Yeast Features: Identifying Significant Features Shared Among Yeast Proteins for Functional Genomics

[1,2,3,4§] Michel Dumontier, [5]James R. Green, [1,2,4]Ashkan Golshani, [1,2]Myron L. Smith, [1]Nadereh Mir-Rashed, [1]Md Alamgir, [1]Veronika Eroukova, [3]Frank Dehne, [1,2]James J. Cheetham

[1]Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6.
[2]Institute of Biochemistry
[3]Department of Computer Science
[4]Ottawa Institute for Systems Biology
[5]Department of Systems and Computer Engineering

[§]Corresponding author

Email addresses:
MD: michel_dumontier@carleton.ca
JRG: james_green@carleton.ca
AG: ashkan_golshani@carleton.ca
MLS: myron_smith@carleton.ca
NM: nmrashed@ccs.carleton.ca
MA: malamgir@connect.carleton.ca
VE: erukova@hotmail.com
FD: frank@dehne.net
JC: james_cheetham@carleton.ca

Abstract

Background

High throughput yeast functional genomics experiments are revealing associations among tens to hundreds of genes using numerous experimental conditions. To fully understand how the identified genes might be involved in the observed system, it is essential to consider the widest range of biological annotation possible. Biologists often start their search by collating the annotation provided for each protein within databases such as the Saccharomyces Genome Database, manually comparing them for similar features, and empirically assessing their significance. Such tasks can be automated, and more precise calculations of the significance can be determined using established probability measures.

Results

We developed Yeast Features, an intuitive online tool to help establish the significance of finding a diverse set of shared features among a collection of yeast proteins. A total of 18,786 features from the Saccharomyces Genome Database are considered, including annotation based on the Gene Ontology's molecular function, biological process and cellular compartment, as well as conserved domains, protein-protein and genetic interactions, complexes, metabolic pathways, phenotypes and publications. The significance of shared features is estimated using a hypergeometric probability, but novel options exist to improve the significance by adding background knowledge of the experimental system. For instance, increased statistical significance is achieved in gene deletion experiments because interactions with essential genes will never be observed. We further demonstrate the utility by suggesting the functional roles of the indirect targets of an aminoglycoside with a known mechanism of action, and also the targets of an herbal extract with a previously unknown mode of action. The identification of shared functional features may also be used to propose novel roles for proteins of unknown function, including a role in protein synthesis for YKL075C.

Conclusions

Yeast Features (YF) is an easy to use web-based application (http://software.dumontierlab.com/yeastfeatures/) which can identify and prioritize features that are shared among a set of yeast proteins. This approach is shown to be valuable in the analysis of complex data sets, in which the extracted associations revealed significant functional relationships among the gene products.

Background

The availability of complete genome sequences has yielded the ability to perform high-throughput, genome-wide screens of gene function (reviewed in [1]). In particular, significant advances have been made in the understanding of proteins and their properties in the yeast model organism *Saccharomyces cerevisiae*, the target of numerous high-throughput experiments, including genome sequencing [2], expression profiling [3], large-scale interaction [4], localization [5], gene deletion and chemical phenotype assays [6-10]. Computational experiments have uncovered similar sequences [11], conserved domains [12], small molecule binding sites [13] and detection of pathways [14]. Annotation of genes and gene products for molecular function, biological process and cellular compartment annotation has also been facilitated by the availability of the Gene Ontology (GO) controlled vocabulary [15]. While much has been discovered in these past few years, many more systems biology experiments intend to uncover how the perturbation of a complex system might yield insight into its behavior [16]. To firmly establish the function of some gene or protein, one must increasingly consider multiple lines of experimental evidence.

High-throughput studies necessarily generate large amounts of functional genomic data which must be analyzed to find meaningful associations and extract biological knowledge. Clustering of microarray data produces groups of genes sharing similar expression profiles, whereas chemical genomics experiments may yield sets of genes linked together by virtue of their shared sensitivity to a particular substance exhibited by the corresponding deletion mutant strains. In many cases, genes of unknown function are linked to genes with well established functions, providing an opportunity to glean new insights into these uncharacterized genes. For this reason, it is essential to include biological information about genes and gene products during the analysis in an attempt to infer the function of all linked genes. The challenge is to determine whether the annotation associated with the set of genes, or some part thereof, might help to explain the observed behaviour. Biologists often start their search by collating the annotation provided about genes found in the Saccharomyces Genome Database (SGD) [17], manually comparing them for similar features, and reducing the list to those that are thought to be significant. Although functional analysis using the Gene Ontology (GO) term finder [18] may be useful to identify common annotations between a few genes, it quickly becomes tedious and error prone to do such analysis when increasing numbers of genes are involved. Moreover, due to the hierarchical nature of the ontology, numerous efforts have been made to cluster and statistically evaluate functionally related or differentially regulated genes using GO terms [19-25].

In addition to GO annotation, there is an increasing desire to combine related information such as domain conservation, molecular and genetic interactions, complexes, pathways, phenotypes, and literature references. Determining commonalities among several genes across such diverse features is time consuming and difficult to assess. FunSpec [26] integrates a diverse set of features with a simple method of statistical assessing significance. Unfortunately, the approach used by FunSpec falls short when applied to the analysis of synthetic genetic interaction or chemical genomics experiments. These experiments typically use gene deletion arrays in which strains containing deleted essential genes result in non-viable phenotype (with the exception of when used in an synthetic rescue experiment). Taking this fact into account will raise the statistical significance of features shared by some set of

genes in synthetic lethal or chemical genomics experiments, but will never be observed for essential genes who may in fact share the feature. FunSpec also uses the full list of GO terms from the Gene Ontology Consortium and the MIPS Comprehensive Yeast Genome Database [27] without doing the more sophisticated analysis that is required, yielding lower levels of significance following statistical correction for multiple tests. Finally, FunSpec does not appear to have been updated since 2002, and many new features have since been described. Another resource, DAVID [28], provides the ability to compare against a user-provided background, thereby improving statistical outcomes, but unfortunately does not include a comprehensive set of identifiers and features for yeast functional genomics research.

In this paper, we describe Yeast Features (YF), a web-based tool for the identification of statistically significant features shared by yeast genes identified via functional genomics experiments, with statistical corrections with the addition of prior knowledge. A condensed set of GO terms is used which results in more significant and interpretable results. We apply the YF analysis tool to two chemical genomics experiments and report novel findings that suggest the functional roles of the indirect targets of an aminoglycoside with a known mechanism of action, and the targets of an herbal extract with a previously unknown mode of action.

Implementation

Software Design

The web-based Yeast Features was implemented using the PHP 5 programming language [29] coupled with a MySQL 5.0.25 [30] database backend. The application accepts as input a list of gene names, systematic (ORF) names or SGD identifier. The gene names and SGD IDs are identifier is converted into corresponding ORF names for further processing. This is done because not all ORFs have gene names, some ORFs have multiple gene names, and almost all sources of annotation data contain the ORF name. The application retrieves and collates the set of observed features for each ORF and computes the observed frequency for each feature. The statistical significance of each feature is then assessed via a hypergeometric distribution (see below for details) by comparing the feature frequency in the input set and a pre-computed 'global frequency'. This global frequency may be affected by the use of a background data set, such as the essential genes. Once the probability has been calculated, the program returns the list of features ranked by p-value, after having been filtered according to user options such as a requirement that some set of genes/protein having been annotated with the feature. The resulting table may be sorted by the number of input proteins that share the feature, or by the total number of proteins having that feature in the entire yeast proteome. Separate tabs are available to categorically view features or to view the full features of each protein in the query. Finally, users can browse the features and view which proteins have been annotated with the feature, from which they can select some for comparison.

Data

Yeast chromosomal features, GO complex data, GO slim annotation, interactions, complexes, pathways, domains, phenotypes and publications are formatted as tab-delimited files were obtained from the SGD FTP site [31]. Scripts were written to properly format the gene/protein interaction and complex data due to their non-standard format so that they could be directly imported into the MySQL tables.

The Hypergeometric Distribution

In addition to providing the user with a list of features that a group of proteins may share, it is also important to give some measure of how likely this grouping is. For example, the fact that a group of proteins shares a feature such as 'molecular function unknown' is not unlikely given that ~25-30% of all ORFs share this feature. Instead, statistically unusual features should be highlighted together with a quantitative measure of likelihood. Therefore, the P-value for each feature present in the set of proteins is computed using the hypergeometric distribution. This distribution is appropriate here since proteins are effectively being sampled 'without replacement' – i.e. if there are only 5 proteins that form a particular complex, it is impossible to observe more than 5 such proteins in any given protein set. The hypergeometric distribution measures the probability that a protein set of size *n* will contain exactly *k* proteins sharing a specific feature, given that there are a total of *K* such proteins in the entire proteome of size *N*. As noted below, for chemical genomic data *N* may be optionally reduced to the set of all genes which may be safely deleted without rendering the organism non-viable. The probability mass function is defined as follows:

$$p(k;N,K,n)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \qquad \text{Eq 1}$$

This distribution is widely used in functional enrichment analysis (e.g. for the identification of sumoylation sites [32] and for clustering gene expression data [19]). The P-value reflects the likelihood of randomly observing a given enrichment level, or one more extreme. The P-value is computed as follows for over-represented features.

$$P-value(k;N,K,n)=\sum_{i=k}^{\min(n,K)} p(i;N,K,n) \qquad \text{Eq 2}$$

The Bonferroni Correction

The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously [33]. While a given confidence interval, or alpha value α, may be appropriate for any one individual comparison, it is not appropriate for the set of all comparisons. In order to reduce the likelihood of false positives, the alpha value needs to be lowered to account for the number of comparisons being performed. The simplest and most conservative approach is the Bonferroni correction, which sets the alpha value for the entire set of *n* comparisons by setting the alpha value for each comparison equal to α/n. The default alpha value set in this application is 0.05, but can be changed by the user.

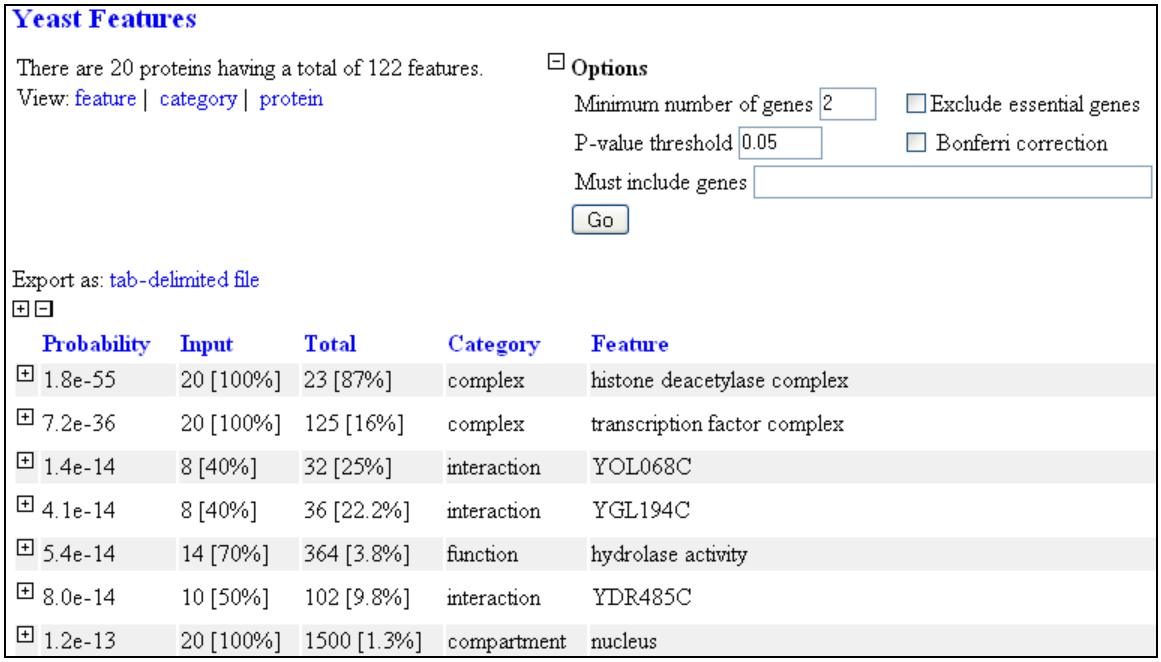
Results and Discussion

Yeast Features identifies the most statistically significant features shared by a set of yeast proteins. A total of 18,786 features were drawn from the SGD and include the GO slim collection (molecular function, biological process, and cellular compartment), conserved domains, interactions, complexes, pathways and publications. This comprehensive set of features exceeds most other applications and

provides a substantially more intuitive software interface, making it a valuable resource for yeast biologists. Our application estimates the significance of shared features from a hypergeometric distribution, a typical strategy for this kind of analysis. The benefit of this approach is that interesting features are highlighted while reducing the significance of common features (e.g. nucleus localization). This allows the biologist to focus on those features that are most likely to explain the common phenotype observed among the protein set, be it chemical sensitivity or gene expression profiles.

Assessing the Significance of Shared Features

An important goal of this project was to design an easy to use, uncluttered graphical user interface. Users can either a) browse the full set of features and their associated genes, or b) find the set of common features between genes by providing their systemic names (yeast open reading frame names), SGD identifiers, or official gene names. Several options are available to modify the search depending on the nature of the data to which the YF tool will be applied. In particular, the Bonferroni correction may be applied to the computed significance scores to account for the number of tests being implicitly performed by YF when a protein set is entered (see Methods). Several other options are available to minimize the number of biologically uninteresting results. One option allows the user to specifically require that a minimum number of genes must share each of the returned features. The default minimum is two, so as to remove features exhibited by only a single gene. The user also has an option to ensure that reported features contain one or more specified genes. By applying this option, it is possible to significantly reduce the number of outcomes, thereby reducing the possibility of a so-called “fishing expedition”, which requires the Bonferroni correction for multiple outcomes. The last option is to exclude essential genes when computing statistical significance, an important consideration for experiments that use deletion arrays (discussed below). With a user provided set of genes and the selected options, YF returns a rank-ordered list of significant features, which can be categorized by the type of feature (i.e. interaction, complex, etc). The list of features and the list of proteins and all of their associated features may be exported in a file for use with any spreadsheet application. YF’s simple interface in combination with novel options to reduce the extent of embarking on a so-called “fishing expedition”, results in a productive environment to identify the most statistically significant and biologically relevant features.



Yeast Features

There are 20 proteins having a total of 122 features.
View: feature | category | protein

Options
Minimum number of genes 2 Exclude essential genes
P-value threshold 0.05 Bonferri correction
Must include genes
Go

Export as: tab-delimited file

Probability	Input	Total	Category	Feature
1.8e-55	20 [100%]	23 [87%]	complex	histone deacetylase complex
7.2e-36	20 [100%]	125 [16%]	complex	transcription factor complex
1.4e-14	8 [40%]	32 [25%]	interaction	YOL068C
4.1e-14	8 [40%]	36 [22.2%]	interaction	YGL194C
5.4e-14	14 [70%]	364 [3.8%]	function	hydrolase activity
8.0e-14	10 [50%]	102 [9.8%]	interaction	YDR485C
1.2e-13	20 [100%]	1500 [1.3%]	compartment	nucleus

Figure 1 - The most significant features for the 20 members of the histone deacetylase complex.
Listed at the top is the number of proteins in the analysis and the total number of shared features, after applying the various criteria. The applicable criteria include the minimum number of genes that must share a feature, the p-value threshold to filter the list, the Bonferri correction for multiple tests, the genes that have each feature, and the exclusion of essential genes to properly assess the significance for gene deletion experiments. The table of features includes the hypergeometric probability, the number of input proteins that share the feature, the total number of proteins in the organism that have this feature, the feature category and the feature name. A user may sort the results by feature, category or by each individual protein. An expand/collapse feature provides a way to view any one of the features in greater detail. Users can obtain these results by exporting the list as a tab-delimited file.

Example: Histone Deacetylase Complex

Once the user enters the input set of proteins and options, the YF program identifies the set of common features and computes the statistical significance of observing them together. As an example, Figure 1 shows the shared features for 20 proteins annotated by SGD as part of the histone deacetylase complex. We see that the most significant feature affirms what was previously known about these proteins mainly that they are part of the histone deacetylase complex. This small p-value indicates that this feature is highly significant because there is only a 1 in $1.8e^{-55}$ chance that the 20 yeast proteins would share this annotation by chance alone. Other significant features shared by all proteins are that they are involved in the transcription factor complex (p-value $7.2e^{-36}$), and are localized in the nucleus (p-value $1.2e^{-13}$). Although the YOL068C (HST1) interaction feature is shared by only 8 proteins, it has a probability of $1.4e^{-14}$, which is more significant than that for localization to the nucleus. The reason for this is that there are significantly more proteins (1500) with the nuclear localization feature as compared to the 32 that interact with HST1 (Figure 2).

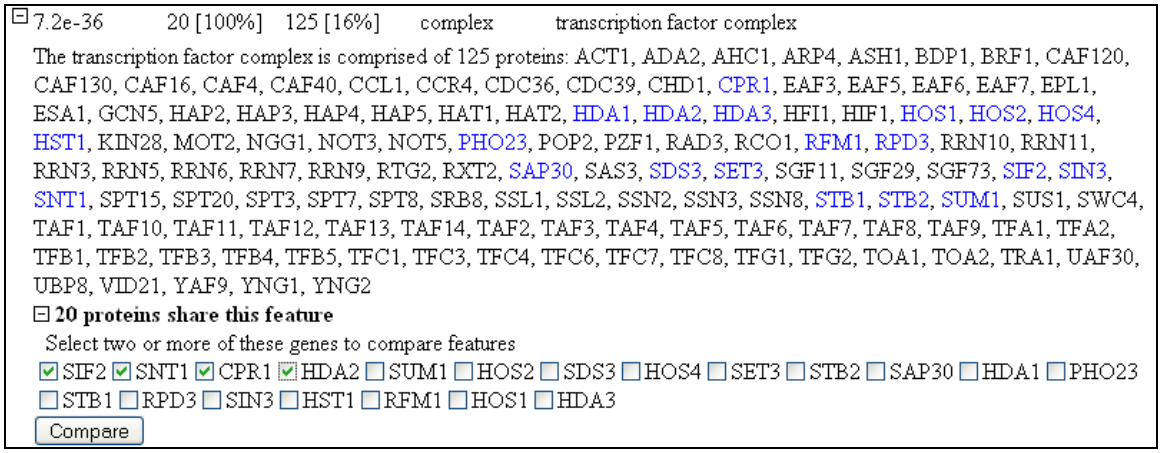


Figure 2 - A detailed look at the transcription factor complex feature.

Using the set of proteins forming the histone deacetylase complex, we see that 20 of the 20 input set of proteins (highlighted in blue) are part of the 125 proteins that comprise the transcription factor complex. Users may also create a smaller subset of the original search by checking the genes and submitting the form.

Using GO Slim to Simplify the Gene Ontology

The Gene Ontology currently makes available ~20,000 terms for the annotation of biological process, cellular compartment and molecular function. The hierarchical nature of the ontology presents a challenge when trying to determine the significance of a shared subset of GO terms. Various applications are devoted to the comprehensive analysis of the full set of GO terms, but the interpretation of results is often aided by reducing or filtering the set of terms to a more manageable set. In a similar line of reasoning, we made use of the species-specific Yeast GO slim terminology maintained by SGD. GO slim collections are reduced term subsets of the GO ontology (79 terms) that broadly cover the ontology without highly specific terms. We find that this collection is of suitable granularity to provide insight into the shared function of an input set while removing excessively specific, redundant or altogether too rare feature categories. One particularly unique feature is that users can browse the terms in each of the categories and select proteins having that annotation to perform the shared feature analysis. For instance, one could ask the question “what complexes are significantly formed by proteins annotated with ‘motor activity’ “, and obtain the answer the kinesin and dynein complexes.

Accurate Evaluation of Gene Deletion Experiments

Functional genomics experiments that utilize gene deletion arrays to identify synthetic sensitive or synthetic lethal conditions will report incorrect p-values on shared features. The reason for this is that essential genes (i.e. those genes that exhibit a lethal phenotype upon deletion) will never be probed for sensitivity, and as such will never form part of the input set of proteins. Yeast Features can optionally correct for this by removing all essential genes from the proteome when computing statistical significance (effectively reducing K and N in Equation 1). Application of this technique results in increased statistical significance of features shared by essential genes. Over one quarter of all features (4887) is associated with essential genes. 3153 features are solely shared by essential genes, and are therefore excluded from consideration in the corrected analysis. For the 1734 features shared by both essential and non-essential genes (Table 1), their statistical significance is reformulated. Thus,

this particular correction may have important consequences for the evaluation of statistical significance for functional genomics experiments that make use of gene deletion arrays.

Category	#Features[a]	Selected Feature[b]	All Genes[c]	Ess. Genes[d]	Change[e]
pathway	48	Fatty acid biosynthesis	12	8	75%
complex	98	Small nucleolar ribonucleoprotein complex	58	44	76%
domain	1511	PF00560: Leucine rich repeat	43	33	77%
function	22	Nucleotidyltransferase activity	73	47	64%
compartment	22	Microtubule organizing center	45	26	58%
process	33	Ribosome biogenesis and assembly	167	127	76%

Table 1 Categorical Summary and Selected Features Having Essential Genes

[a] Number of features shared by essential and non-essential genes

[b] Example features that had at least 4 genes remaining after removal of essential genes

[c] Total number of genes, including essential genes

[d] Number of essential genes having feature

[e] Percent change in the total number of genes having feature, excluding essential genes

Rapid Assessment of Function for Chemical Genomics Experiments

Paromomycin is an aminoglycoside that is known to interact with the protein synthesis machinery of the cell to assert its antimicrobial activity [34]. More specifically, this drug binds to the small ribosomal subunit and causes defects in protein synthesis fidelity. To investigate the gene deletions that cause hypersensitivity to this drug we screened the entire yeast non-essential gene deletion array (~ 4000 gene deletion strains) with sub-minimal inhibitory concentration (sub-MIC) of paromomycin and compared their growth phenotype with those grown in the absence of the drug, as described in [35]. Among the deletion strains, we selected the top 5% that showed increased sensitivity to this treatment for further studies. We then analyzed the selected genes for their common features using YF. As expected from the molecular mechanism of the activity of paromomycin, it was observed that of the 45 genes that have been placed by YF into functional complexes, 25 were grouped as being involved in protein synthesis and its machinery. Other complexes that also came up in this study belonged to the processes of transcription (14 genes) nuclear pore structure (2 genes) and mitochondria membrane structure (2 genes). These complexes may represent alternative pathways, which are targeted by paromomycin, and hence may explain the side effects associated with this drug. Altogether these observations confirm the ability of YF to find the common features between proteins, and that the data analyzed by YF can be directly applied to biological investigations.

Investigation of a Functionally Unknown Protein

One benefit of using YF is to gain insight about the proteins that are interesting as a result of the biological experiment, but have no functional characterization. For

example, YKL075C is a previously uncharacterized ORF that, when deleted, confers hypersensitivity to paromomycin. Because of the known mechanism by which paromomycin affects the cell, we reasoned that YKL075C might be a novel gene that plays a role in the process of protein synthesis. To examine this possibility we performed ribosome profiling analysis as in [36]. If YKL075C is a protein synthesis related gene then we may expect that, based on its molecular function, the deletion of this gene may alter the distribution of ribosomal subunits within a cell. Interestingly we observed that deletion of YKL075C reduced the level of polysomes, relative to free subunits, by about 15%, suggesting a role for YKL075C in protein synthesis. Further investigations are required to fully characterize the molecular function of YKL075. However, given that approximately 25 – 30% of all ORFs in sequenced genomes are of unknown function, this example demonstrates the value of YF in providing insight into the roles of such genes.

Determining Mode of Action of New Antibiotics

Plants are a rich source of medicinal compounds due to a host of secondary metabolites and the activity of these bioactive chemicals can be elucidated using gene array technologies. As an example, Echinacea is one of the top commercial herbal medicines and is employed as a general tonic and in the treatment and prevention of colds and flu. Echinacea extracts were also shown to have antifungal and antiviral activities [37, 38]. However, the mode of action(s) in these medicinal applications of Echinacea is poorly understood. In studying the antifungal activity of Echinacea extracts, we used 16 different treatments based on ethanol extracts of roots, flowers or leaves of *Echinacea* plants. Similar to the previous experiments, growth of *S. cereviseae* strains from the entire non-essential gene deletion array were monitored with and without sub-MIC levels of *Echinaceae* extracts incorporated into the media. A set of 23 mutants was identified that were significantly sensitive to five or more extract treatments and further analyzed using YF. The significant pattern to emerge from the YF analysis was that at least nine of these supersensitive deletion mutants appear to have impaired cell wall functions. Table 2 lists five functions ordered by P-value which identify the cell wall-associated processes for each of the nine deletion mutants. For example, the strain deleted for YPR159W (KRE6) is defective in the biosynthesis of the cell wall component β-1,6 glucan and had significantly reduced growth with at least 6 of the 16 Echinacea treatments. Three strains (YDR245W, YMR307W, YLR338W) also exhibit a phenotype that has increased levels of chitin when systematically deleted. This result is not observed using FunSpec. An additional group of four mutants (not listed) with deletions in genes of unknown function may also be involved in cell wall processes since at least two, YPL264C and YPR071W, are similar to integral membrane proteins. Fungi are recognized as a sister group to animals and more distantly related to plants, and the development of compounds that inhibit fungal growth without harm to the plant or animal host is difficult since all three eukaryotic groups have much in common biochemically. One of the defining characteristics of fungi, however, is the structure and makeup of the cell wall, and therefore the development of compounds that target fungal cell walls are attractive since these may offer a high degree of specificity to fungal pathogens. The hypothesis that Echinacea extracts interfere with fungal cell wall function suggested by this YF analysis is therefore of interest and should be further tested.

	Features				
	Cell wall org & syn[a]	Calcofluor sensitivity[b]	Mannosyl-transferase[c]	chitin[d]	budding[e]
Category	GO Process	Phenotype	Complex	Phenotype	GO Process
ORFs					
YPR159W	+				
YER155C	+				
YJR075W	+		+		
YDR245W	+	+	+	+	
YMR307W	+	+		+	
YLR338W		+		+	
YJR073C		+			
YIL140W					+
YIL008W					+
P-value	5.5×10^{-5}	6.6×10^{-5}	1.7×10^{-4}	1.4×10^{-3}	3.9×10^{-3}
P-value rank	1	2	3	4	6

Table 2 - Selected deletion strains supersensitive to Echinacea extracts.

[a] cell wall organization and biogenesis

[b] decreased resistance to calcofluor white (systematic deletion)

[c] α-1,6-mannosyltransferase complex

[d] increased levels of chitin (systematic deletion)

[e] cell budding

Comparison with Other Tools

While many tools exist to evaluate the significance of genes sharing Gene Ontology annotation (see [20] for comparison), Yeast Features considers a compact, informative subset of GO along with numerous other data sets that are simply not available with other applications (Figure 3). The ease of use of this online tool along with novel browsing, and search space restrictions including a background deletion set makes this tool a natural successor to FunSpec. Since YF is solely focused on yeast, it will continue to incorporate useful annotation for this organism as it becomes available. For instance, the anti-cancer drug Methotrexate was used to identify sensitive genes including one of unknown function which subsequently proposed to be involved in the transport of Methotrexate due to the prediction of a transmembrane segment [10]. In the future, we aim to increase the diversity of our features along predicted lines of evidence.

YCR033W (SNT1)
Gene Name: SNT1
ORF Name: YCR033W
SGD Id: [S000000629]
Chromosome: 3
Description: Subunit of the Set3C deacetylase complex; putative DNA-binding protein

compartment (1)
- nucleus (shared by 20 proteins)

complex (2)
- histone deacetylase complex (shared by 20 proteins)
- transcription factor complex (shared by 20 proteins)

domain (9)
function (1)
phenotype (3)
process (2)
publication (14)
interaction (1)
- The SNT1 (YCR033W) protein has 29 interactions:

Experiment	Proteins/Genes
Affinity Capture-MS	YBR103W|SIF2, YCR028C-A|RIM1, YDR155C|CPR1, YDR432W|NPL3, YGL194C|HOS2, YIL112W|HOS4, YKR029C|SET3, YMR273C|ZDS1, YOL068C|HST1
Affinity Capture-Western	YBR103W|SIF2, YDR155C|CPR1, YIL112W|HOS4
Synthetic Growth Defect	YJL168C|SET2
Synthetic Lethality	YAL011W|SWC3, YAL013W|DEP1, YBR231C|SWC5, YDR334W|SWR1, YDR392W|SPT3, YDR485C|VPS72, YGR078C|PAC10, YJL168C|SET2, YLR055C|SPT8, YLR085C|ARP6, YLR200W|YKE2, YML041C|VPS71, YNL107W|YAF9, YOL012C|HTZ1
Synthetic Rescue	YDR155C|CPR1
Two-hybrid	YGR172C|YIP1

Figure 3 - A detailed look at the SNT1 protein entry and its features.
For each protein, a number of features are listed including, the gene name (if available), the standard name, the SGD identifier, any aliases, the chromosome location and a description. What follows is the set of features, organized by category and indicating the number of other proteins in the input set that also share the feature. Note that the interactions are categorized by their experimental method of determination.

Conclusions

Yeast Features is an easy to use, online tool to identify shared features among a set of yeast genes and assess their statistical significance, with new corrections to accommodate prior knowledge about the experimental system. The methodology identifies known functional attributes of genes using several independent lines of information including molecular function, biological process and cellular compartment from the yeast Gene Ontology slim collection, as well as conserved domains, interactions, complexes, metabolic pathways, phenotypes and publications. We demonstrate the utility of this approach via two functional genomics experiments: In one experiment, the chemical perturbation led to assignment of functional roles for associated genes, and allowed us to pursue new avenues for research on functionally unknown genes; the second analysis defined the biochemical activity of functionally unknown compounds, including herbal medicines. As Yeast Futures is expanded to accommodate new background knowledge, we expect that it will continue to provide

important insight into the biological roles of yeast proteins in functional genomics experiments.

Availability and requirements

Project name: Yeast Features

Project home page: http://software.dumontierlab.com/yeastfeatures/

Operating system(s): Platform independent

Programming language: PHP

Other requirements: MySQL 5+, PHP 5+, Apache HTTP server 2.0+

License: Free for academic and commercial users under the GNU Lesser General Public License (LGPL)

Any restrictions to use by non-academics: None

Authors' contributions

MD, AG, MLS jointly conceived of the study, all authors participated in its design, MA and AG provided the paromomycin data, VE and AG provided the ribosome profiling data, MLS and NM provided the *Echinaceae* extract data, MD implemented the software, JRG identified the hypergeometric distribution as a scoring metric, AG determined the requirement for excluding essential genes, MLS provided insight on design and usability, and MD, AG, MLS, and JG drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Matthew Jessulat for excellent technical assistance. We would also like to thank Joanna Biernacka for her expert advice on statistics, and Gary Bader for his excellent suggestions. This research was funded in part by grants from the Natural Sciences and Engineering Research Council of Canada. J. T. Arnason and V. Treyvaud (U Ottawa) provided MLS and NM with Echinacea extracts.

References

1. Friedman A, Perrimon N: **Genome-wide high-throughput screens in functional genomics**. *Curr Opin Genet Dev* 2004, **14**(5):470-476.
2. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*: **Life with 6000 genes**. *Science* 1996, **274**(5287):546, 563-547.
3. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast**. *Nature* 2003, **425**(6959):737-741.
4. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B *et al*: **Proteome survey reveals modularity of the yeast cell machinery**. *Nature* 2006, **440**(7084):631-636.
5. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast**. *Nature* 2003, **425**(6959):686-691.
6. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al*: **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis**. *Science* 1999, **285**(5429):901-906.
7. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H *et al*: **Systematic genetic analysis with ordered arrays of yeast deletion mutants**. *Science* 2001, **294**(5550):2364-2368.
8. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B *et al*: **Functional profiling of the Saccharomyces cerevisiae genome**. *Nature* 2002, **418**(6896):387-391.
9. Baetz K, McHardy L, Gable K, Tarling T, Reberioux D, Bryan J, Andersen RJ, Dunn T, Hieter P, Roberge M: **Yeast genome-wide drug-induced haploinsufficiency screen to determine drug mode of action**. *Proc Natl Acad Sci U S A* 2004, **101**(13):4525-4530.
10. Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP, Davis RW: **Chemogenomic profiling: identifying the functional interactions of small molecules in yeast**. *Proc Natl Acad Sci U S A* 2004, **101**(3):793-798.
11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
12. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z *et al*: **CDD: a Conserved Domain Database for protein classification**. *Nucleic Acids Res* 2005, **33**(Database issue):D192-196.
13. Snyder KA, Feldman HJ, Dumontier M, Salama JJ, Hogue CW: **Domain-based small molecule binding site annotation**. *BMC Bioinformatics* 2006, **7**:152.
14. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
15. **The Gene Ontology (GO) project in 2006**. *Nucleic Acids Res* 2006, **34**(Database issue):D322-326.

16. Auffray C, Imbeaud S, Roux-Rouquie M, Hood L: **From functional genomics to systems biology: concepts and practices**. *C R Biol* 2003, **326**(10-11):879-892.
17. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: **SGD: Saccharomyces Genome Database**. *Nucleic Acids Res* 1998, **26**(1):73-79.
18. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G *et al*: **Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO)**. *Nucleic Acids Res* 2002, **30**(1):69-72.
19. Shah NH, Fedoroff NV: **CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology**. *Bioinformatics* 2004, **20**(7):1196-1197.
20. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B: **GOToolBox: functional analysis of gene datasets based on Gene Ontology**. *Genome Biol* 2004, **5**(12):R101.
21. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005, **21**(16):3448-3449.
22. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S *et al*: **GoMiner: a resource for biological interpretation of genomic and proteomic data**. *Genome Biol* 2003, **4**(4):R28.
23. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data**. *Genome Biol* 2003, **4**(1):R7.
24. Berriz GF, King OD, Bryant B, Sander C, Roth FP: **Characterizing gene sets with FuncAssociate**. *Bioinformatics* 2003, **19**(18):2502-2504.
25. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes**. *Bioinformatics* 2004, **20**(4):578-580.
26. Robinson MD, Grigull J, Mohammad N, Hughes TR: **FunSpec: a web-based cluster interpreter for yeast**. *BMC Bioinformatics* 2002, **3**:35.
27. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V: **MPact: the MIPS protein interaction resource on yeast**. *Nucleic Acids Res* 2006, **34**(Database issue):D436-441.
28. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
29. **PHP** [http://php.net]
30. **MySQL** [http://mysql.com]
31. **SGD Data Download FTP site** [ftp://genome-ftp.stanford.edu/pub/yeast/data_download/]
32. Zhou F, Xue Y, Lu H, Chen G, Yao X: **A genome-wide analysis of sumoylation-related biological processes and functions in human nucleus**. *FEBS Lett* 2005, **579**(16):3369-3375.
33. Shaffer JP: **Multiple Hypothesis Testing**. *Ann Rev Psych* 1995, **46**:561-584.
34. Fourmy D, Recht MI, Blanchard SC, Puglisi JD: **Structure of the A site of Escherichia coli 16S ribosomal RNA complexed with an aminoglycoside antibiotic**. *Science* 1996, **274**(5291):1367-1371.

35. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, Brown GW, Kane PM, Hughes TR, Boone C: **Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways**. *Nat Biotechnol* 2004, **22**(1):62-69.
36. Foiani M, Cigan AM, Paddon CJ, Harashima S, Hinnebusch AG: **GCD2, a translational repressor of the GCN4 gene, has a general function in the initiation of protein synthesis in Saccharomyces cerevisiae**. *Mol Cell Biol* 1991, **11**(6):3203-3216.
37. Binns SE, Purgina B, Bergeron C, Smith ML, Ball L, Baum BR, Arnason JT: **Light-mediated antifungal activity of Echinacea extracts**. *Planta Med* 2000, **66**(3):241-244.
38. Vimalanathan S, Kang L, Amiguet V, Livesey J, Arnason J, Hudson J: **Echinacea purpurea aerial parts contain multiple antiviral compounds**. *Pharmaceutical Biology* 2005, **43**:740-745.