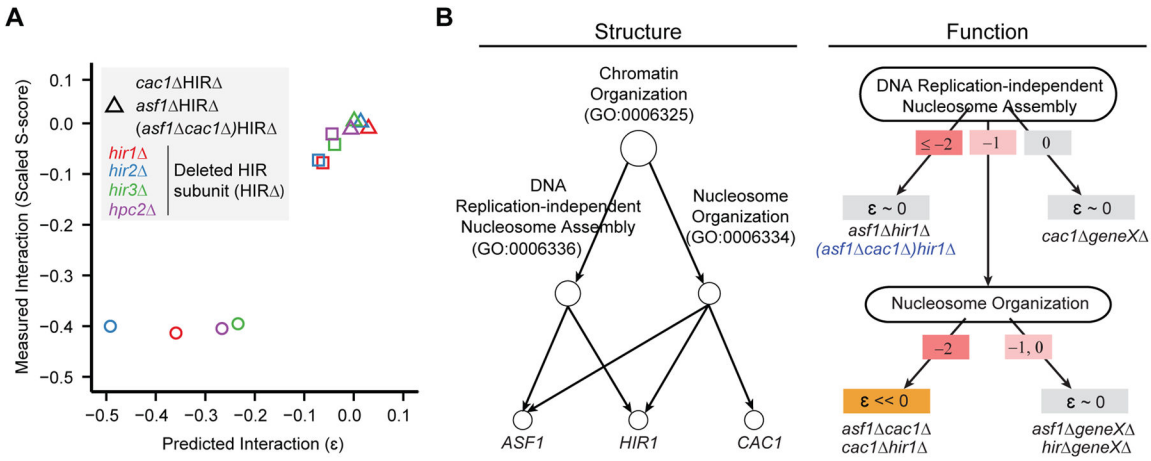

Figure 6. Prediction of triple mutants

(a) Measured versus predicted interaction scores for genotypes involving pairwise and three-way deletions involving *ASF1*, *CAC1*, and genes in the 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).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Top new functional relationships in F_{GO}. See also Supplemental Table S1.

	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