ANALYSIS OF LARGE GENOMIC DATA COLLECTIONS

Curtis Huttenhower

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

COMPUTER SCIENCE

Adviser: Olga G. Troyanskaya

January 2009

Abstract

Modern computational biology draws on the historical strengths both of computer science and of molecular biology. It requires careful attention to algorithmic development, data structures, storage, and manipulation, efficient software engineering, and machine learning; but it also drives towards a deeper understanding of the causes and cures of disease, the organization of life at both micro- and macroscopic levels, and the molecular systems governing every living organism. In particular, the field must take advantage of the ever-increasing availability of experimental assays that provide whole-genome measurements: the sequences of the human and other genomes, the abundances of every transcript in a cell, the distribution of gene activity across cell or tissue types, and the interactions among proteins and protein complexes. These data can be combined and analyzed in an integrated manner to enable biological discoveries not obtainable from single experiments, but this requires both the computational ability to manipulate large, heterogeneous, and noisy datasets as well as the biological ability to ask targeted questions of these diverse data. This manuscript presents four broad solutions to these challenges. First, we discuss the closer integration of computational techniques with laboratory experimentation to study *S. cerevisiae* mitochondria. This confirms that biological discoveries can be made much more efficiently in the laboratory if informed by computational inference and that computational algorithms can be much more accurate if informed by appropriate experimental design. Second, we present a set of specific software tools created to address biological needs, ranging from the efficient analysis of very large data collections to the visualization of dense biological networks. Third, we detail several ways in which statistical models can be applied to genomic data so as to describe specific biological phenomena, including cellular growth rate, aneuploidy, and phosphorylation. Finally, we provide methods for integrating heterogeneous genomic data designed specifically for very large data collections and complex organisms (e.g. human beings); this allows one to study not only traditional biological questions, such as the interactions between individual genes and proteins, but also the interactions between entire pathways and processes at a systems level. These tools, techniques, and discoveries provide a basis from which further computational research can readily develop as biological experimentation explores more data, examines new organisms, and brings us ever closer to understanding and eliminating disease.

To my parents, who don't harass me too much for spending over twenty years in school.

To my sister, for also braving the trek through graduate school.

To Olga Troyanskaya, my adviser, for making it all possible.

To Hilary Coller, for giving me a second laboratory home.

To Dannie Durand, for clueing me in to this whole bioinformatics thing.

To coauthors, collaborators, and friends;
thanking you all enough would double the length of this weighty tome.

To my committee, for having the perseverance to read it.

To Boy and Girl, for moral support.

And, always, to Cécile.

This work was supported by NSF CAREER award DBI-0546275, NIH grant R01 GM071966, NIH grant T32 HG003284, and NIGMS Center of Excellence grant P50 GM071508.

Table of Contents

Large Scale Computational Biology: An Introduction

You are most likely reading this using some combination of the 10^8 cells making up your optic system, in conjunction with a portion of the 10^{11} neurons in your central nervous system (or, possibly, using your millions of auditory or tactile cells instead). These represent only a few of the roughly 200 different cell types into which the 10^{14} cells of the human body can be classified; as you read, some 10^{11} cells of your immune system are fending off perhaps 10^{10} bacteria and other foreign organisms, but allowing the 10^{13} symbiotic bacteria in your digestive tract to continue their work unharmed. Inside each of your cells lies just over three billion pieces of genetic information, encoding the ~25,000 different genes that are used to construct each cell's ~10^9 protein components. And that's just in one of you, not taking into account the other six billion human inhabitants and over 10^{30} other organisms on the planet (Blinkov and Glezer 1968; Janeway, Travers et al. 2001; Schiffman 2001; Sears 2005; Lodish, Berk et al. 2007).

Needless to say, biology - the study of this massive diversity of life - is complicated. Fortunately, this document is being written on a machine containing about 10^9 transistors, running programs that together account for maybe 10^8 lines of code, all of which we understand (more or less) perfectly. Computer science is a field intimately acquainted with the study, management, and organization of complexity, from the engineering of the systems making up a modern processor to the statistics of uncovering informative signals in noisy data. As technological advancements in biology have made us increasingly able to detect, measure, and quantify the activities of the molecular components of life, the field has turned to computer science to process and understand this data. Remarkably, most of these developments have happened only within the last ten years - the first complete genome of a free-living organism (*Haemophilus influenzae*) was sequenced in

1995 (Fleischmann, Adams et al. 1995) - and computer science and biology alike are still growing into their new roles.

The interdisciplinary field that has grown out of these joint efforts is referred to as computational biology or bioinformatics. In some usages, the former refers specifically to biological discoveries made by applying computational techniques and the latter to algorithmic discoveries that happen to employ biological data; this manuscript will use the two terms interchangeably. Ideally, no such distinction is necessary, as the most successful research will advance our knowledge of both scientific areas. It is the challenge of bioinformatics to collect, organize, and understand the ever-increasing variety of data produced by our measurements of the biological world, ranging from population studies in ecological communities to the activities of individual molecules within single cells. Much bioinformatic work, particularly in this manuscript, focuses particularly on molecular biology and on the assessment of subcellular activity. The tiny scale on which the molecular activity of life is carried out makes it difficult to observe, and thus much of computer science's role in computational biology lies in using machine learning to discover meaningful signals in experimental data and in using these to infer the purpose and structure of cellular processes.

This manuscript will discuss four specific areas in which we have contributed to computational biology. First, one of the ways in which the field is still adapting is in the optimization of laboratory investigations; often, experimental work is time-consuming and expensive, while computational analyses - even those requiring sophisticated algorithmic design and complex data processing - can be performed more quickly and cheaply. This implies that using computational predictions to guide laboratory work can greatly increase the efficiency with which new biology is discovered, and, conversely, that by designing experimental assays appropriately, their results

can more easily be incorporated into computational algorithms to improve their accuracy. We verify this hypothesis using yeast mitochondria as an experimental system; in approximately six person-months, we tripled the number of known mitochondrial proteins, incorporated this new knowledge into our computational predictions, and showed that such predictions can be extremely accurate even in areas of biology where our prior knowledge is very sparse (e.g. newly sequenced or previously unstudied organisms).

Second, we have created a variety of software systems to address specific biological questions. The core of these tools is the Sleipnir library, which provides C++ implementations of many algorithms developed to allow rapid processing of large biological data collections. Public data repositories have already grown to contain hundreds of thousands of experimental results, and Sleipnir represents the first set of computational tools capable of integrating and learning from this data efficiently. Other software we discuss includes the Nearest Neighbor Networks clustering algorithm, designed to find genes performing similar functional roles in biological data, and the Graphle application for interactive, collaborative exploration of large biological networks. Finally, we also present COALESCE, a system for integrating many different data types and experiments to infer comprehensive regulatory networks in higher organisms.

Our third section will address ways in which statistical models can be applied to understand biological data and to predict cellular behavior. By applying linear models to measurements of the cellular response to various perturbations (specifically changes in growth rate and chromosomal copy number), we can explicitly catalog which of an organism's genes are activated (or inactivated) under those conditions. Conversely, these models also allow us to predict how a cell perceives its environment if we are given only information about its genetic activity (e.g. if we know which genes are active in some population, we can estimate how quickly the

population is growing). Other statistical analyses applied to data on specific protein modifications reveal how the cell uses those modifications to regulate its activity and, when compared to similar modifications in other organisms, how the regulatory network has evolved over time.

Finally, we will present several ways in which we have leveraged very large biological data collections in order to better understand entire biological systems. In particular, by integrating a substantial portion of all currently available experimental data pertaining to human beings (~30 billion data points), we can make accurate predictions of gene function, interactions, and involvement in human disease. An algorithm for doing this is integrated in the HEFalMp (Human Experimental/Functional Mapper) system. In the organism *S. cerevisiae*, similar data is used by the MEFIT system to understand gene function and the yeast response to dataset-specific environment perturbations. We discuss ways in which large scale genomic data collections such as these can be used to understand the overall molecular and cellular activity in any organism, allowing meaningful maps of biological regulation and interactions to be constructed by integrating thousands of experimental results.

Writing from a computational perspective, we provide a brief overview below of the areas of computer science most involved in computational biology, as well as a more in-depth introduction to the cellular processes and biological assays common in the field and in this manuscript. In addition to detailed descriptions of our specific contributions, we will conclude with several possible future directions for this research and for the field as a whole.

An Overview of Computational Tools

The areas of computer science most intimately related to computational biology are machine learning and related fields: statistics, information retrieval, artificial intelligence, numerical computation, modeling, and computer vision (Aluru 2005). This manuscript assumes a general computational background, so this section will touch briefly on the main terminology and techniques used throughout the subsequent text. For other broad reviews, see (Kitano 2002; Ray, Chong et al. 2002), and for recent reviews particularly relevant to this work, see (Troyanskaya 2005; Markowetz and Troyanskaya 2007; Troyanskaya 2007). Computational biology as a whole is still in its infancy and, as a result, is one of the fastest paced fields in modern science; citations more than ten or even five years old are almost uniformly outdated, and a decade from now, this text may seem equally archaic.

Machine Learning, Classification, and Performance Metrics

Machine learning (Mitchell 1997) is, broadly, the task of constructing algorithms capable of improving their performance (under some evaluation measure) as they are exposed to increasing training information. Many machine learning tasks of bioinformatic relevance are classification problems: given some set of records (each representing an entity of interest using one or more data features), partition the set into two or more classes (some formulations of the problem do not require a strict partitioning and are referred to as multiclass). Classification can be supervised, in which case the target class (i.e. label) for some (or all) of the records is known a priori; or it can be unsupervised, in which case the records are separated based solely on characteristics of their data features. We will briefly review supervised classification here, and unsupervised classification is discussed below in its incarnation in graph clustering algorithms.

Formally, consider a set of records R in which each $r \in R$ is an ordered list of features $(f_1, f_2, ..., f_n)$. Each feature f_i is drawn from some set F_i, which can be categorical (i.e. $|F_i|$ is finite, also referred to as discrete) or continuous ($|F_i|$ infinite, often integral or real valued). One feature is often set aside as the label or dependent variable to be predicted, with the others serving as data or independent variables used during learning; when labels are present and binary, they are often referred to as positive and negative examples. A set of records can contain missing values, in which case not every record possesses a value for every feature, or it can contain errors, in which case some record/feature values are numerically or categorically incorrect. Typically, records are split into a training set from which a classifier is learned and a test set used to evaluate its performance; records from the training set alter the classifier's behavior (presumably improving its performance), but during evaluation of the test set, the classifier becomes a read-only system. Classifier evaluation itself is a complex research area whose details will not be covered here (Mitchell 1997).

Classification algorithms tend to fall into two categories: generative models, which describe the entire feature set of the records R, and discriminative models, which only fully describe the output label as it depends on other features. Probabilistically, this is the difference between generatively modeling the joint distribution $P(F_1, F_2, ..., F_n)$ and discriminatively modeling, for some label F_1, the conditional distribution $P(F_1 | F_2, F_3, ..., F_n)$. In bioinformatics, this distinction often boils down to the difference between modeling the world, i.e. constructing a generative model that describes how a biological system behaves, and predicting some outcome, i.e. constructing a discriminative model that uses observed data to infer some unobserved property.

Bayesian Networks

A generative model that will be used repeatedly in this text is the Bayesian network, one of the class of graphical models (Neapolitan 2004) that capture dependencies between data in terms of conditional probability distributions. A Bayesian network can be represented as a directed acyclic graph $G=(V, E)$ in which each vertex v_i represents an event or random variable. Independence among these events is captured by the edge structure of the graph, such that the joint probability over all vertices $P(v_1, v_2, ..., v_n)$ is equal to the product of all conditional probabilities $P(v_i \mid u_1, u_2, ..., u_m$ such that $(u_j, v_i) \in E)$. In other words, the probability of an event is completely described by the values of its parent events, independently of all other events in the graph.

Bayesian networks are Bayesian in the sense that Bayes' theorem is central to their definition and implementation:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

That is, the posterior probability of some event A given some evidence B is proportional to the product of $P(B \mid A)$ and the prior probability of A. For example, the prior probability of a fire in one's apartment is very low; however, the posterior probability of a fire given that the smoke alarm is going off is substantially higher. Conversely, it is much more likely that one burns dinner than that one's apartment is on fire, an event that can also activate the smoke alarm. These relationships and probabilities are summarized in the example Bayesian network in Figure 1. Given evidence for zero or more of the events in a Bayesian network, the probability of all other events can be inferred (in a process referred to as Bayesian inference) using Bayes' theorem. For example, the probability of the smoke alarm going off under any circumstances is:

$$P(SA) = P(SA \mid F, BD)P(F, BD) + P(SA \mid F, \sim BD) + P(SA \mid \sim F, BD) + P(SA \mid \sim F, \sim BD) =$$

$$0.999 \cdot 0.001 \cdot 0.1 + 0.95 \cdot 0.001 \cdot 0.9 + 0.9 \cdot 0.999 \cdot 0.1 + 0.001 \cdot 0.999 \cdot 0.9 = 0.0918$$

Suppose that one hears the smoke alarm go off while visiting one's neighbor. The probability that your apartment is in trouble is:

$$P(F \mid SA) = P(SA \mid F)P(F)/P(SA) = [P(SA \mid F, BD)P(BD) + P(SA \mid F, \sim BD)P(BD)]P(F)/P(SA) =$$

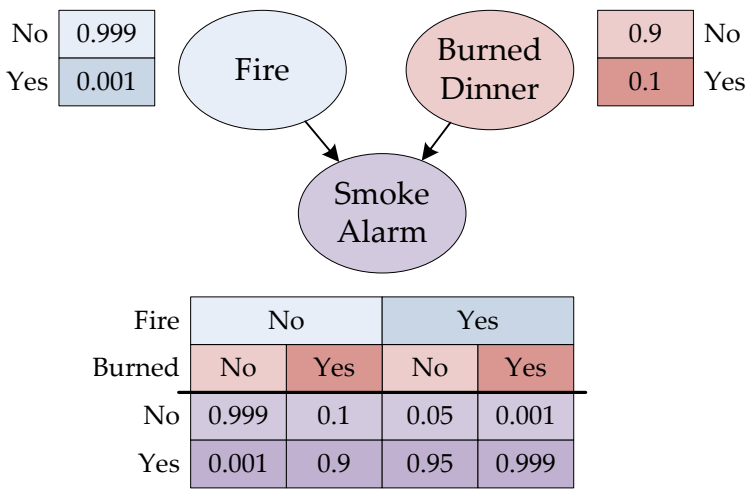
$$(0.999 \cdot 0.1 + 0.95 \cdot 0.9) \cdot 0.001 \,/\, 0.0918 = 0.0104$$

Sadly, the probability that your dinner is in trouble is in this case $P(BD \mid SA)=0.981$. Fortunately, however, one can be reassured that the smoke alarm is "explained" by one's dinner being burned if this is later discovered to be true:

$$P(F \mid SA, BD) = P(SA, BD \mid F)P(F)/P(SA, BD) = P(SA \mid F, BD)P(BD)P(F)/[P(SA \mid BD)P(BD)] =$$

$$0.999 \cdot 0.1 \cdot 0.001/((0.999 \cdot 0.001 + 0.9 \cdot 0.999) \cdot 0.1) = 0.00111$$

These calculations remain essentially identical for non-binary discrete probability distributions, and they can be extended similarly to linear combinations of normal distributions in continuous Bayesian networks (Fu and Tsamardinos 2005).



Fire	No		Yes	
Burned	No	Yes	No	Yes
No	0.999	0.1	0.05	0.001
Yes	0.001	0.9	0.95	0.999

Figure 1: Example Bayesian network capturing three binary random events. A fire might occur with probability P(F)=0.001, and one's dinner might be burned with probability P(BD)=0.1; either of these events influences the probability of the smoke alarm being activated, P($SA \mid F$, BD). Probabilities are, hopefully, illustrative and not an accurate reflection of reality; example inspired by (Pearl 1988).

This type of probabilistic explanation is formalized by the concept of a Markov blanket (Pearl 1988). A node v's Markov blanket consists of its parents, children, and children's parents; any other set of nodes in a Bayesian network is conditionally dependent of v given the values of events in its Markov blanket. This is also referred to as d-separation, in that if two sets of nodes X and Y are conditionally independent given a third set Z, they are said to be d-separated by Z. While exact Bayesian inference (and many types of approximation) are known to be NP hard (Cooper 1990), this type of decomposition has led to a variety of efficient algorithms taking advantage of network structure, decomposition, message passing, parallelism, and approximation (MacKay 2003). This ability to rapidly estimate one or more event probabilities given any amount of prior evidence has contributed greatly to the utility of Bayesian networks in computational biology.

Another important factor in the use of Bayesian networks for bioinformatics is the ease with which their structure and parameters can be learned from data (Neapolitan 2004). Given a predefined network structure, one can consider a set of records R in which features correspond to nodes in the Bayesian network. If these records are discrete and contain no missing values, then we can obtain maximum likelihood estimates for the network parameters (also referred to as conditional probability tables or CPTs) by counting: $P(v_i=x_i|u_1=y_1, u_2=y_2, ..., u_m=y_m)$ is simply the number of records containing $(x_i, y_1, y_2, ..., y_m)$ divided by the number containing $(y_1, y_2, ..., y_m)$. These estimates can be further modified by techniques such as Laplace smoothing (Jurafsky and Martin 2008). If the records do contain missing values, maximum likelihood estimates can still be obtained using the Expectation Maximization (EM) algorithm, which iteratively infers probability distributions (expectations) over the missing values using the network's current parameters and then updates (maximize the likelihood of) these estimated parameters using the resulting distributions (Neapolitan 2004). This is guaranteed to converge to a local maximum (since no one iteration can decrease the probability of the observed data), but it is thus sensitive to initial conditions (i.e. initial parameter values) and to the computational complexity of calculating the maximum likelihood values. In the case of Bayesian networks, this depends multiplicatively on the maximum in-degree of the network's nodes and can thus be a significant factor.

Bayesian structure learning is, in general, a much more difficult problem due to the tremendous size of the search space: given a set of event vertices V, the number of possible graphs is superexponential in $|V|$. The first class of algorithms dealing with this problem (Rebane and Pearl 1987) begin by establishing undirected dependence between variables (i.e. determining that two events are joined by an edge without assigning directionality to it) by statistical tests for significant mutual information. Directionality (i.e. causation) is assigned by, first, considering all

triplets of connected variables and directing pairs of edges so as to preserve the observed conditional independencies; remaining undirected edges are directed so as to avoid cycles. A second class of algorithms instead treats the problem explicitly as a search, moving through the space of possible graph structures (and parameter settings) by making small, directed modifications and evaluating the likelihood (or other scoring function) of the resulting network (Hastie, Tibshirani et al. 2001). While Bayesian structure learning has been successfully employed in computational biology (Sachs, Perez et al. 2005), it often suffers from a lack of appropriate data (i.e. noise) and from intractability when many variables are involved, and it will not be a focus of this manuscript.

Instead, we take advantage of a special class of Bayesian networks that are particularly suited for large, noisy data collections: naive Bayesian networks. Also referred to as naive Bayes classifiers, these comprise the class of graph structures in which a single root or class node is the only parent of all other nodes (the data or feature variables), i.e. $G=(V, E)$, $V=\{c, f_1, f_2, ..., f_n\}$, $E=\{(c, f_1), (c, f_2), ..., (c, f_n)\}$. This is thus a direct encoding of a supervised classification problem in which all data is assumed to be independent - an assumption that is almost never strictly true but which, for various potentially defensible reasons, performs remarkably well in practice (Hand and Yu 2001). Learning the maximum likelihood parameters of such a model is efficient even in the case of missing data (requiring only counting and no iteration), and regression models have even been proposed which coopt the Bayesian structure to provide maximum a posteriori (i.e. discriminative rather than generative) parameter estimates (Greiner and Zhou 2005). Likewise, inference given any combination of observed or missing data is a simple multiplication $P(c|f_1, f_2, ..., f_n) \propto \prod_i P(f_i|c)$. Naive Bayesian classifiers are thus similar in spirit, if not mathematics, to linear

models as discussed below, and they represent the primary generative model employed in this text.

Support Vector Machines

A canonical discriminative classification algorithm is the support vector machine or SVM (Vapnik 1998). SVMs are typically discussed as binary classifiers over one or more real valued data features, but multiclass and continuous algorithms also exist, as well as means of dealing optimally with categorical data (Schölkopf and Smola 2002). SVMs integrate a number of concepts, but in the simplest case reduce to a problem of maximum margin linear classification. As pictured in Figure 2, consider a number of positive and negative training examples in some high dimensional, real valued space. These can be separated by infinitely many different hyperplanes (each with parameters equal to the dimensionality of the space), but only one such hyperplane will do so while also maximizing the distance to the nearest examples. Thus, consider a set of training records $R=\{(c_1, \boldsymbol{d}_1), ..., (c_n, \boldsymbol{d}_n)\}$ where $\boldsymbol{d}_i=(d_{i,1}, ..., d_{i,m})$, $c_i \in \{0, 1\}$, and $d_{i,j} \in \Re$. A hyperplane in this m-dimensional space can be described by all $\boldsymbol{x}$ such that $\boldsymbol{w} \cdot \boldsymbol{x} - b = 0$, where $\boldsymbol{w}$ is a perpendicular vector specifying the hyperplane's orientation and b indicates its offset from the origin. The boundaries of the space separating the training examples can in turn be described by parallel hyperplanes $\boldsymbol{w} \cdot \boldsymbol{x} - b = 1$ and $\boldsymbol{w} \cdot \boldsymbol{x} - b = -1$; to maximize the intervening margin $2/|\boldsymbol{w}|$, we thus minimize the norm $|\boldsymbol{w}|$. To exclude training examples from the intervening space, we constrain $\boldsymbol{w} \cdot \boldsymbol{d}_i - b \geq 1$ if $c_i=1$ and $\boldsymbol{w} \cdot \boldsymbol{d}_i - b \leq 0$ if $c_i=0$, which taken together forms a constrained optimization problem that can be solved using quadratic programming (Vapnik 1998).

Further extensions are required if the data are not linearly separable, i.e. to correctly deal with misclassification errors when they are unavoidable. This can be addressed (Cortes and Vapnik 1995) by introducing a soft margin parameter C that balances the optimization between margin

maximization and misclassification by minimizing $|\boldsymbol{w}|^2/2 + C\sum_i \varepsilon_i$, where ε_i is zero if r_i is correctly classified and equal to the distance between it and the hyperplane if it is not. Similarly, the method was limited to linear classification until the introduction of kernel methodology to transform the problem space (Boser, Guyon et al. 1992). This replaces the optimization dot products $\boldsymbol{w}\cdot\boldsymbol{d}_i$ with any positive semi-definite kernel function $k(\boldsymbol{w}, \boldsymbol{d}_i)$, which can be shown to be equivalent to a dot product of two vectors $\varphi(\boldsymbol{w})\cdot\varphi(\boldsymbol{d}_i)$ for some function φ (which does not have to be explicitly known to the SVM) (Aizerman, Braverman et al. 1964). Commonly used numerical kernels include polynomials $(\boldsymbol{w}\cdot\boldsymbol{d}_i)^d$ and Gaussian radial basis functions $\exp(-0.5|\boldsymbol{w} - \boldsymbol{d}_i|^2/\sigma^2)$, while custom kernels specific to biological data have also been used in bioinformatics (Lanckriet, Deng et al. 2004).

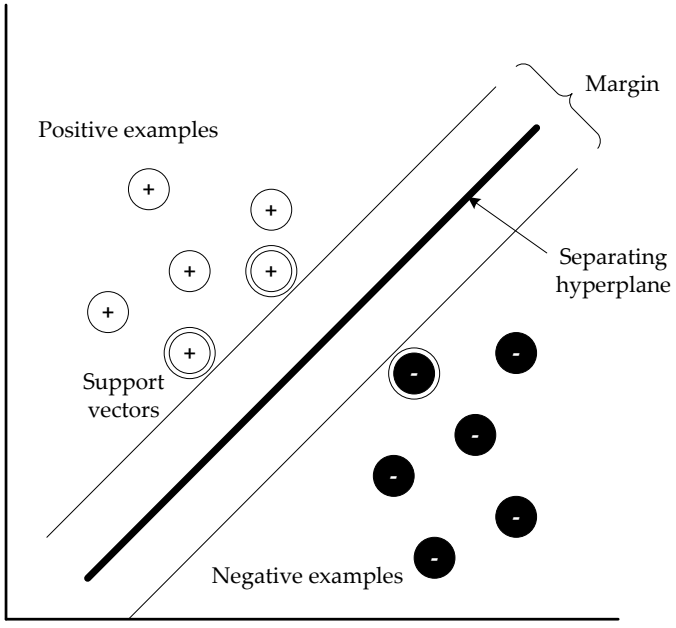


Figure 2: Schematic diagram of a support vector machine (SVM). The most basic SVM is a hyperplane separating sets of positive and negative examples in some high dimensional, real valued space. The hyperplane is chosen to maximize the margin between the nearest training examples, referred to as the support vectors.

Performance Metrics

Performance evaluation techniques in computational biology are essentially a combination of those found in information retrieval (Jurafsky and Martin 2008) and in medical informatics (Altman and Bland 1994). Supervised techniques rely on a "gold standard" collection of example records labeled with a known class; predictions can then be made on held out, unlabeled examples, and the predicted labels compared with the gold standard. This is most easily considered in the binary case, where all labels are either positive (one) or negative (zero). In this situation, any set of predictions can be partitioned into four sets: true positives (records predicted to be positive and labeled as such in the gold standard), false positives (records predicted to be positive but labeled as negative), true negatives (predicted and labeled negative), and false negatives (predicted negative but labeled positive). Given some set of predictions, the fraction of correct labels is the precision (also referred to as positive predictive value):

$$precision = \frac{TP}{TP + FP}$$

The recall is the fraction of all positive labels included in the predicted set:

$$recall = \frac{TP}{TP + FN}$$

Typically, predictions are generated in some rank order, e.g. decreasing confidence or probability. This allows a precision/recall curve to be constructed by moving a cutoff through this rank order and calculating the precision and recall for each induced partition, ranging from high confidence/low recall/high precision to low confidence/high recall/low precision; for an example, see Figure 17. In some fields, recall is referred to as sensitivity, and is generally paired with specificity (in place of precision) as a measure of accuracy:

$$specificity = \frac{TN}{TN + FP}$$

Specificity thus represents the fraction of incorrect labels that are excluded; while analogous to precision, the two measures can behave very differently as the ratio of positive to negative training examples varies. Like precision and recall, sensitivity and specificity are often plotted as a curve over a sliding rank cutoff. This is referred to as a Receiver Operating Characteristic (ROC) curve (Zweig and Campbell 1993), and the area under such a curve, or AUC, is often used as a summary statistic for overall performance over an entire ranked list of predictions. This can be calculated analytically using the Wilcoxon rank-sum test (also referred to as the Mann-Whitney U or Mann-Whitney-Wilcoxon test (Wilcoxon 1945; Mann and Whitney 1947)); for examples, see Figure 23, Figure 24, Figure 39, etc. Rarer variants of these metrics include the area under a precision/recall curve, AUPRC, consisting of the integral over a precision/recall curve just as AUC integrates over a sensitivity/specificity curve, and the log-likelihood score (LLS), consisting of a logarithm of the ratio of actual to random predictive performance:

$$LLS = \log\left(\frac{TP / FP}{(TP + FP)/(TN + FN)}\right)$$

Linear Models

A supervised statistical tool that we employ to model biological measurements is the linear model (Davidson 2003), consisting of a vector of observations $\boldsymbol{y}$, a matrix of observed values X, a vector of unknown parameters $\boldsymbol{b}$, and a vector of residual errors $\boldsymbol{\varepsilon}$:

$$\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{\varepsilon}$$

X typically (but not always) includes at least one constant term (e.g. a column of ones) and one additional variable, in which case this clearly models a simple linear regression. A full linear model thus predicts a set of output variables as a sum of linearly weighted input variables, with the attractiveness of the model arising from its simplicity, the frequency of linear relationships in natural data, and the ease with which unknown parameters $\boldsymbol{b}$ can be estimated algorithmically. Under the assumption of independent errors $\boldsymbol{\varepsilon}$ normally distributed about zero, the maximum likelihood parameters $\boldsymbol{b}$ can be estimated using the method of least squares. A related calculation that will not appear here, but which is frequently useful in computational biology, is the analysis of variance (ANOVA), which partitions the variability of $\boldsymbol{y}$ into components correlated which each component of X using equivalent techniques.

In a biological setting, such a model is useful (when applied to appropriately linear data) for two reasons. The first is descriptive: if the components of $\boldsymbol{y}$ represent biological measurements and the components of X experimental conditions, the inferred parameters $\boldsymbol{b}$ provide estimates of the relationship of each measurement to the environment. If these measurements correspond, for example, to protein activity within the cell under various environmental perturbations (e.g. chemical exposure, nutrient availability, temperature, disease, etc.), this can establish strong evidence that proteins with high magnitude parameters under some conditions are specifically responsible for cellular tasks relating to those conditions. All proteins with large $\boldsymbol{b}$ under heat shock, for example, might catalog the cell's machinery for responding to high temperatures. Conversely, a learned linear model can also be predictive of a protein's activity (or other observed quantity) under novel conditions. In other words, by learning such a model $\boldsymbol{b}$ from experimental measurements, we can then extrapolate how we expect a cellular system $\boldsymbol{y}$ to respond for new values of the environmental conditions X.

Graphs and Clustering

The usage of graph structures in mathematics, computer science, and biology varies so widely and with such extensive histories in each field that we will give only a cursory introduction here; for details, see (Diestel 2005), (Cormen, Leiserson et al. 2001), and (Junker and Schreiber 2008), respectively. Briefly, a graph or network (terms which this document will use interchangeably) $G=(V, E)$ consists of a set of nodes or vertices V and a set of edges E; in most cases (and all cases in this document), each pair of nodes has at most one edge connecting them, making $E \subseteq V \times V$. Likewise, although graphs can formally allow self connections (i.e. $e \in E$, $v \in V$, $e=(v, v)$), these will generally not appear in this manuscript (although they can be relevant in biological networks, e.g. for self-regulation). Graphs can be directed, in which case an edge (v, u) implies a directional relationship between nodes v and u and is distinct from (u, v); or they can be undirected, in which case these two edges are equivalent (or, equivalently, $(v, u) \in E$ implies $(u, v) \in E$, or edges can be represented as sets $\{u, v\}$). Graphs can also be weighted, in which case each edge possesses some numerical weight $w(e)$.

In computational biology, graphs are typically used to represent gene or protein interaction networks: each vertex represents a protein (or gene), and each edge represents some type of relationship between the proteins: physical binding (i.e. do those two proteins ever directly interact in the cell), regulation, sequence similarity, etc. Particularly in unweighted protein-protein interaction networks, this has led to a focus on network topology. Defining the degree of a vertex as the number of incident edges (i.e. $d(v)=|\{e \in E \mid e=(v, u) \text{ or } e=(u, v), u \in V\}|$), one can imagine a random graph in which edges are placed by selecting nodes at random (Erdos and Renyi 1960), resulting in a Poisson degree distribution with few outliers. Interestingly, most "natural" networks from biology (and other fields) do not appear to behave this way; instead,

their distribution of degrees follows a power law with a "long tail" of high-degree nodes (Barabasi and Oltvai 2004). These high-degree vertices are referred to as hubs, and in biology can represent, for example, proteins that act to coregulate diverse processes or to tie together multiple cellular components. Graphs of this class are in turn referred to as scale-free or as having the small-world property (i.e. any two nodes are connected by a relatively short path). This family of network structures has additional implications in biological modeling: scale-free or near-scale-free networks can arise as the result of specific evolutionary processes (Middendorf, Ziv et al. 2005) such as genome duplication, giving computational biology a way to test evolutionary hypotheses using modern measurements of cellular networks.

Another way in which biological networks are often characterized is by searching the network for subsets of nodes exhibiting specific interconnectivity properties. The most classical such property analyzed in unweighted graphs is the presence of fully connected subgraphs, referred to as cliques. For example, in an undirected graph $G=(V, E)$, a three-clique is any set of three vertices $\{t, u, v\} \subseteq V$ in which all edges $\{t, u\}$, $\{t, v\}$, and $\{u, v\}$ are in E. Again, these can be indicative of specific biological features, e.g. cliques appearing in protein-protein interaction networks may indicate protein complexes; in computer science, they are of interest since clique finding is NP hard (Karp 1972), leading to a variety of optimizations and approximation algorithms. A related problem in computational biology is that of network motif finding (Milo, Shen-Orr et al. 2002), in which graphs are characterized based on other connectivity properties of subgraphs (e.g. by counting the number of triangles (three-cliques), squares, vees, or other subgraphs indicative of specific biological regulatory relationships). Likewise, the term modules in this context often refers to subgraphs exhibiting unusually high interconnectivity and/or low connectivity to the rest of the network, which can be indicative of a cohesive biological pathway or function (Ravasz, Somera et

al. 2002); these are thus often called functional modules, and particularly in weighted graphs, their discovery is closely linked to the problem of clustering.

Clustering, or the problem of finding closely related groups of entities, has been at the core of computational biology since its inception. Mathematically, clustering is an unsupervised classification operation on the vertices of weighted graphs that organizes them i) into strictly partitioned sets, referred to as hard clustering, ii) by assigning them in a weighted or probabilistic manner to zero or more sets, referred to as soft clustering, or iii) into hierarchical subsets, referred to as hierarchical clustering (MacKay 2003). For largely historical reasons, analysis of biological data has been dominated by two forms of clustering: k-means (Tavazoie, Hughes et al. 1999) and agglomerative hierarchical clustering (Eisen, Spellman et al. 1998). The former is a hard clustering algorithm in which every vertex in a graph is assigned to exactly one of k clusters, where k is an integral input parameter. Briefly, a typical implementation resembles:

1. Input integer $k>0$ and weighted graph $G=(V, E)$.
2. Randomly select $H \subseteq G$, $|H|=k$, and seed $C=\{\{h\} \mid h \in H\}$.
3. Loop until convergence:
4. For each $g \in G$, insert g into the set $\underset{c \in C}{\arg\max} \frac{1}{|c|} \sum_{h \in c} w(g,h)$.

This calculation will, of course, rely on the semantics of the edge weights w and can be modified to use the minimum, maximum, or centroid in place of the average. Hierarchical clustering, as the name implies, instead embeds each vertex in a hierarchy of subsets in which the innermost subsets indicate the most (or, depending on semantics, least) heavily weighted connections:

1. Input weighted graph $G=(V, E)$.
2. Loop until $|V|=1$:
3. Let $(v, u)=\text{argmax}_{u,v \in V}\, w(v, u)$.
4. Remove u and v from V; insert a new node $\{u, v\}$.
5. Remove all edges incident to u or v from E.
6. For all $t \neq u,v$ insert new edges $(t, \{u, v\})$ with weight $\frac{1}{|u \| v|} \sum_{x \in u \cup v} w(t, x)$.

Again, maxima and minima are dependent on semantics, and averaging can be replaced with minimization, maximization, or a centroid calculation. Clustering is particularly useful in microarray analysis, in which the specific distance or similarity measure used to calculate weights also plays a critical role; this is discussed below in more detail.

Biological Background and Terminology

Molecular biology is the study of the interactions and functional roles of biomolecules within the cell (Alberts, Johnson et al. 2007). Cells, the basic building blocks of life, are microscopic structures that represent the most basic self-reproducing machines. Using remarkably few fundamental building blocks, cells combine a small number of organic molecules into biological macromolecules; these go on to perform the basic needs of the cell, forming structural components, performing mechanical work, storing and transporting information, and carrying out regulatory programs. A computer is an example of an electronic machine that accepts input from its environment (e.g. from the keyboard), processes it using a set of predetermined relationships between transistors (e.g. programs), and uses this information to maintain a consistent internal state (e.g. a functional operating system) and to influence its surroundings

(e.g. by displaying appropriate output on a monitor). Likewise, a cell is a molecular machine that detects input from its environment, processes it using a set of predetermined relationships between molecules, and uses this information to maintain a consistent internal state (referred to as homeostasis) and to influence its surroundings - all by relying on biomolecular interactions in place of transistors and programming.

There are three main types of macromolecules that carry out these cellular tasks, the most famous being DNA: deoxyribonucleic acid. As popularized by James Watson and Francis Crick (Watson and Crick 1953), DNA consists of a double helical structure (i.e. a "twisted ladder") in which the "rungs" are made up of base pairs. An individual base pair is made up of two nucleotides, each one a small organic molecule consisting of a base, a sugar (deoxyribose in DNA), and a phosphate linking one adjacent nucleotide to the next. While each side or strand of DNA is made up of nucleotides linked covalently by phosphate groups, the bridging rungs of the base pairs are formed by noncovalent, chemical attractions between complementary bases. Four specific nucleotides are found in natural DNA, differentiated by the composition of their bases: the purines adenine (A) and guanine (G) containing large, double-ringed bases, and the pyrimidines cytosine (C) and thymine (T) with smaller, single-ringed bases. While any base pairs can appear adjacent to each other within one strand of DNA, only complementary base pairs appear across from each other to form the base pair rungs of a DNA double helix. This complementarity between specific large and small bases results in DNA "rungs" being made up of adenine-thymine (A-T) and guanine-cytosine (C-T) base pairs. As a double helix, DNA molecules are extremely large (containing potentially billions of base pairs), stable, and intrinsically error correcting (since each base is mirrored by its complement in the opposite strand). These characteristics make DNA ideal for information storage and transmission within the cell; like the

source code of a computer program, DNA by itself performs essentially no cellular work, but when read and implemented by other machinery, it contains all of the information necessary to replicate a cell and carry out its vital tasks.

The second major cellular biomolecule is RNA, ribonucleic acid, which is similar to DNA in both name and structure. RNA is normally single stranded, such that an RNA molecule is roughly equivalent to a single half or side of a DNA molecule; however, RNA nucleotides (which are thus unpaired) employ ribose as their sugar and incorporate the base uracil (U) in place of thymine (the two bases are structurally very similar). Also unlike DNA, RNA is typically an unstable molecule, easily degraded both passively by entropic forces and actively by cellular machinery; this makes it ideal for carrying short-lived messages, which is one of its main cellular tasks. Since RNA can represent exactly the same information as DNA (consisting of an equivalent four base code), its main purpose in the cell is to copy short fragments of information from DNA and transport it for use to various locations in the cell, leaving the important DNA information storage mechanism secure and unmodified. RNA made to perform this task is referred to as messenger RNA or mRNA. However, since RNA molecules are single stranded and flexible, they can form various three-dimensional structures to carry out other mechanical tasks, leading to a minor zoo of RNA subtypes: transfer or tRNA, which transport individual cellular building blocks; ribosomal or rRNA, which is a structural component of the ribosome (discussed below); small interfering siRNA and micro miRNA, which bind to and interfere with mRNA; and many types of ribozymes, which have arbitrary structures specialized to carry out specific interactions within the cell.

The third and most prevalent biomolecule is the protein; these form the main structural building blocks and functional machinery of all cells and represent the vast majority of biomass on the

planet. If DNA represents the source code of the cell, proteins are its executables. Like RNA, they are single stranded chains of repeating units, but in proteins, these units are amino acids: one of 20 molecules consisting of an amine, a carboxyl group, and a variable molecular side chain that determines the identity and chemical activity of the amino acid. Also like DNA and RNA bases, each amino acid is assigned a single letter code: tyrosine is represented as Y, lysine as K, and so on for all 20 residues. Chains of amino acid (referred to as peptides) fold into extremely complex three-dimensional structures based on noncovalent chemical forces, and the structure and chemical composition of the resultant protein determines how it interacts with other molecules in the cell and what tasks it can carry out. Structural proteins make up the membranes, supports, cables, and glue from which the cell is built; enzymes catalyze specific chemical reactions to modify other proteins or small molecules; and signaling peptides communicate and receive information intra- and intercellularly. Proteins can work together physically in groups referred to as complexes, building larger machines out of smaller ones, or they can work together conceptually in groups called pathways; proteins in the same pathway generally transmit information to each other by various signaling mechanisms, modify related molecules or metabolic products, or represent related functions necessary to carry out the pathway's overarching cellular role. Proteins themselves are also modified by the covalent attachment and removal of various small molecules, and the modification state of a protein can increase or decrease its functional activity, providing an important regulatory mechanism for the cell.

The relationship among these three biomolecules is described by a process known as the central dogma of molecular biology: DNA stores information that is transcribed (i.e. copied) for use into messenger RNA. This RNA carries the information to protein complexes called ribosomes that translate the sequence of bases into a corresponding sequence of amino acids based on the

genetic code. The resulting peptide folds into an active protein that goes on to carry out tasks necessary for the survival and reproduction of the cell. The genetic code itself is shared almost identically among all life, and maps three base pair codons to individual amino acids: three adenines in a row (AAA) are translated into a single amino acid lysine (K), two adenines and a cytosine (AAC) dictate the incorporation of asparagine (N), and so forth. Thus, an organism's genome can be thought of as a single large sequence of DNA normally totaling between a million and several billion bases; genes are specific subsequences of the DNA (a particular physical location within a genome is referred to as a locus) that are transcribed to RNA and, if protein coding (some genes encode tRNA, miRNA, and other non-mRNA features), translated into proteins. Interestingly, neither an organism's total amount of DNA nor the number of encoded genes have been found to correlate with organismal complexity, and many higher organisms have huge amounts of repetitive, non-protein-coding DNA whose function is still poorly understood.

Critically for the cell - and for computational biology - all three steps of the central dogma provide opportunities for regulation, since decreasing the transcription, translation, or posttranslational efficiency of a protein can effectively turn it "off" and stop its biological activity (or vice versa for increased activity). At the transcriptional level, specific cellular machinery (a protein complex generically referred to as polymerase) moves along a DNA strand encoding a gene, constructing a new RNA strand that replicates the DNA base sequence. The initiation, frequency, and speed of this process can be increased by proteins known as transcriptional activators or decreased by repressors, referred to collectively as transcription factors or TFs; these proteins typically interact with DNA near a gene's locus to regulate its transcription at sites called transcription factor binding sites (TFBSs). After transcription, the quantity of mRNA for a

particular gene (referred to as a transcript) and the efficiency with which it is translated can also be modulated by mRNA interactors, typically miRNAs or RNA binding proteins; these are currently an area of very active research. Post-translationally, proteins' activities are modulated by a wide variety of modifications: they can interact with other proteins to form complexes (thereby being activated or inactivated), be covalently modified by small molecule side chains, be physically moved to specific locations in the cell (preventing or allowing them to interact with other molecules), degraded, or secreted from the cell. The combination of these regulatory activities is what dictates the systems level operation of the cell - that is, the engineering that allows it to survive, grow, replicate, and interact with its environment. Individual proteins represent the building blocks of the cell, just as individual transistors are the building blocks of a processor; counting transistors will tell you less about a computer than will a schematic overview, though, and likewise, one of the goals of computational biology is to build regulatory schematics of the molecular systems of the cell.

This task is complicated by an array of factors with which molecular biologists have wrestled for decades. Real life is, unsurprisingly, more complex than the summary captured by the central dogma. The core of the problem is perhaps captured by the "one gene, one enzyme" hypothesis advanced by George Beadle and Edward Tatum (Beadle and Tatum 1941). This is the idea that each gene, located at a specific position within an organism's genome, encodes one transcript, which is translated to one protein, which performs one specific cellular function. This idea was close enough to the truth to elicit a Nobel Prize and to drive molecular biology for nearly 50 years, but as is so often the case with natural phenomena, we continue to discover that reality is far more convoluted. Genes encoded in DNA are typically made up of two subparts, introns and exons: introns are sequences of DNA that are transcribed but removed from the mRNA (spliced

out) before translation, and exons go on to be translated into proteins. Not only have we found that transcription can start and stop at slightly different loci around a gene, but transcripts are often alternatively spliced, resulting in multiple splice forms (and thus multiple translated proteins) per gene. Likewise, specific proteins can perform multiple cellular roles, either by participating in multiple distinct pathways, by responding differently to different stimuli, or by demonstrating different activity depending on subcellular localization or modification state.

An additional layer of complexity is introduced by the temporal and physical necessities of the cell cycle, which is the repeating set of tightly regulated processes by which a cell replicates itself. Reproduction is one of the defining characteristics of life, and this is equally true at a cellular level: the cell cycle, and particularly its interrelationship with the information stored in DNA, is central to many aspects of molecular biology. To reproduce, in addition to building the physical structure of a new cell out of protein components, a cell must duplicate its DNA precisely, exactly once, and ensure that each of the two new cells (referred to as mother and daughter) receive one identical copy apiece. This process is organized in temporal phases, including the controlled replication of the DNA itself. In order to package and organize billions of base pairs of DNA, it is typically wrapped around special proteins called histones in a structure resembling beads on a string; this structure in turn condenses into large bundles of DNA called chromosomes, and these are segregated to the mother and daughter cells during division. However, critically for computational biology, histones also play a (currently only partially understood) regulatory role in the normal transcription process. A short segment of DNA looped around a single histone is referred to as a nucleosome; these nucleosomes can be further organized into a regular structure called chromatin, which bundles the DNA and histones together into a compact form. Not only can chromatin on a large scale silence the transcription of genes within the structure, the

placement of individual nucleosomes can influence transcription by blocking access of TFs or polymerase to a gene's DNA. Determining the placement of nucleosomes (which may or may not be static over time) and its influence on gene regulation is another area of current research (Segal, Fondufe-Mittendorf et al. 2006).

Finally, while all life shares the basic features of the central dogma, its specific organization and complexity can vary greatly over the various taxonomic kingdoms. This manuscript will deal almost exclusively with eukaryotes, the class of organisms to which essentially all multicellular life belongs; eukaryotic cells are characterized by the presence of intracellular organelles defined by membranes, most notably the nucleus that contains the cell's DNA. Prokaryotes are single celled organisms without a nucleus and, generally, with a single circular chromosome. While eukaryotes can be unicellular (e.g. the yeast *Saccharomyces cerevisiae*, a single celled fungus responsible for fermentation in brewing and baking), the kingdom also comprises multicellular plants and animals (metazoans), including human beings. Even though genome size does not strictly correlate with organismal complexity, higher organisms do introduce increasingly many layers of regulation and interactions and are also much more difficult to study holistically, part of the reason it is still difficult to understand human biology and disease at the molecular level.

Computational biology is thus the application of computer science to the understanding of this dazzling array of biological complexity. Throughout most of the centuries-long history of biology, we have only been able to gather information about the workings of life through observation at the macroscopic level; microscopes with cellular resolution have been available since the Renaissance, investigations into molecular composition and interactions began more or less in the 18th century, and our understanding of the central dogma and its consequences has only crystallized since the early 1900s (Morange 1998). Fortunately, and not coincidentally, the

advent of whole-genome sequencing and the consequent explosion of biological measurements at the molecular level has coincided with the widespread availability of computational analysis and algorithms for data processing. Having provided an extremely high-level overview of the molecular forces at work in the cell, it remains to discuss how this flurry of infinitesimally small activity is observed and assayed in modern biology.

High- and Low-Throughput Assays in Modern Biology

While the contributions of computation to biology and of biology to computer science continue to grow more and more diverse, the inspiration for bioinformatics lies in the analysis of biological data. Due to the miniscule scale on which molecular biology proceeds, essentially all biological assays are more or less indirect; similarly, due to molecular biology's tremendous complexity, they are all more or less imprecise. These factors, fortunately, are exactly what machine learning has been created to deal with, with its foundations in robustness to noise, missing data, and hidden variables. This is perhaps epitomized by the problem of genome sequencing: given a population of cells, determining the sequence of As, Cs, Gs, and Ts making up their chromosomes and overall DNA was, as of only a few decades ago, a daunting experimental challenge. The earliest DNA sequencing was carried out (and modern sequencing is still inspired by) a method referred to as Sanger sequencing (Sanger and Coulson 1975). Briefly, once DNA is purified from a sample, it can be duplicated *in vitro* by adding DNA polymerase in combination with deoxynucleotide monomers, individual A, C, G, and T triphosphates. By repeatedly performing individual reactions in which one of these monomers is replaced with a non-extensible (terminating) dideoxy- form, a population of DNA fragments can be built up, consisting of every possible substring of the isolated DNA beginning from one end and

extending, in the longest fragments, to the other. These fragments can be radioactively or fluorescently labeled and appropriately detected, resulting in A-, C-, G-, or T-specific signals at increasing lengths (and thereby positions within the genome).

In practice, nothing is as simple as it sounds in theory. Due to differences in incorporation of the four bases and the limited dynamic range of differentiating DNA fragment lengths, individual sequencing reads using these methods were originally limited to several dozen bases and are currently limited to roughly 1,000. In order to sequence whole genomes (consisting of billions of bases), this led to a strategy called shotgun sequencing in which a genome is randomly fragmented and these fragments are individually amplified and sequenced. This leads to literally billions of tiny, overlapping DNA sequences that must be stored, organized, and reassembled computationally (Ewing, Hillier et al. 1998). Modern high throughput sequencing techniques (also referred to as deep sequencing) have largely replaced termination and length-based assays; these employ reversible termination, fluorescence quantification, or various ligation/hybridization strategies and are currently developing extremely rapidly (Strausberg, Levy et al. 2008). However, as a whole, these strategies uniformly exchange higher throughput for shorter reads, increasing the opportunity for algorithmic improvement in genome assembly.

Once genomes are assembled, a host of computational data collection and analysis possibilities emerge. The detection of genes within a genome and of substructure within genes has been performed using a host of techniques, primarily based on Hidden Markov Models (Durbin, Eddy et al. 1998). Genomes and individual genes can be compared between organisms over evolutionary timeframes, leading to models of evolution, mutation, selection, and speciation. Since the genetic code is constant, DNA sequences can be computationally translated to amino acid sequences, which can in turn be compared, examined for substructure (e.g. reused or

conserved protein domains, small recurring functional units), or algorithmically folded to estimate three-dimensional protein structure using physical models (Kryshtafovych, Fidelis et al. 2007). Transcription factor binding sites (and other regulatory binding motifs) can also be predicted directly from DNA sequences, although these are more often assayed experimentally as described below.

Although genome sequences were one of the first types of high-throughput genomic data, classical low-throughput laboratory experiments were the other. Molecular biology has a rich history of hundreds of experimental types spanning hundreds of thousands of experiments, collected in literature and community knowledge over the course of decades. When computational biology started to become a reality with the advent of genome sequencing, it became practical to collect these individual data points into large repositories for analysis as well. These data repositories (which have only become more important as they accumulate modern high-throughput data alongside classical low-throughput results) generally catalog a few specific types of results: the participants in specific pathways, direct protein-protein interactions, transcriptional regulators, and so forth. They also tend to divide into curated databases focusing on high-level functional information summarized from literature and experimental databases focusing on lower-level, raw experimental results. Some primary examples of the former are the Gene Ontology (Ashburner, Ball et al. 2000) (GO, cataloging protein roles, biochemical functions, and subcellular localization in a semihierarchical ontology), the Kyoto Encyclopedia of Genes and Genomes (Kanehisa, Araki et al. 2008) (KEGG, specific pathways, mainly metabolic), the Munich Information center for Protein Sequences (Ruepp, Zollner et al. 2004) (MIPS, strictly hierarchical protein roles), Reactome (Vastrik, D'Eustachio et al. 2007) (individual curated enzymatic reactions), and Online Mendelian Inheritance in Man (Hamosh, Scott et al. 2005) (OMIM, genes

linked to known disorders). Common examples of the latter include the Human Protein Reference Database (Mishra, Suresh et al. 2006) (HPRD, curated protein-protein interactions and modifications), the Biomolecular Interaction Network Database (Bader, Donaldson et al. 2001) (BIND, protein-protein and synthetic interactions), the General Repository for Interaction Datasets (Stark, Breitkreutz et al. 2006) (bioGRID, protein-protein and synthetic interactions), the Database of Interacting Proteins (Salwinski, Miller et al. 2004) (DIP, protein-protein interactions), the InterAction database (Kerrien, Alam-Faruque et al. 2007) (IntAct, various experimental interactions), and the Molecular Interaction database (Chatr-aryamontri, Ceol et al. 2007) (MINT, protein-protein interactions).

Although traditional entries in these databases have been made using a tremendous assortment of experimental techniques, a number of high-throughput assays have been recently developed specifically for detecting protein-protein physical interactions (Fields and Song 1989; Walhout and Vidal 2001; Vasilescu and Figeys 2006). Genome-scale varieties of these assays tend to fall roughly into two categories: affinity/mass spectrometry and yeast two-hybrid (Y2H). Both rely on the framework of screening a collection of proteins (the prey, usually extracted from some cellular population of interest) with some known protein or compound (the bait). Many baits are used in affinity-based methods, ranging from antibodies developed to attract specific proteins (and their interaction partners) to small molecule interactors to metal ions. Once the bait has been used to separate interacting (and only interacting) prey from the sample, the proteins in the retained prey are characterized, often by mass spectrometry (Gavin, Bosche et al. 2002; Ho, Gruhler et al. 2002; Krogan, Cagney et al. 2006). Y2H screens rely on genetically manipulated yeast strains in which several modified proteins have been inserted. The first is a reporter gene placed downstream of a promoter sequence that is normally inactive. The second is a

transcription factor that will activate this promoter, but it is broken into two parts: a binding domain (BD) that will physically bind this promoter sequence, and an activation domain (AD) that will initiate transcription. The BD is fused to one protein (e.g. the bait), the AD to another (e.g. all prey), and only if the two fusion proteins physically interact will the transcription factor be active and the reporter transcribed. Since Y2H relies on several assumptions and technical abnormalities (e.g. strong overexpression of the target proteins, nuclear localization to initiate transcription, etc.), its accuracy has still not been reliably quantified, but it has been used with great success for several genome-wide screens for protein-protein interactions in yeast (Uetz, Giot et al. 2000; Ito, Chiba et al. 2001).

A second type of relationship between two proteins that is less physically based, but equally biologically important, is referred to as a synthetic (or genetic) interaction. A protein-protein interaction or PPI refers specifically to two proteins coming into physical contact in the cell, either transiently or semi-permanently in a complex. A synthetic interaction instead refers to two proteins that, when both disrupted, produce a phenotype that is unusual relative to the effects of their two individual disruptions (Novick, Osmond et al. 1989; Tong and Boone 2006). An illustrative example is synthetic lethality: two proteins that, when individually removed from an organism's genome (often referred to as deletion or knockout), cause no extreme phenotype, but when both removed, render the organism inviable. Synthetic interactions can also be beneficial (e.g. synthetic rescue, in which deletion of individual genes is lethal and deletion of both genes is not) or quantitative, in which case beneficial and detrimental interactions are referred to as alleviating or aggravating (also synthetic sickness). While the physical basis of synthetic interactions is not fully characterized, they are thought to reflect mainly on two types of protein relationships: operation within the same pathway and membership in parallel redundant

pathways. The former is expected to result in alleviating interactions; once the pathway is disrupted by the removal of one protein, the removal of the second will have no effect. The latter is expected to incur aggravating interactions, as the cell can fall back to the redundant pathway when one is disrupted, but cannot recover from the loss of both pathways. Interestingly, synthetic interactions between cocomplexed proteins can be either alleviating (if deletion of one protein immediately disrupts the complex) or aggravating (if the complex remains stable after losing one protein but destabilizes with the removal of both). While quantitative models of the expected effects of synthetic lethality are still being developed (Tong, Lesage et al. 2004; St Onge, Mani et al. 2007), whole-genome synthetic screens have also been quite successful in yeast (Tong, Evangelista et al. 2001; Giaever, Chu et al. 2002).

A final experimental method for genomic data collection that we will discuss is high-content microscopy. This family of techniques is qualitatively fairly different from protein interaction assays, relying on either microscopic observation of cellular phenotypes or subcellular structures to quantify the effects of protein disruption, chemical exposure, or other environmental perturbations (Carpenter and Sabatini 2004; Edwards, Oprea et al. 2004). A simple example of such a technique is flow cytometry, in which some cellular component (e.g. an individual protein or the cell's DNA) is labeled with a fluorescent marker; single cells flow over a scanner, which records the amount of fluorescence in each cell. This is most often used to determine the distribution of cell cycle phases within a population of cells, since labeling DNA will cause twice as much fluorescence in dividing cells, but it can be used to quantify the per-cell frequencies of any subcellular marker. Flow cytometry and other cell sorting techniques (e.g. robotic deposition of specific strains on high density plates) can be coupled with automated microscopy to capture individual cellular images and thus phenotypes beyond overall fluorescence levels. These can be

at the whole cell level (e.g. cell size, shape, or division characteristics (Ni and Snyder 2001; Giaever, Chu et al. 2002; Jorgensen, Nishikawa et al. 2002)) or at a subcellular level (e.g. various fluorescent tagging techniques (Huh, Falvo et al. 2003; Sprague, Pego et al. 2004; Sapsford, Berti et al. 2006; Sieben, Debes Marun et al. 2007)) and combined with automated image processing to extract detailed statistics on the cell's response to experimental perturbations.

Microarrays

At present, the single experimental assay that has provided the richest, most abundant genomic data and the greatest impetus to computational biology is the microarray. Although microarrays have since evolved to quantify an assortment of biological phenomena, their original purpose (which also remains their most common current use) was to measure the abundance of many individual mRNAs within a mixed population (Kulesh, Clive et al. 1987; Schena, Shalon et al. 1995). Microarrays for this purpose, sometimes more specifically referred to as gene expression microarrays, rely on the fact that two isolated DNA strands will anneal to form a double-stranded helix only if their respective bases are complementary. This can be taken advantage of in a manner similar to the bait/prey techniques used to detect protein-protein interactions: many copies of a particular single-stranded DNA sequence can be constructed as bait, and if allowed to interact with a pool of unknown single-stranded DNA, only sequences that are complementary will "stick" as prey. Measuring the amount of prey attracted to a particular bait reveals how much DNA with the bait's sequence (or, more properly, a complementary sequence) was present in the original population. Since, as discussed above, it is relatively easy to synthesize artificial DNA with a known sequence, this can easily be repeated with many different baits to fully characterize all sequences of interest within a population.

In any cellular culture, genes are transcriptionally up- or down-regulated under specific conditions in order to produce more (or fewer) proteins appropriate to those conditions: when provided with a particular nutrient, cells will produce proteins to metabolize it; when stressed by a particular chemical, cells will produce proteins to render it harmless or to export it; and so forth. When genes are thus regulated at the transcriptional level, the amount of mRNA corresponding to some gene (and thus to one or several of its protein products) is an excellent indicator of its activity (and, often, functional relevance) under some condition. However, mRNA (which is inherently single-stranded) does not form stable helices in the same manner as DNA and is thus not directly amenable to the bait/prey technique. Thus, in order to reliably assay mRNA from some cellular sample, an enzyme (reverse transcriptase) is used to construct a complementary DNA strand (referred to as cDNA) from the RNA template, just as RNA is transcribed in the cell from a complementary DNA template. This process is quantitative, so by measuring the amount of each cDNA prey in such a pool using many known baits, we can observe the abundance of each mRNA transcript in our original cellular sample (Lashkari, DeRisi et al. 1997).

An assortment of physical platforms has been developed to perform this quantification; commercialization of microarray technology has led to a host of competing (and often complementary) techniques. The earliest microarrays used robotic pinning of known DNA libraries onto glass microscope slides (Duggan, Bittner et al. 1999). These DNA fragments are typically selected to anneal specifically to a single gene's transcript (or, in higher organisms, to a specific mRNA splice form). Referred to as spotted cDNA arrays, these can achieve a density of thousands to tens of thousands of probes (bait sequences) per slide, and are thus capable for many organisms of quantifying the transcripts of every gene in the genome. Prey abundances are

measured by incorporating a fluorescent dye into the sample cDNA during the reverse transcription process; since the position and sequence of every bait probe are known, a laser scanner then determines the level of fluorescence at each spot and thus the amount of that sequence's corresponding mRNA transcript. Since hybridization efficiency can vary depending on experimental conditions and the specific probe sequence used to target a gene, an internal control is generally employed. This usually consists of mRNA sampled from some neutral, control condition and labeled with a second dye fluorescing at a different frequency (cyanine 3 and cyanine 5, or Cy3 and Cy5, are by far the most common such dyes (Stears, Martinsky et al. 2003)). By simultaneously hybridizing both samples to the same microarray and comparing the intensities of the two dyes, one can precisely determine whether and by how much each gene has been up- or down-regulated in the experimental condition relative to the control condition; such microarrays are correspondingly referred to as dual-channel or two-color.

Other technologies exist for constructing probes with known sequences and locations on a slide and for quantifying single-channel fluorescence intensities. Affymetrix (Lockhart, Dong et al. 1996) provides a single-channel microarray technology in which 25-base oligonucleotide (synthetic single-stranded DNA) sequences are built directly on the chip substrate using mask-based photolithography. This can achieve much greater probe density (currently ~10^6 probes per array), and a combination of careful probe design and the use of many probes per gene allows quantification of absolute transcript levels in a single-channel hybridization. Similarly, Nimblegen (Singh-Gasson, Green et al. 1999) uses individually controlled micromirrors to photolithographically build approximately two million 25- to 85-base oligomer probes on each array; these are also generally hybridized using a single channel. Since building a new photolithographic mask is a time consuming process, and micromirrors can be reconfigured

instantaneously, Nimblegen arrays are much more amenable to custom probe design than are Affymetrix arrays. Agilent (Lee, Rinaldi et al. 2002) uses techniques derived from inkjet printing to deposit 60-base oligomers onto arrays at (currently) ~10^5 probes per array; these are typically hybridized as dual-channel arrays. Finally, Illumina (Oliphant, Barker et al. 2002) constructs 50-base oligomers on the surfaces of extremely small silica beads; these are then affixed to a chip substrate randomly at a density between 10^6 and 10^7 probes per array and hybridized using either one or two channels. This is interesting in that the probe locations are not known a priori and are decoded by incorporating sequence-specific, detectable tags into the synthesized oligomers. Given this remarkable diversity of manufacturing processes, each platform unsurprisingly offers its own strengths, weaknesses, and bioinformatic analysis opportunities.

The reality of any microarray is messy: concentrations of reagents can vary, hybridization can be more or less efficient, fluorescent dyes can decompose, bubbles can block access to specific probes, cross-hybridization can occur between slightly mismatched sequences, or a host of other systematic or unsystematic biases can be introduced. Thus, while a fluorescent scanner will blithely report the intensity of each probe, it is up to the computational biologist to correctly interpret the millions of resulting numbers as specific transcript abundances (Quackenbush 2002). A scanner will typically report many individual pixel intensities per probe; these must be reconciled and, if in disagreement, discarded as inconsistent data. Similarly, the intensity of the background fluorescence level reported by the scanner and any position-specific biases (e.g. bubbles of non-hybridization) will vary spatially, and affected probes must be removed or corrected. In dual-channel arrays, the intensity distributions of the two dyes are typically quite different and must be reconciled before comparison. When multiple probes per gene (or other biological target of interest) are present, they must also be resolved; and when multiple

microarrays are used as part of a single dataset, inter-array variation must be taken into account. Finally, the data is typically logarithmically transformed (either as raw single-channel intensities or as ratios of dual-channel intensities, referred to as log ratios) and one resulting value assigned to each gene; if desired, values missing as a result of this processing can often be accurately re-imputed based on the remainder of the data (Troyanskaya, Cantor et al. 2001).

Since their inception, the concepts underlying microarrays have developed to quantify many biological targets other than mRNA abundance. Many assays rely on tiling arrays; rather than using one (or a few) probes specific to particular genes, these microarrays use as many probes as possible spaced linearly along an organism's genome (or portions thereof). Tiling arrays thus tend towards higher density and shorter probe lengths; for example, current Affymetrix arrays tiling the *S. cerevisiae* genome place 25-mer probes every five bases along the genome, resulting in a 20-base overlap between probes (Gresham, Ruderfer et al. 2006). One use for such an array is comparative genomic hybridization (CGH), in which the genomic DNA of a sample is fragmented and hybridized directly to an array in place of reverse transcribed cDNA (Schwaenen, Nessling et al. 2004). When compared with a reference sample containing exactly one copy of every genomic locus, this can be used to detect loci or entire chromosomes that have been duplicated; amplification or deletion of loci and chromosomes in this manner is a hallmark of several genetic disorders, particularly cancer (Alberts, Johnson et al. 2007).

Another application of tiling arrays is chromatin immunoprecipitation on chip, or ChIP-chip, in which a sample's genomic DNA is chemically cross-linked (i.e. covalently bound) to any proteins interacting with it in some environment: transcription factors specific to that environment, DNA replicating proteins during the cell cycle, and other regulators and structural components (Ren, Robert et al. 2000). The DNA is then fragmented and all unbound fragments discarded. A protein

prey of interest (e.g. a specific transcription factor) is then captured using an antibody bait, other proteins (and their bound sequence) are allowed to escape, and the DNA sequences bound specifically to the prey are released. These can then be quantified on a microarray to determine all of the target protein's genomic binding sites. This technique can be extended to discover the locations of all nucleosomes in an organism's genome (Segal, Fondufe-Mittendorf et al. 2006) or to determine where an organism's DNA has been epigenetically modified by methylation (Schumacher, Kapranov et al. 2006). Finally, since the probes on such tiling arrays are typically short, single base mismatches will cause a quantifiable decrease in binding efficiency; this can be taken advantage of in order to find single nucleotide polymorphisms (SNPs) in a manner akin to very high resolution CGH (Gresham, Ruderfer et al. 2006).

Location-based arrays have also seen increasing use for substances other than DNA probes. Protein arrays, for example, spot specific bait proteins in place of DNA oligomers and quantify the binding of protein prey (MacBeath and Schreiber 2000); this provides a more direct assay of protein abundance than does mRNA level detection. These arrays generally employ antibodies both as bait and in the detection of bound protein, the former being developed to bind specific known proteins and arrayed accordingly, the latter in order to quantitatively associate various fluorescent dyes with the captured prey. Whole cells can also be printed in a similar manner (Wheeler, Carpenter et al. 2005), referred to as cell or transfection arrays, with specific genes up- or down-regulated at each spot; these can then be quantified using methods similar to those described above for cytometry, e.g. fluorescent markers or whole cell imaging. Even gene expression microarrays are themselves slowly being supplemented by high throughput (also called deep) sequencing, in which cDNAs are individually sequenced, identified, and counted

rather than being indirectly quantified by (possibly relative) fluorescence levels (Strausberg, Levy et al. 2008).

Regardless of the experimental platform, most of these assays share the same goal and, in concept, produce the same type of information: numbers representing the activity levels of each gene in an organism's genome under a specific experimental condition. This reduces the genome (under one condition) to a vector of numbers; when multiple assays are performed under several related conditions, this produces a matrix of continuous values, typically with rows corresponding to genes and columns to experimental conditions. This vector- and matrix-oriented view of microarray data is itself amenable to a tremendous array of bioinformatic techniques, one of the most canonical of which is referred to as differential expression analysis (Golub, Slonim et al. 1999; Alizadeh, Eisen et al. 2000). This is a supervised problem in which a set of microarrays is partitioned into one or more subsets with known class labels. For example, a group of tumor samples might be broken into specific cancer subtypes. The problem lies in the discovery of one or more genes with characteristic, different expression patterns within each class. Such a set of marker genes can then be used to predict the class of new, unlabeled samples. A naive approach in the binary case is to t-test each gene's expression vectors between the two classes; all significantly differing genes can then be assembled into a composite diagnostic signature. Many techniques that are more sophisticated have also been developed for discovering both clinical and scientific biomarkers (Cui and Churchill 2003).

Differential expression analysis is a supervised technique in that it requires labeled examples (i.e. preclassified microarrays) in order to determine diagnostic gene sets; one unsupervised technique for discovering common patterns in gene expression is singular value decomposition (SVD, a specific instance of principle component analysis, PCA (Wall, Rechtsteiner et al. 2003)).

SVD is a general mathematical technique that transforms a collection of high-dimensional data points into equivalent points in a new space; the bases of this new space (called singular vectors or, slightly improperly, eigenvectors) are the orthogonal axes of greatest variance in the data. This is demonstrated in two-dimensional space in Figure 3; the data does not change, but its representation does, as each point is now constructed out of a weighted combination of singular vectors. Algebraically, SVD decomposes an mxn matrix X (in which, for microarray data, each row represents a gene and each column a condition) into three matrices: an mxn U, left singular vectors or eigenconditions, encoding the normalized weights of each new basis vector for each gene; a diagonal nxn matrix Σ, singular values or eigenvalues, representing the global weights of the new basis vectors in the data (i.e. the commonality or strength of that expression pattern); and a transposed nxn matrix V^t, right singular vectors or eigengenes, which are the basis vectors themselves (i.e. the expression patterns capturing the most common variability in the data):

$$X_{mxn} = U_{mxn}\Sigma_{nxn}V^t_{nxn}$$

Since the singular vectors V represent the axes of greatest variation in the data, they often (but certainly not always) capture some degree of significant biological activity (Alter, Brown et al. 2000). It must be stressed that, as an unsupervised method, these axes of greatest variation (eigengenes) may or may not represent meaningful biological variability; each eigengene might represent a single specific biological response, an overlapping combination of many such responses, or a completely unrelated concordance of noise in the data. Reprojecting gene expression vectors onto such a basis when it is not biologically motivated is no better - and often worse - than analyzing the original microarray directly; conversely, when a few specific signals are disproportionately strong in the data, this reprojection can accurately recover underlying biological information that would otherwise have been difficult to detect (Hibbs, Hess et al. 2007).

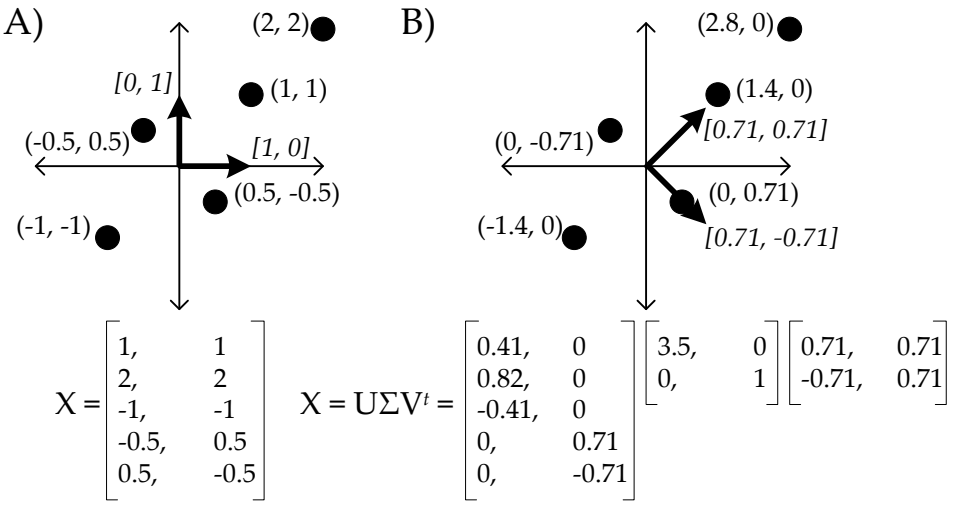


Figure 3: An application of singular value decomposition (SVD) to a two-dimensional sample dataset. A) Five data points are represented in standard Cartesian coordinates using the usual $\boldsymbol{x}$=[1, 0] and $\boldsymbol{y}$=[0, 1] bases. B) The same five data points are transformed using SVD to coordinates based on the axes of maximum variation, $\boldsymbol{x}$'=[0.71, -0.71] and $\boldsymbol{y}$'=[0.71, 0.71] as encoded in the matrix V. The singular values Σ indicate the overall magnitude of this variability, i.e. the range of the data is 3.5 times greater along $\boldsymbol{y}$' than along $\boldsymbol{x}$'. The transformed data points themselves can be recovered by multiplying UΣ as illustrated in the diagram. This decomposition can be performed equivalently for data of arbitrarily high dimension, e.g. transforming a matrix of many microarray samples into an equivalent combination of eigenconditions, singular values, and eigengenes (the latter representative of the strongest patterns of variability in the data).

Finally, by far the most common high-level analysis performed on microarray data is the unsupervised operation of clustering. As described above, clustering can be either hard (e.g. k-means), soft, or hierarchical, and all forms have been applied in endless permutations to microarray data. Both the first and most lasting microarray clustering algorithms have been hierarchical (Eisen, Spellman et al. 1998); see Figure 21, Figure 28, Figure 41, etc. for examples. Hierarchical clustering has the benefit of leaving the values of an entire input genes-by-experiments matrix unchanged and simply reordering the rows and/or columns to improve

visualization. Its two main drawbacks, however, are that it can be driven by the strongest few signals in the data (thus leaving genes sharing weaker signals scrambled among the hierarchy) (Sherlock 2000) and that it has no allowance for genes participating in multiple clusters (Tanay, Sharan et al. 2002). Any form of microarray clustering depends, of course, on the distance or similarity measure used to transform a matrix of expression values into a weighted graph of gene pair scores. Most commonly, Pearson correlation or Euclidean distance are used to compare gene expression vectors, with the assumption being that genes expressed at similar levels over many conditions are more likely to be functionally related. Substantial evidence indicates that this is generally true (Huttenhower, Flamholz et al. 2007), and while countless other clustering algorithms and similarity measure have been proposed, hierarchical clustering using simple correlation remains one of the most powerful and widespread means of global microarray analysis.

Wet Lab, Dry Lab: Applying Computational Biology to Laboratory Experiments

One of the greatest current opportunities in modern biology is the closer integration of computational analyses with laboratory efforts. The last decade has seen the addition of *in silico* experimentation to the canon of *in vivo* and *in vitro* assays, but while experimentalists have long taken advantage of the complementarity of *in vivo* and *in vitro* results, the place of *in silico* models and predictions remains less clear. Computational biology offers many advantages: it can integrate large quantities of noisy and heterogeneous data to make systems-level predictions, it can statistically discover small but consistent effects, and it can do so more quickly and cheaply than many laboratory assays. But its obvious disadvantage is that, at the end of the day, any computational result is simply a prediction that must be reconciled with biological reality.

Cementing the relationship between computational and experimental biology involves at least two important aspects: the optimization of computational protocols so they are maximally biologically relevant, and the sociological reconciliation of computational and biological techniques. Put simply, experimental biology is a field whose history spans centuries; bioinformatics, even with the great strides made in the past decade or two, has not yet completely adapted itself to the needs of biologists, and biologists have not yet fully determined how to best take advantage of computational tools. Solving this problem itself requires further experimentation to see how computational predictions can best direct laboratory work and how the results of this work can, in turn, feed back into computational methodologies.

Here, we provide a detailed analysis of one end-to-end solution to these problems: we design a computational protein function prediction system specifically geared toward informing laboratory work; this system directs quantitative experimental assays themselves designed to inform future computational work; we demonstrate that computational input vastly improved our capacity for biological discovery; and we analyze the impact of biological input on the behavior of computational systems in general. While a complete understanding of how best to integrate computational and biological experimentation will certainly take additional decades of effort from both fields, this study provides both an initial demonstration of success and a framework within which to continue future endeavors.

We would like to thank David C. Hess and Amy A. Caudy for their extensive experimental collaboration on this project, as well as Matthew A. Hibbs and Chad L. Myers for their work on the computational aspects and the assembly of related manuscripts.

Investigating *S. cerevisiae* Mitochondria: Computational and Experimental Design

In order to better understand the interplay between computational and experimental biology, we chose protein function prediction as a computational environment and *S. cerevisiae* mitochondria as an experimental system. The former is an established problem in bioinformatics that dates from the earliest days of the field (Fleischmann, Moller et al. 1999; Ashburner, Ball et al. 2000; Rost, Liu et al. 2003): given some amount of experimental data, potentially ranging from DNA sequence to protein-protein interactions to gene expression measurements, identify the cellular pathways and processes within which a protein participates. Given that i) the amount of publicly

available experimental data has increased exponentially since this early work and ii) many proteins even in simple organisms still have no characterized functions (let alone in higher organisms or considering multiple functions per protein), identifying protein functions is still an area of very active research (Pena-Castillo, Tasan et al. 2008).

We focused on three computational systems (Myers, Robson et al. 2005; Huttenhower, Hibbs et al. 2006; Hibbs, Hess et al. 2007) for protein function prediction, with the goals that i) they be particularly capable of guiding laboratory experiments with high precision (possibly at the expense of recall) and ii) they be able to incorporate the results of experiments into their algorithmic pipeline in order to iteratively improve performance. In other words, they must make a few good predictions, and after these predictions are tested in the laboratory, the algorithms must learn from the results and make a new set of better predictions. bioPIXIE (Myers, Robson et al. 2005), the first of the three systems, weights heterogeneous genomic data probabilistically in order to predict protein-protein functional interactions; given a set of known mitochondrial genes, these predicted relationships can be turned into predicted functions using various measurements of "guilt by association." Likewise, MEFIT (Huttenhower, Hibbs et al. 2006), the second system, probabilistically weights large collections of microarray data to predict interactions; for both algorithms, laboratory results can be incorporated by expanding the set of "known" mitochondrial genes. Finally, SPELL (Hibbs, Hess et al. 2007) uses a query-based approach to dynamically search through databases of gene expression values; new mitochondrial predictions can be made by querying on known mitochondrial proteins, and new laboratory results can be incorporated into these queries.

Likewise, yeast mitochondria represent a complex and dynamic area of biology with both basic scientific (Dimmer, Fritz et al. 2002; Westermann and Neupert 2003; Shutt and Shadel 2007) and

clinical (Foury and Kucej 2002; Tomaska 2002; Schwimmer, Rak et al. 2006) relevance. Unlike most eukaryotes, yeast can survive without functioning mitochondria; this condition (referred to as "petite" (Ogur and St John 1956)) is both easily detectable through reliable assays (Ogur, St. John et al. 1957) and directly indicative of a disruption of mitochondrial function. Moreover, petite colonies occur in standard laboratory yeast at a baseline rate of ~20% (Baruffini, Ferrero et al. 2007); when a protein involved in mitochondrial function is impaired in any way (e.g. by deletion from the genome), this rate can increase or decrease, and precise quantification of a yeast mutant's petite frequency thus represents an opportunity to observe both subtle and catastrophic defects. In combination with *S. cerevisiae*'s genetic tractability, this provides an ideal system for coupling laboratory experiments with computation: given proteins predicted to be involved in mitochondrial function, they can be deleted from the genome and the resulting phenotypes assayed, and the quantitative nature of these assays can be incorporated into future predictions.

We successfully implemented this system, screening 193 of the most confident computational predictions to discover 109 proteins required for proper mitochondrial function. This was achieved in less than one person-year, demonstrating the computational input can vastly decrease the amount of time required to make substantial, rigorously confirmed laboratory discoveries. These 109 proteins include functions not only in general mitochondrial activity but also specific confirmations in the areas of mitochondrial motility, aerobic respiration, and several multi-protein synthetic interactions. Moreover, these discoveries represent multiple iterations of computational prediction, in which the results of the first iteration (141 assays) were provided to the computational systems in order to improve the results of the second iteration (52 assays). Not only does this represent a tremendous biological advance, but by examining our computational performance using standard evaluation techniques, we observe that computational function

prediction can often produce even more accurate results than would be expected by most numerical (as opposed to laboratory) evaluations. This study thus represents not only a substantial advance for yeast mitochondrial biology, but also expands our understanding of computational function prediction as a field and provides a framework for any future integration of computational and laboratory techniques.

Biological Ramifications: Computation Works

In order to understand molecular biology at a systems level, it is first necessary to learn the functions of genes by identifying their participation in specific cellular pathways and processes. While protein sequence and structural analyses can provide valuable insights into the biochemical roles of proteins, it has proven much more difficult to associate proteins with the pathways where they perform these roles. Recently, high-throughput and whole-genome screens have been used to form basic hypotheses of protein participation in biological processes. However, the results of these studies are not individually reliable enough to functionally associate proteins with pathways. Many computational approaches have been developed to integrate data from such high-throughput assays and to generate more reliable predictions (Sharan, Ulitsky et al. 2007), but protein function cannot be confidently assigned without rigorous experimental validation targeted specifically to the predicted pathway or process. Surprisingly few follow-up laboratory efforts have been performed on the basis of computational predictions of protein function, and as such, these approaches remain largely unproven, and consequently underutilized by the scientific community (Murali, Wu et al. 2006; Pena-Castillo, Tasan et al. 2008). Here, we demonstrate that computational predictions can successfully drive the characterization of protein roles using traditional experiments. To test the approach, we

systematically measured the mitochondrial inheritance rates of a tractable set of *S. cerevisiae* strains carrying deletions of genes predicted to be necessary for this biological process.

The mitochondrion is an organelle central to several key cellular processes including respiration, ion homeostasis, and apoptosis. Proper biogenesis and inheritance of mitochondria is critical for eukaryotes, as 1 in 5,000 humans suffers from a mitochondrial disease (Schaefer, Taylor et al. 2004). *Saccharomyces* has proven to be an invaluable system for studying a variety of human diseases (Botstein, Chervitz et al. 1997; Smith and Snyder 2006), including cancer (Hartwell 2004), neurologic disorders (Walberg 2000), and mitochondrial diseases (Foury and Kucej 2002; Tomaska 2002; Schwimmer, Rak et al. 2006). Yeast is a particularly attractive model system for studying mitochondrial biology due to its ability to survive without respiration, permitting the characterization of mutants that impair mitochondrial function. In fact, experimental efforts in yeast have identified genes crucial for mitochondrial organization and biogenesis (Dimmer, Fritz et al. 2002), including mutants that affect the sub-processes of mitochondrial genome inheritance (Contamine and Picard 2000), protein targeting and import (Pfanner and Geissler 2001), protein synthesis (Myers, Pape et al. 1985), protein complex assembly (Model, Meisinger et al. 2001), and actin-based transmission of mitochondria to the daughter cell (Boldogh and Pon 2007). In addition to the experimental utility of yeast, it is well suited for the application of computational prediction approaches due to the availability of manually-curated annotations of yeast biology and the available wealth of genome-scale data.

Previous efforts have focused on identifying mitochondria-localized proteins through laboratory techniques such as mass spectrometry and 2D-PAGE (Sickmann, Reinders et al. 2003; Reinders, Zahedi et al. 2006) and through computational predictions of cellular localization (Calvo, Jain et al. 2006; Prokisch, Andreoli et al. 2006). These approaches have resulted in the identification of

over 1,000 mitochondria-localized proteins in *S. cerevisiae* (Westermann and Neupert 2003). However, despite yeast's convenience as a model system, mitochondrial phenotypes of ~370 of these 1,000 localized proteins have not been characterized, so the mitochondrial role of these predictions is unknown (over half of these 370 have no known function in any cellular process). Previous computational efforts have attempted to address this problem by predicting putative mitochondrial protein modules (Perocchi, Jensen et al. 2006) and examining expression neighborhoods around mitochondrial proteins (Mootha, Bunkenborg et al. 2003). While valuable, these efforts have not been reliably confirmed through comprehensive experimental follow-up, which is required in order to convert these types of predictions into concrete knowledge (Shutt and Shadel 2007).

Here, we describe a strategy that combines computational prediction methods with quantitative experimental validation in an iterative framework. Using this approach, we identify new genes with roles in the specific process of mitochondrial inheritance by directly measuring the ability of cells carrying deletions of candidate genes to propagate functioning mitochondria to daughter cells. We assayed our 193 strongest predictions with no previous experimental literature evidence of phenotypes and interactions establishing a function in mitochondrial inheritance. By these assays we experimentally discovered an additional 109 proteins required for proper mitochondrial inheritance at a level of rigor acceptable for function annotation. Further, we identified more specific roles in mitochondrial biogenesis for several predicted genes through mitochondrial motility assays and measurements of respiratory growth rates. We also discovered genes with redundant mitochondrial inheritance roles through targeted examination of double knockout phenotypes. This demonstrates that using an ensemble of computational function

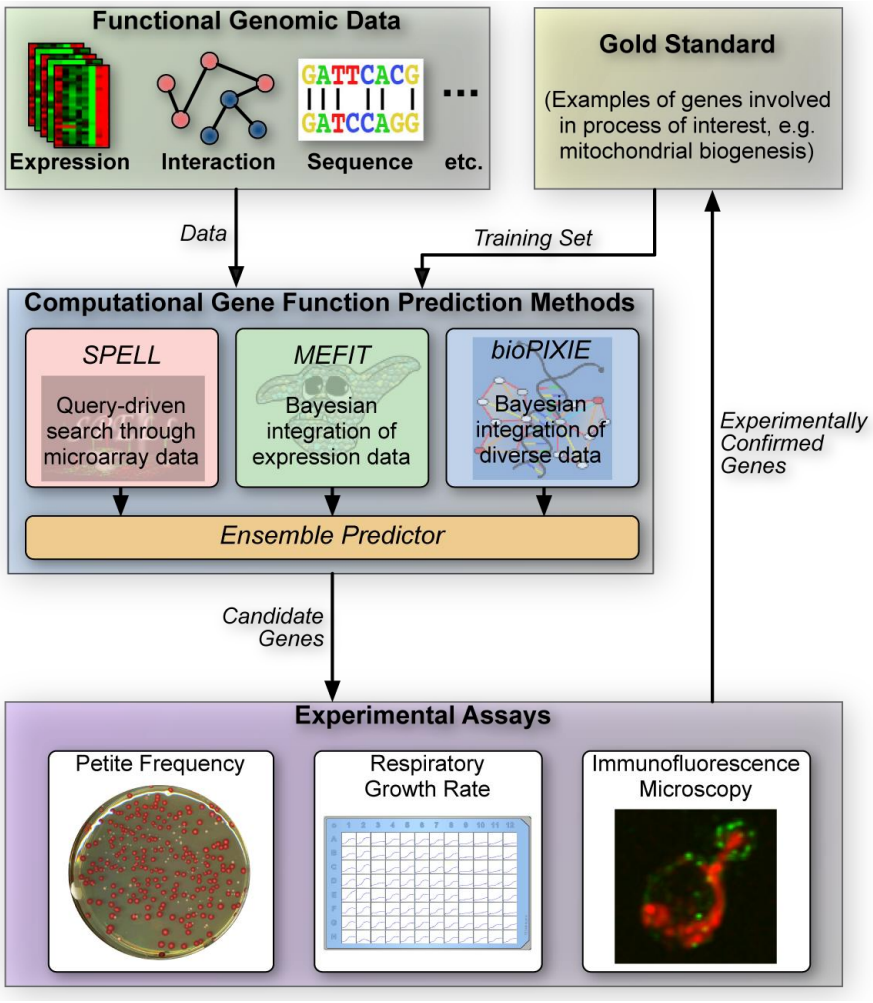


Figure 4: An overview of our iterative framework integrating computational and experimental methodologies for discovery of gene function. Our study uses an ensemble of computational gene function prediction methods (bioPIXIE (Myers, Robson et al. 2005), MEFIT (Huttenhower, Hibbs et al. 2006), and SPELL (Hibbs, Hess et al. 2007)), each of which predicts new genes involved in mitochondrial function based on high-throughput data and examples of known mitochondrial proteins (the gold standard). We selected test candidates by integrating these approaches based on estimated precision of each method and tested these predictions experimentally using three biological assays (see Methods for details). Upon evaluating these experimental results, the proteins newly discovered to be involved in mitochondrial function were added to the known examples, and the process was iterated to comprehensively characterize additional mitochondrial proteins. See Table 1 for an overview of our results.

prediction methods to target definitive, time-consuming experiments to a tractably sized set of candidate proteins can result in the rapid discovery of new functional roles for proteins. Our results also show that most mutants resulting in severe respiratory defects have already been discovered. This is likely to be the case for mutant screens in many fundamental biological processes, because saturating screens have discovered mutations with strong phenotypes. Even in a well-studied eukaryote like *S. cerevisiae*, there are many processes that need to be fully characterized by identifying all proteins required for its normal function (Pena-Castillo and Hughes 2007). As such, most of the remaining undiscovered protein functions are only identifiable by rigorous, quantitative assays that can detect subtle phenotypes, such as those used by our study.

Results

We utilized an ensemble of computational gene function prediction approaches to systematically identify candidates for involvement in mitochondrial inheritance. These candidates were experimentally assayed, and the confirmed predictions were then utilized as inputs for a second round of prediction and experimentation. A schematic overview of this approach is shown in Figure 4.

An ensemble of computational function prediction methods was used to iteratively target experiments

We trained an ensemble of three computational prediction methods (bioPIXIE (Myers, Robson et al. 2005; Myers and Troyanskaya 2007), MEFIT (Huttenhower, Hibbs et al. 2006), and SPELL (Hibbs, Hess et al. 2007)) using genomic data that we collected from many sources and a set of 106 genes known to be involved in *mitochondrial organization and biogenesis* based on published experiments as curated by the *Saccharomyces* Genome Database (SGD) (Issel-Tarver, Christie et al.

2002). Genes are assigned by SGD to this biological process if published experiments have definitively demonstrated functions involved in the formation, assembly, or disassembly of a mitochondrion. This includes genes that affect mitochondrial morphology and distribution, replication of the mitochondrial genome, and synthesis of new mitochondrial components.

An intuitive description of our computational methods is that each employs "guilt by association" to identify genes exhibiting similar data patterns to the genes used for training (further details in Methods). The ensemble was used to rank all genes in the genome from most likely to be involved in mitochondrial inheritance to least likely. We selected the top 183 most confident genes that were not included in the training set for experimental validation. Of these, we found existing experimental literature evidence of involvement in mitochondrial inheritance for 42 proteins, and as such we included these in our set of positive controls (along with 6 genes from the training set). The remaining 141 proteins comprised our set of first iteration predictions, as none of these proteins appeared in published experiments that demonstrated their requirement for proper mitochondrial inheritance. We assayed these predicted genes experimentally as described below. We then augmented our training set of genes known to be involved in mitochondrial inheritance with the experimentally verified predictions (using both our experiments and the uncurated published literature, see methods) and repeated this process to generate a second iteration of predictions. From this second iteration, we selected the 52 most confident predictions that were not previously tested and performed the same experimental assays.

Petite frequency assay quantitatively detected defects in mitochondrial inheritance

In order to confirm the potential roles of our candidate genes in mitochondrial biogenesis and inheritance, we employed an experimental assay that measures the rate of generation of cells

lacking respiratory competent mitochondria (called "petite" cells (Ogur and St John 1956)). This assay reliably detects defects in mitochondrial inheritance, but it is too time consuming to perform on a whole genome scale. Wild-type yeast from the S288C genetic background produces petite daughter cells at a baseline frequency of ~20% (Baruffini, Ferrero et al. 2007), but mutation of genes involved in mitochondrial inheritance and biogenesis can significantly alter this rate.

We measured the frequency of petite formation for single gene deletion strains of all 193 candidate genes (141 from the first iteration, 52 from the second) and for 48 positive control genes. To reduce the effects of suppressor mutations and aneuploidy associated with the yeast deletion collection (Game, Birrell et al. 2003), we sporulated the heterozygous Magic Marker deletion collection (Pan, Yuan et al. 2004) and isolated six independent haploid deletion mutants for every gene tested. Individual deletion strains were grown in media requiring aerobic respiration for growth (glycerol), and strains completely unable to grow were deemed respiratory deficient and did not continue in the assay. The remaining mutants were then assayed by measuring the ratio of petite cells to total cells in a colony founded from a single cell. At least twelve matched wild-type sporulation isolates were assayed on each day of experiments in order to establish baseline frequencies. For each gene tested, petite frequencies were measured for at least eight colonies and compared to the distribution of wild-type frequencies measured in parallel on each day of experiments, which allowed us to quantitatively detect subtle phenotypes with statistical rigor. A schematic of this assay is shown in Figure 5 and further details are available in Methods.

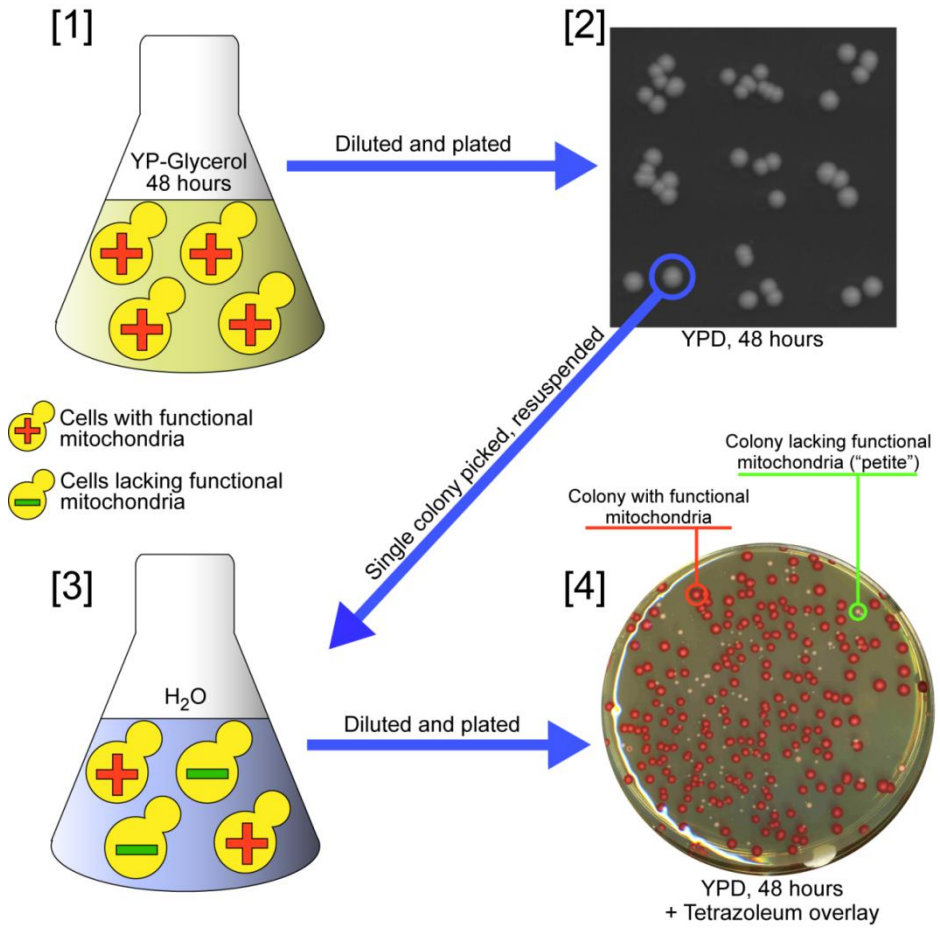


Figure 5: Schematic overview of the petite frequency assay. 1) Initially, strains were grown in a non-fermentable carbon source (liquid YP-Glycerol) for 48 hours. All cells growing under this condition must be respiratory competent and contain functional mitochondria. Any strain with no viable cells after this step was deemed respiratory deficient and did not continue in the assay. 2) Cell cultures were serially diluted and plated on a fermentable medium (YPD) and grown for 48 hours to form colonies founded from a single cell. At this point, the requirement for respiratory competency is lifted, so that daughter cells can survive while losing respiratory function. 3) A single colony is picked and briefly re-suspended in water. 4) The suspension is diluted and plated on YPD and grown to form colonies founded from single cells. After 48 hours, soft agar containing tetrazolium is overlaid on the plates. Colonies founded by respiratory competent cells will take up the tetrazolium and appear large and red. Colonies founded from respiratory deficient cells ("petite" colonies) appear smaller and white.

Computationally-driven experimentation discovered a new role in mitochondrial inheritance for 109 proteins

In our first iteration of prediction and experimental testing, 82 of our initial 141 predictions (58%) were confirmed to play a role in mitochondrial inheritance as they exhibited a significantly altered petite frequency rate compared to the wild-type distribution (FDR corrected Mann Whitney U-test p-value < 0.05; see Figure 6A). These 82 newly confirmed predictions were added to the training set, and we then performed another iteration of prediction and experimentation. In this second iteration, 17 of the 52 predictions (33%) were experimentally confirmed (Figure 6A). Based on the second iteration predictions, we also examined a targeted set of double knockout mutants and experimentally confirmed 10 more proteins that exhibit synthetic petite frequency defects (full details below). Further, the petite frequency assay demonstrated a high level of sensitivity as 44 of our 48 positive controls (92%) exhibited a significant phenotype (the remaining 4 are discussed further below). All together, after both iterations of our approach we discovered a role in mitochondrial inheritance by demonstrating significant phenotypic alterations for 109 of our 193 (56%) total predictions (see Table 1 for breakdown).

These newly characterized functions include 42 genes with other previously known functions (not in mitochondrial inheritance) and 67 genes with no previously characterized cellular role. For example, we observe that mutation in the functionally uncharacterized TOM71 causes a 44% increase in petite frequency. While Tom71 has been co-localized with the translocase complex responsible for protein import through the mitochondrial outer membrane, previous work (largely in vitro) has not identified a strong functional defect associated with Tom71 in translocase activity (Schlossmann, Lill et al. 1996). Our confirmation that TOM71 significantly affects mitochondrial inheritance rates strongly suggests that it does indeed play a role in

mitochondrial import, at least for some subset of proteins required for mitochondrial inheritance or biogenesis. The identification of a functional role for 67 previously uncharacterized proteins is particularly striking as this covers roughly 1 in 18 of the remaining ~1,200 proteins in yeast that still have no known functional roles (Pena-Castillo and Hughes 2007).

Subtle phenotypes are predominant among our new discoveries

We observed a striking difference in the severity of petite frequency phenotypes in single gene knockouts between the confirmed gene predictions and the positive controls (Figure 6B). Of the 44 positive controls demonstrating a significant phenotype, the majority exhibited a complete loss of respiratory function (28 of 44, 64%) as opposed to the more subtle phenotype of altered mitochondrial inheritance (16 of 44, 36%). The proportions of subtle and severe phenotypes were reversed in our predictions experimentally confirmed by single gene knockouts, in which 79 of 99 mutants (80%) showed altered mitochondrial inheritance while only 20 of 99 mutants (20%) were

	Number tested	**Number with significant mitochondrial phenotype**
Positive Controls	48	44
First Iteration Predictions	141	82
Second Iteration Predictions	52	17
Synthetic Interaction Predictions	27	10

Table 1: Summary of results. We iteratively employed an ensemble of computational function prediction methods to select candidate genes for experimental testing. Confirmed predictions from the first iteration were added to the training set for the second iteration. Promising candidates for synthetic interactions were also selected after our second iteration for testing with double mutant assays.

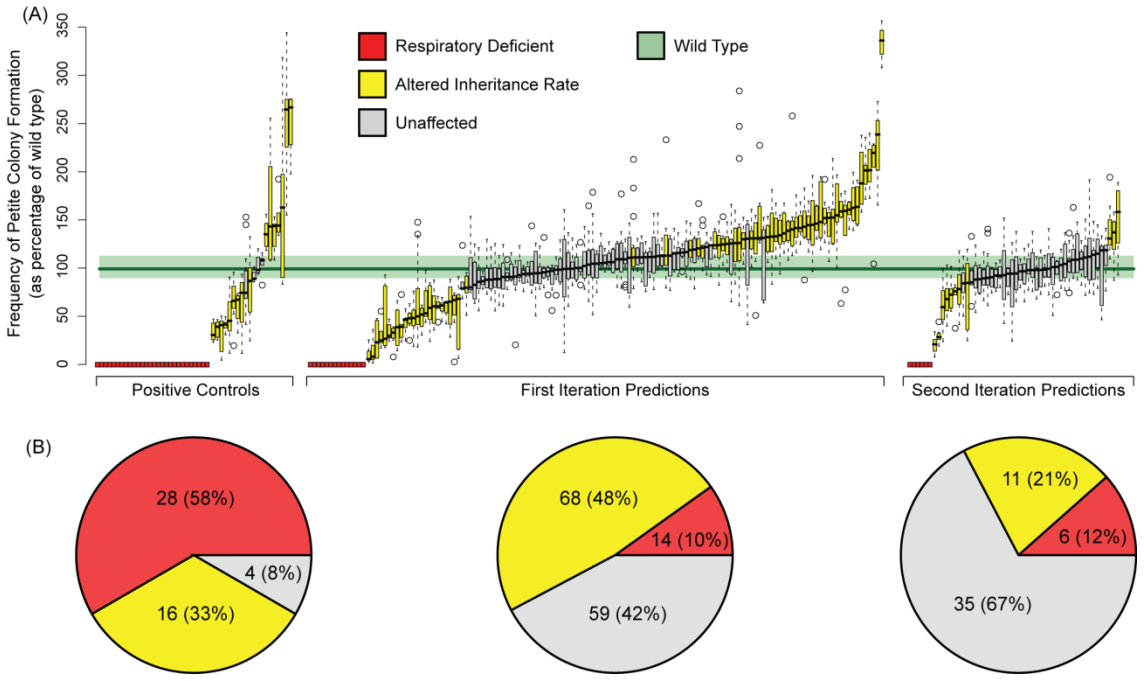


Figure 6: The combination of computational predictions and quantitative assays discovers novel genes involved in mitochondrial function. A) Mitochondrial inheritance rates of single gene knockouts were determined for 193 genes predicted to be involved in mitochondrial function and for 48 control genes known to be involved. A box plot is shown for each deletion strain tested; red indicates the inability to grow on a non-fermentable carbon source (glycerol), yellow indicates a mitochondrial inheritance rate significantly altered from wild type, and gray indicates no significant difference from wild type. Significance was determined using a Mann-Whitney U-test comparing at least 12 independent measurements of wild type to at least 8 independent measurements of each mutant strain. The green shaded region indicates one quartile above and below the median rate for all 358 wild type replicates. A total of 99 of the 193 prediction candidates were confirmed (an additional 10 genes were confirmed through double knockout analysis, see Figure 7). B) Distribution of petite frequency phenotypes among positive controls (left), first iteration predictions (center), and second iteration predictions (right) with colors as in A. Severe phenotypes (red) were more prevalent among positive controls, while the majority of confirmed predictions exhibited an intermediate phenotype (yellow). We hypothesize that this difference is due to a bias towards detection of extreme phenotypes in classical genetic screens and high throughput methodologies.

respiratory deficient. The quantitative nature of these phenotypes among our novel discoveries may indicate why they have not been previously associated with mitochondrial inheritance by either classical genetic screens or high-throughput techniques (Shutt and Shadel 2007; Perocchi, Mancera et al. 2008), which generally assay extreme rather than subtle phenotypes. In further support of this observation, since undertaking this study, 7 of our 99 confirmed candidates have been associated by other groups to mitochondrial inheritance (COA1 (Pierrel, Bestwick et al. 2007), IBA57 (Gelling, Dawes et al. 2008), GUF1 (Bauerschmitt, Funes et al. 2008), ATP25 (Zeng, Barros et al. 2008), QRI5 (Barros, Myers et al. 2006), GRX5 (Rodriguez-Manzaneque, Tamarit et al. 2002), REX2 (van Hoof, Lennertz et al. 2000)), and 3 of these 7 exhibited the most extreme phenotype of respiratory deficiency in our study.

Decreased petite frequency identifies petite negative mutants

Among our deletion strains exhibiting the subtle phenotype of altered petite frequency, we observed mutants with both statistically significant increases and decreases in frequency. Increased petite formation clearly indicates a failure in normal mitochondrial biogenesis or inheritance, while decreased petite frequency is a distinct phenotype referred to as "petite negative" (Chen and Clark-Walker 2000). Petite negative mutants display synthetic lethality or sickness with respiratory deficiency, which impairs the survival of petite cells and thus decreases their frequency. Known petite negative mutations occur in mitochondria-localized proteins that normally support the maintenance of the mitochondrial membrane potential in the absence of respiration (Chen and Clark-Walker 2000). Decreased petite frequency was observed in nine (19%) of our positive controls, two of which (FMC1 and PHB1) are known petite negative mutants (Dunn, Lee et al. 2006). Previously, traditional genetics and genome-wide screens have identified 21 petite negative mutations that result in synthetic lethality (Dunn, Lee et al. 2006).

Among our 99 discoveries in mitochondrial biogenesis from single gene knockouts, we found 32 additional mutants exhibiting a decreased petite frequency indicative of non-lethal synthetic interactions. Many of the characterized petite negative genes have roles in the assembly and turnover of ATP synthase complexes, and so these genes may be a rich target for further study (Dunn, Lee et al. 2006).

Computational iteration identifies candidates with redundant mitochondrial function verified through double mutant analysis

The confirmation rate from our second iteration decreased from our first iteration (58% to 33%), which suggests we may be nearing the limit of predicted genes that can be verified using the single knockout petite frequency assay. In particular, examining single gene deletion strains prohibits characterization of the roles of redundant proteins or genes that only exhibit synthetic phenotypes. In fact, all four of our 48 positive controls that did not exhibit a significant petite frequency phenotype are known to synthetically interact with at least one other gene involved in mitochondrial biogenesis and inheritance (Rep, Nooy et al. 1996; Mozdy, McCaffery et al. 2000; Saracco and Fox 2002; Sesaki, Southard et al. 2003). Furthermore, our most confident unconfirmed prediction (RMD9) was recently discovered to have a redundant role with Oxa1 in the processing and stability of mitochondrial mRNA (Nouet, Bourens et al. 2007). However, our second iteration prediction results indicate which of our unconfirmed predictions are worthy of further investigation with double mutant analysis or additional assays, particularly in light of additional localization evidence. Following the second round of computational prediction, 27 of the 59 initially unconfirmed predictions persisted as highly ranked candidate genes while the remainder decreased in confidence. Of these, 23 (85%, hypergeometric p-value $<10^{-9}$) candidates

are known to localize to the mitochondria, while only 1 of the remaining 32 unconfirmed candidates (3%) is similarly localized.

To test the hypothesis that these 27 high-confidence unconfirmed predictions represented genes that had redundant mitochondrial function, we performed targeted double mutant analysis looking for synthetic interactions. We chose 4 deletion mutants (*aim17*Δ, *rvs167*Δ, *tom6*Δ, and *ehd3*Δ) confirmed to be involved in mitochondrial inheritance with modest petite frequency phenotypes to cross with these 27 candidates. Choosing mutants with modest phenotypes was necessary to allow for a strong synthetic interaction to be observed. We tested 103 double mutant strains and observed 15 significant synthetic phenotypes (FDR corrected Wilcoxon rank-sum p-value < 0.05) spanning 10 of 27 mutants that did not display a single mutant phenotype (Figure 7). While some of our double mutants exhibit suppression, we did not focus on these interactions because of the modest nature of the single mutant phenotypes. Instead, we focused on synthetic defects that we could rigorously define as the double mutant petite frequency being significantly different from both single mutants and the wild-type petite frequency. Of the genes exhibiting significant double mutant phenotypes, 2 were synthetic respiratory deficient and 8 demonstrated altered petite frequency. Among these 10 genes showing synthetic phenotypes was the previously mentioned RMD9. An *rmd9*Δ is hypothesized in Nouet et al. to have a pleiotropic effect on mitochondrial biogenesis (Nouet, Bourens et al. 2007). Our results support this hypothesis as RMD9 has synthetic respiratory deficiencies in all 4 of the double mutant strains we examined (Figure 7). In contrast to *rmd9*Δ, the remaining 9 genes showed more specific patterns of synthetic phenotypes, as 7 interacted with only 1 of the 4 known mitochondrial inheritance genes used to generate double mutants. These specific synthetic interactions suggest the functions these genes may perform in mitochondrial inheritance. For example, the four genes

(AIP1, MPM1, YDL027C, and YDR286C) that specifically interact with *rvs167*Δ are potentially involved in the actin-based transmission of mitochondria to the daughter cell, as Rvs167 is a regulator of actin polymerization (Balguerie, Sivadon et al. 1999). In fact, the only known actin-localized protein among our 27 candidates, Aip1, had a genetic interaction only with the *rvs167*Δ (Figure 7).

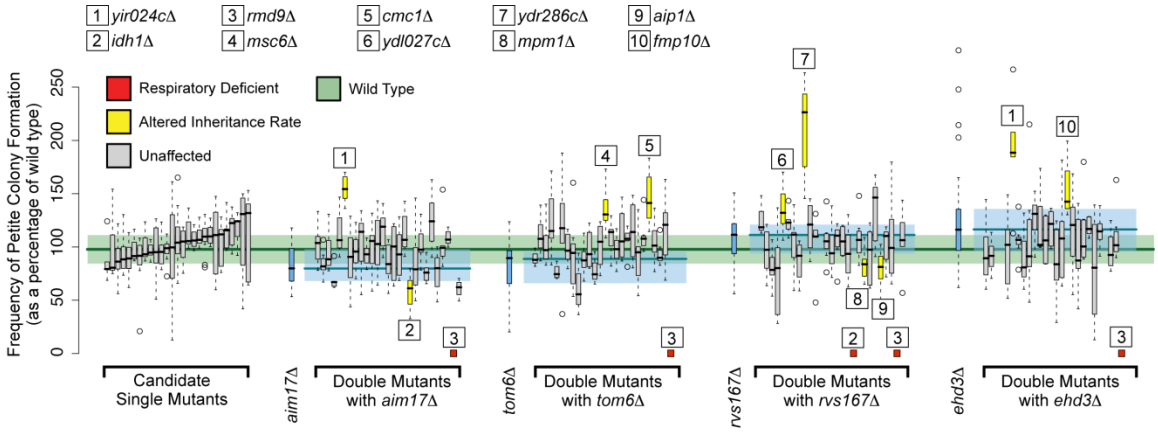


Figure 7: Double mutant petite frequency phenotypes. Based on their persistence as strongly predicted candidates during our second iteration, we selected 27 genes unconfirmed by single mutant analysis for investigation of synthetic phenotypes. The single mutant petite frequency is shown for each of these strains on the left. Each of the 27 strains was crossed with 4 genes known to be involved in mitochondrial inheritance (*aim17*Δ, *tom6*Δ, *rvs167*Δ, and *ehd3*Δ) to create ~100 double mutant strains. Results are shown in blue for each of the 4 strains crossed into, followed by all 27 double mutants constructed against that strain. The order of the double mutants is the same as in the 27 single mutants shown on the left. Colors are as in Figure 6. Significantly altered double mutant strains are marked with numbers, corresponding to the key above the box plots.

The high rate of synthetic phenotype recovery (10 out of 27 candidates tested) was made possible by the use of computation to limit the number of double mutants queried. There were 59 unconfirmed predictions from the first round of our analysis, and 95 genes tested in this study have the quantitative petite frequency phenotypes necessary for double mutant analysis. Combining these 95 confirmed genes with the 59 unconfirmed genes yields 5,605 possible double mutants to assay, which is far too large to reasonably test with the quantitative petite frequency assay. However, we used computation in two ways to reduce the number of double mutants screen to ~100. First, we used computational iteration to identify the subset of unconfirmed predictions most likely to be involved in mitochondrial inheritance. Second, we used the functional networks generated by the bioPIXIE algorithm (Myers, Robson et al. 2005) to select four genes from different sub-functions in mitochondrial inheritance. This allowed us to test less than 2% of the possible double mutants, but still identify phenotypes for 10 of 27 candidates (37%) due to the efficiency of our computational approach.

Computationally targeted experiments characterize new protein functions regardless of known localization

While we expect high correlation between localization to the mitochondria and involvement in mitochondrial inheritance, many non-mitochondria-localized proteins are vital for regulating mitochondrial function and inheritance (Boldogh and Pon 2007). Thus, a candidate gene approach based solely on protein localization would neglect many important participants in normal mitochondrial biogenesis. Our use of computational predictions to drive experimental discovery is unbiased with respect to any one genomic feature or assay. In this study, 47 (43%) of our 109 newly confirmed discoveries are not known to localize to the mitochondria (Issel-Tarver, Christie et al. 2002; Prokisch, Scharfe et al. 2004) and would have been overlooked in a screen of

mitochondria-localized proteins lacking known functions. Further, the accuracy of our predictions for non-mitochondria-localized proteins is comparable to that for mitochondria-localized proteins (44% vs. 59%, respectively). Thus, computational predictions can broaden the scope of potential discoveries beyond a more restricted candidate gene approach based on a single experimental technique or data source.

AIM21 is required for proper mitochondrial motility

Specific examples of non-mitochondria-localized proteins critical for mitochondrial biogenesis include proteins linking mitochondria to the actin cytoskeleton. Several of our novel discoveries have literature evidence associating them to the actin cytoskeleton but no evidence suggesting a role in mitochondrial inheritance (Huh, Falvo et al. 2003; Gavin, Aloy et al. 2006; Collins, Miller et al. 2007). One of these genes, the uncharacterized ORF YIR003W (AIM21), has been shown to co-localize with actin in high-throughput studies (Huh, Falvo et al. 2003) and was predicted as an interactor with the actin cytoskeleton with high confidence by our system bioPIXIE (Myers, Barrett et al. 2006). We found that strains carrying a deletion of YIR003W grow normally on glycerol but form petites at a frequency of 166% of wild type cells, one of the highest petite frequencies observed in our experiments.

To better understand the mitochondrial inheritance defect in this mutant, we used our computational predictions to direct experiments targeting the role of the actin cytoskeleton in mitochondrial inheritance. The morphology of the actin cytoskeleton and of the mitochondria in this mutant was visualized by dual immunofluorescence (Figure 8A and B). In the *yir003w*Δ mutants, the actin skeleton appears relatively normal, with typical polarization of actin patches toward the daughter, and the mitochondria show no gross perturbation in these mutants. However, by observing sustained mitochondrial movement events, we assessed mitochondrial

motility for this mutant and found severe defects comparable to a *puf*3Δ strain (Figure 8C), a gene known to be involved in mitochondrial motility (Garcia-Rodriguez, Gay et al. 2007). Even though this mutant displayed no overt morphological phenotypes, detailed analysis of YIR003W uncovered a subtler, specific defect in mitochondrial motility.

Respiratory growth fitness indicates specific roles in mitochondrial biogenesis

To further characterize our predictions, we assayed single gene knockout mutants for respiratory growth defects, as assembly of the complexes required for respiration is a critical step in mitochondrial biogenesis. We quantitatively measured growth profiles of most of our single gene deletion mutants under respiratory growth conditions (glycerol) comparing them to growth in fermentative conditions (glucose) as a control. A 96-well plate incubator and optical density reader was used to determine growth profiles for six independent replicates of each deletion strain tested and for two matched wild-type isolates of each strain (24 control wells per plate, see Methods for details). Exponential growth rates and saturation densities were calculated for each strain (Figure 9A), and both of these parameters were assessed for statistical significance relative to the distribution of all wild-type controls. Significant phenotypes were only reported if the defect was unique to the glycerol growth condition (i.e. was not present in the glucose growth curve) in order to ensure that the growth defect is respiration specific. By combining the growth rates and saturation densities (Figure 9C), we arrived at a respiratory growth phenotype that classifies each mutant as severe, moderate, weak, or unaffected. An example growth curve of each class is shown in Figure 9B.

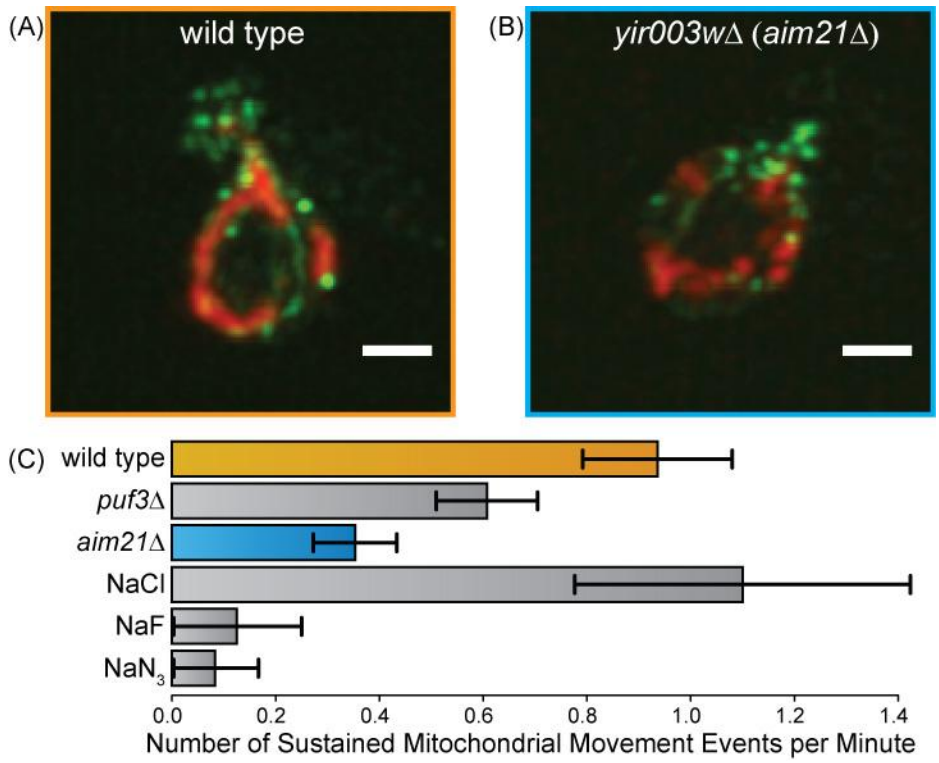


Figure 8: YIR003W (AIM21) is required for mitochondrial motility. A, B) Dual immunofluorescence of mitochondria (outer membrane protein porin stained in red) and actin (total actin, stained in green) in the indicated yeast strains (scale bar 2μm). C) Mitochondrial motility was measured in strains carrying an integrated mitochondrially-targeted GFP by tracking the movement of the tip of a mitochondrion within a budding cell every second for two minutes. A sustained mitochondrial movement is defined as movement in the same direction for at least three consecutive seconds. PUF3 is a gene with known involvement in mitochondrial motility (Garcia-Rodriguez, Gay et al. 2007). To determine the frequency of sustained mitochondrial movement resulting from Brownian motion or other passive processes, sustained mitochondrial movement was measured in the presence of the metabolic inhibitors sodium azide (NaN_3) and sodium fluoride (NaF). 10mM concentrations of these inhibitors were compared to a control of 10mM sodium chloride (NaCl). Due to its lack of static actin or mitochondrial phenotypes, AIM21's motility defect would be difficult to find without integrative computational predictions driving specific experimental assays.

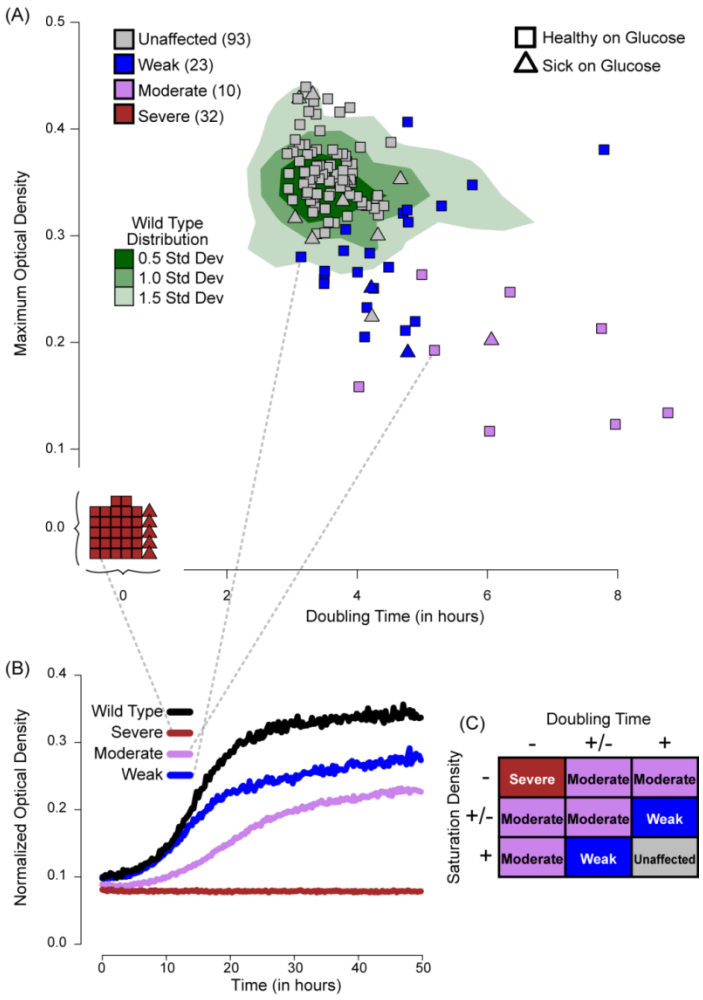


Figure 9: Respiratory growth phenotypes. A) Scatter plot of growth rate and saturating density measured from growth curves in minimal non-fermentable media. The vertical axis indicates the maximum (saturating) optical density achieved by the strain, and the horizontal axis represents the estimated doubling time based on an exponential fit to the growth curve. Green shading indicates the distribution of all 536 wild type measurements. Triangles represent strains with saturation density and/or doubling time significantly altered on glucose, while squares represent strains that showed normal growth on glucose. Each point is colored by the strength of its respiratory growth phenotype (see part C). B) Example growth curves for wild type and strains representing each of the three phenotypic classes: weak, moderate, and severe respiratory growth defects. C) Determination of respiratory growth phenotype. Each growth parameter (saturation density and doubling time) was statistically scored as no effect (+), intermediate effect (+/-), or extreme effect (-). The combination of saturation density and doubling time results produces a final respiratory growth phenotype, with maroon representing a severe defect, purple a moderate defect, blue a weak defect, and gray no defect.

As expected, nearly all mutants classified as respiratory deficient in the petite frequency assay were classified as severely defective in the respiratory growth assay. However, we also observed significant respiratory growth phenotypes for 29 mutants without previously reported respiratory impairments in the literature. Of these, 22 exhibited a weak or moderate defect that may have been difficult to observe in whole-genome screens assaying respiratory growth (Giaever, Chu et al. 2002; Steinmetz, Scharfe et al. 2002); the remaining 7 severe phenotypes might have been previously overlooked due to suppressor mutations in the systematic deletion collection. While employing multiple replicates in such assays lowers overall throughput, these results suggest that testing many replicates enables more complete discovery of subtle respiratory growth phenotypes.

Mitochondrial biogenesis and respiratory growth are partially overlapping processes that intersect in the translation and assembly of respiration complexes. As such, 55 of the 67 assayed mutants (82%) that exhibited an altered petite formation frequency had only weak or unaffected phenotypes in the respiratory growth assay (Table 2). The remaining 12 mutants exhibiting altered inheritance rates were classified as either severe or moderate in the respiratory growth assay; thus, these mutants demonstrate both an inheritance defect and a strong defect in respiration. These include four positive controls (CIT1, COX14, FMC1, and MRP49) known to be directly involved in the translation and assembly of respiratory complexes (Suissa, Suda et al. 1984; Fearon and Mason 1992; Glerum, Koerner et al. 1995; Lefebvre-Legendre, Vaillier et al. 2001). Additionally, since the beginning of this study, two of the eight additional genes in this class (MAM33 and COA1) have been shown to function in aerobic respiration (Muta, Kang et al. 1997; Mick, Wagner et al. 2007; Pierrel, Bestwick et al. 2007). This suggests that the remaining six genes newly characterized by this study (AIM8, AIM23, AIM24, AIM34, CTK3, and UBX4) are

also functioning in the assembly of respiration complexes. Thus, performing additional experimental assays allows us to determine more specific roles several genes that we have newly associated with mitochondrial inheritance.

Computationally directing experimental efforts can accelerate discovery rates

We employed thorough assays performed in replicate in order to detect important but subtle phenotypic variations. As such, it is impractical to scale these assays to the entire genome at the same level of rigor. In fact, given our rate of experimental efforts, it would require nearly 7 years for us to apply the petite frequency assay to all viable single gene deletion strains. However, by using computational predictions of protein function as a form of initial genetic screen, we were able to target our efforts towards the most promising candidates first. This is important for testing single gene deletions, but it is imperative for assaying potential synthetic defects. There are 18 million possible double gene knockouts in *Saccharomyces*, a number far too large to comprehensively test for a broad range of phenotypes. However, we were able to discover 15 synthetic mitochondrial inheritance defects by assaying a small, computationally chosen fraction of this available space. In all, by utilizing computational predictions of proteins involved in mitochondrial inheritance, we have rapidly characterized new functional roles for 109 genes.

Discussion

We have used computational predictions of gene function to direct focused, non-high-throughput laboratory experiments, confirming 109 proteins required for normal mitochondrial inheritance in *S. cerevisiae*. These discoveries include 67 genes with no previously known function (5% of the remaining ~1,200 uncharacterized *S. cerevisiae* genes) and 47 proteins not currently known to localize to the mitochondria. For several genes, our results provide evidence of involvement in

		Petite frequency phenotype	
Respiratory growth phenotype	**Respiratory deficient**	**Altered inheritance rate**	**Unaffected**
Severe	28	4	0
Moderate	0	8	2
Weak	2	12	9
Unaffected	0	43	50

Table 2: Overlap between petite frequency and respiratory growth phenotypes. We observe that the majority of single deletion strains deemed respiratory deficient in our petite frequency assay exhibited severe respiratory growth defects as well. Interestingly, 12 mutants with altered mitochondrial inheritance rates exhibited either severe or moderate respiratory growth defects, indicating that these genes may be involved in respiratory complex assembly.

specific sub-processes of mitochondrial inheritance (e.g. AIM21/YIR003W in mitochondrial motility). No previous study has systematically tested computational predictions using non-high-throughput laboratory techniques; the 56% accuracy established by our study demonstrates the potential of such computationally driven genetic investigations for direct future biological discoveries. Of our newly characterized mitochondrial genes, 51 have strictly defined human orthologs, 5 of which are associated with known disease.

Computation identifies subtle phenotypes confirmed by experimentation

Computational function prediction and non-high-throughput laboratory experiments complement each other in another important way highlighted by these results: the combination of these two techniques can rapidly identify subtle, quantitative phenotypes that are difficult to detect with high-throughput assays. When investigating well-studied processes (such as mitochondrial biology), most genes for which loss of function completely disrupts the process have already been discovered, since such extreme phenotypes are relatively easy to detect. This is evidenced by the strong enrichment for severe phenotypes among our positive control set. Many important biological functions also tend to be redundant, such that disruption of a single gene

results in only a mild (but quantifiable) perturbation of the process rather than loss of function. This is likely to be even more prevalent in higher organisms, which employ far more redundancy than does *Saccharomyces*, and it is also key to understanding the molecular mechanisms of many diseases. Deletion of yeast orthologs of human mitochondrial disease genes is significantly more likely to cause a modest respiratory growth defect than a severe defect (Perocchi, Mancera et al. 2008); similarly, since aerobic respiration is essential for mammalian viability, many disease-related mutations are unlikely to completely disrupt human mitochondrial function. Rather, these mutations tend to cause diseases by partially compromising the mitochondria (Shutt and Shadel 2007). Recently, (Fan, Waymire et al. 2008) compared several mouse models of mitochondrial disease, and found that subtle mutations caused disease in adult animals, while more severe mutations were suppressed at a high frequency. Subtle defects accrued over time have also been of increasing recent interest as related to aging in human beings (Lambert and Brand 2007). As the field continues to investigate the molecular basis of human disease and aging, the relationship between diseases and mutations incurring subtle functional perturbations is likely to extend far beyond mitochondrial biology.

Computational approaches quickly provide accurate, unbiased predictions of protein function

Using computational techniques to generate candidate gene lists for further investigation has several advantages relative to individual high-throughput experimental screens with comparable accuracy. First, computational data integration has the capacity to take advantage of large collections of existing publicly available experimental data; this can reveal information on a process of interest (e.g. mitochondrial function) by simultaneously examining many previous results. Additionally, computational predictions can often be generated in days or weeks, in contrast to the months or years required to conduct many traditional experimental assays.

Computational integration of multiple data sources can also be less biased to any one biological feature of the candidate genes. For example, high-throughput localization studies have identified hundreds of mitochondrial genes without known functions (Westermann and Neupert 2003; Prokisch, Scharfe et al. 2004), but this approach would have missed the 51 genes (~50%) discovered in this study that do not have known mitochondrial localization. This lack of bias assisted us in discovering functions for 67 of the uncharacterized genes in *S. cerevisiae*, all of which represent healthy and viable mutants in the yeast deletion collection with no extreme single mutant phenotype detected by previous screens. Thus, while genetic screens are important and valuable for candidate selection, particularly in areas with prior knowledge, computational prediction approaches integrating existing data are a viable, accurate alternative.

Mitochondrial phenotypes are unlikely to represent pleiotropic effects

The 51 genes we confirm to be necessary for mitochondrial inheritance that have no known mitochondrial localization raise the possibility that these mutants are somehow indirectly affecting inheritance. Several lines of evidence argue against this possibility. First, we expect that many of these 51 proteins will localize to specific cellular structures controlling inheritance outside of the mitochondria. For example, 13 of the 51 are known to localize to actin cytoskeleton and/or the bud neck, both structures that play intimate roles in mitochondrial inheritance. Of the remaining 38 proteins, 3 were computationally predicted to localize to the mitochondria by another study (Prokisch, Scharfe et al. 2004), 11 have no known localization, and 7 have only been localized to the cytoplasm by high-throughput microscopy (which does not exclude mitochondrial localization). Further study of these 38 proteins may identify as-yet-undiscovered mitochondrial localization or highlight the importance of other cellular processes necessary for

mitochondrial inheritance (e.g. transcriptional regulation of nuclear-encoded mitochondrial genes).

Extensions to specific mitochondrial sub-processes and other biological systems

Mitochondrial inheritance and biogenesis comprises a number of sub-processes that work together to ensure that new mitochondria are generated and segregated to a daughter cell. This starts with the nuclear genes encoding mitochondrial proteins being transcribed, translated, and targeted to the mitochondria for import. The mitochondria must also replicate its own genome and assemble the numerous membrane-bound complexes necessary for proper function. During inheritance, the mitochondria themselves move along actin cables to the bud neck, where they are then segregated between the mother and daughter cells. While additional work will be necessary to associate all of the proteins discovered in this study with specific sub-processes, we have already identified two groups with interesting potential responsibilities in mitochondrial inheritance.

The first group of 8 proteins (AIM8, AIM23, AIM24, AIM34, COA1, CTK3, MAM33, and UBX4), which are likely to be involved in cellular respiration, was identified by comparing our glycerol growth data with the petite frequency results. This comparison identified mutants with respiratory growth rates far lower than would be predicted by their petite frequencies alone, suggesting a proximal role in respiration. Though the components of the mitochondrial complexes that generate ATP have been identified for some time in yeast, extensive chaperone, assembly, and turnover machinery for these complexes remains to be fully elucidated. The assembly and maintenance of these respiratory complexes is thus a likely role for these 8 proteins. The second group consists of 11 genes known to be associated with the actin cytoskeleton, including AIM21 as described in Results. The biochemical functions of the other 10

proteins with respect to actin have been previously described (Ayscough and Drubin 1996; Riezman, Munn et al. 1996; Goode, Drubin et al. 1998; Sekiya-Kawasaki, Groen et al. 2003), but they had no previously known mitochondrial roles. For example, Cap2p has been characterized in vitro to bind the barbed ends of actin filaments and prevent further polymerization (Kim, Yamashita et al. 2004), but it has not been previously implicated in mitochondrial inheritance. Interestingly, many of this specific subgroup of actin-associated proteins have also been implicated in actin/membrane interactions for endocytic trafficking (Toret and Drubin 2006). This raises the intriguing possibility that these proteins have specialized in interactions between actin and intracellular membranes.

Our general approach can be successfully extended to other processes beyond mitochondrial inheritance in yeast and to other organisms. We have applied our computational ensemble (Myers, Robson et al. 2005; Huttenhower, Hibbs et al. 2006; Hibbs, Hess et al. 2007) to 388 other processes in *Saccharomyces* with promising results, and we report functional predictions for these processes. Computational methods have also been successfully applied in other organisms with readily available genomic data collections (Sharan, Ulitsky et al. 2007; Guan, Myers et al. 2008; Pena-Castillo, Tasan et al. 2008), and the iterative nature of our approach may be particularly useful in higher eukaryotes where current functional knowledge is relatively sparse.

These results demonstrate the utility of employing computation to direct quantitative, functionally definitive assays. Here, we have used this technique to newly confirm the involvement of 109 proteins in the process of mitochondrial inheritance in *S. cerevisiae* by assaying the frequency of petite colony formation. A subset of these proteins was also characterized using growth profiling and immunofluorescence microscopy, revealing participation in specific sub-processes of mitochondrial biogenesis. In particular, AIM21 was

shown to be required for proper mitochondrial motility, a discovery which would have been difficult to make without specifically targeted computational predictions. As these techniques can be naturally extended to additional organisms and processes, close integration of computational function prediction with experimental work in other biological systems promises to quickly direct experimenters to novel facets of their areas of interest.

Materials and Methods

Petite frequency assay

This protocol is adapted from the original petite frequency (Ogur and St John 1956) and tetrazolium overlay (Ogur, St. John et al. 1957) assays. For each mutant strain tested, we grew several replicates of the strain for 48 hours in liquid YP Glycerol at 30C (Amberg, Burke et al. 2005). Strains able to grow on glycerol were diluted and plated for single colonies on YPD plates, which releases the requirement for functional mitochondria. Thus, as colonies formed, cells without functional mitochondria were generated. When the colony is fully formed, it is a mixture cells with functional mitochondria and cells without functional mitochondria. We measured this ratio by re-suspending a colony and plating a dilution of this re-suspension such that 100-300 colonies are formed on a YPD plate. By overlaying with soft agar containing tetrazolium, cells with functional mitochondria were stained red, while cells without functional mitochondria remained white. The final mixture for agar overlay contains: 0.2% 2,3,5-triphenyltetrazolium chloride (available from Sigma), 0.067M phosphate buffer pH 7.0 and 1.5% bacto agar. The ratio of white cells to total cells gives the petite frequency. Eight independent petite frequencies (biological replicates) were measured for each strain tested. The distribution of these frequencies was compared to the frequency of petite generation in wild-type yeast. Strains identified as having the altered mitochondrial inheritance phenotype in this assay exhibit at least a 20%

change in petite frequency from wild type, and have a p-value of less than 0.05 when comparing the petite frequency distributions of that strain to the wild-type petite frequency distribution, using a Mann-Whitney U test.

Computational prediction ensemble methodology

The three computational systems employed in our study were bioPIXIE (Myers, Robson et al. 2005; Myers and Troyanskaya 2007), MEFIT (Huttenhower, Hibbs et al. 2006), and SPELL (Hibbs, Hess et al. 2007). Each was used to analyze genes involved in the GO biological process *mitochondrion organization and biogenesis* (GO:0007005). All methods were initially trained and/or evaluated using the 106 annotations to this process as of April 15th, 2007. Detailed descriptions of these methods can be found in their respective publications.

Identification of additional control genes with literature evidence

42 of our initial computational predictions had strong literature evidence for involvement in mitochondrial biogenesis and inheritance and were determined to be "under-annotated," meaning that they already had strong literature evidence for their involvement in mitochondrial organization and biogenesis, but were not yet annotated to the corresponding GO term. These 42 genes, along with 6 genes already annotated, were included as our positive control set of 48 genes. In most of these 42 cases the information was already curated by SGD in the form of annotations to other GO terms, such as *integral to the mitochondrial membrane* or *mitochondrial protein import*. In addition to these 42 genes, we identified an additional 95 genes that we believe have enough literature evidence to warrant their inclusion in this process without further laboratory testing, for a total of 137 under-annotated genes. All 137 of these genes were included in the training set for our second iteration of computational predictions.

Selection of prediction candidates for experimental testing

Novel candidates for laboratory evaluation were chosen on the basis of both the three individual computational approaches as well as the ensemble of their predictions. We limited ourselves to consider only those genes with viable knockouts available in the heterozygous deletion collection (Giaever, Chu et al. 2002). Furthermore, we chose to evaluate predictions to both genes with no previously known function as well as genes known to be involved in a biological process other than *mitochondrial inheritance and biogenesis*. We chose the 20 most confident genes of unknown function and the 20 most confident genes with existing annotations from each of the three individual methods for validation. Due to overlaps between the predictions of each method, there were 87 genes in this group; however, 20 of these genes we determined to be under-annotated and were tested as positive controls, leaving 67 genes used as novel candidates without any prior literature evidence. We then chose an additional 74 genes from the ensemble list of predictions with no previous literature evidence to arrive at our total of 141 test candidates in our first round of laboratory evaluation.

Iterative re-training, prediction, and verification

After our first round of testing, 82 of the 141 novel predictions were discovered to have involvement in *mitochondrial inheritance and biogenesis*. Combined with the original 106 annotated genes and the 137 genes identified as under-annotated, this results in a total of 325 genes. Each of the three computational methods was re-applied using this updated training set of 325 genes and the same procedure was used to form an updated ensemble list of predictions. We selected the 52 genes with the highest confidence from the updated results that were not previously tested for laboratory investigation. The petite frequency assay was used, and an additional 17 genes demonstrated a significant phenotype.

Double mutant construction and testing

Deletion alleles marked with the ClonNAT resistance gene (rather than the G418 KanMX resistant marker) were prepared for the four tested strains (*aim17*Δ, *rvs167*Δ, *tom6*Δ, and *ehd3*Δ). A ClonNAT marked *ura3*Δ allele was prepared as a control (all other strains contained a *ura3*Δ*0* allele. These ClonNAT restant strains contained the Magic Marker reporter (Pan, Yuan et al. 2004) as well as *can1*Δ and *lyp1Δ* mutations to reinforce haploid selection. These five strains were crossed to a set of deletion strains marked with the G418 resistant KanMX marker, and diploids were selected on YPD-G418-ClonNAT. The diploids were then sporulated as described for our single mutant assays, except that double mutants were selected on media containing G418 and ClonNAT, and three controls were isolated for each sporulation: G418 resistant mutant, ClonNAT resistant mutants, and wild-type strains. The petite frequency assay was applied to these double mutant strains as described above. Phenotypic calls were determined for the double mutants based on the significance of the difference between the distributions of petite frequencies of the double strain versus both of the corresponding single strains. If the FDR corrected joint Wilcoxon rank sum p-value of both of these comparisons was <0.05, and the distribution of the double mutant strain was significantly different from wild type, then we scored the double mutant as significantly altered.

Yeast strains and media

All *S. cerevisiae* strains used in this study are descended from the S288C derivative used for the deletion consortium project (Giaever, Chu et al. 2002). Methods for individual mutant manipulation are described below. Standard methods for media preparation were used as previously described (Amberg, Burke et al. 2005).

Deletion set manipulation

The Magic Marker heterozygous yeast deletion set (Pan, Yuan et al. 2004) was pinned from glycerol stocks onto enriched sporulation agar as described (Tong and Boone 2006). Single colonies developing on these random spore plates were re-struck for single colonies on the same medium and tested for presence of the G418 resistant KanMX marker (Pan, Yuan et al. 2004) to identify the spore as wild-type or a deletion mutant. Single colonies that grew from this re-streaking process were picked and arranged in 96 well plates containing YPD. Each set of strains for a given candidate gene of interest were placed in a single column (1-12 of a 96 well plate); mutant isolates were placed in the first six wells (A-F) and sister wild type isolates were placed in wells G and H. These 96 well plates were glycerol stocked.

Growth rate assay

Strains were measured for their ability to grow in both respiratory (2% glycerol as carbon source) and fermentative (2% glucose as carbon source) conditions in minimal media supplemented for auxotrophies. Cultures were grown at 30C. Growth curves were generated in a 96-well plate format (described above in "Deletion set manipulation") that tests 12 mutants per run. For each mutation tested, 6 independent deletion mutants of that gene were grown in separate wells. Twenty-four replicate wild-type strains were also present in each 96-well plate format. Plates were grown and measured using a Tecan GENios plate incubator and reader, which recorded densities every 15 minutes for 24 hours for glucose cultures and 48 hours glycerol cultures.

Growth rate data processing

Growth rates were derived from these curves by using Matlab to fit an exponential model:

$$y = a2^{bx}$$

For each well, this model was fit over the entire curve, the first 2/3, and the first half; whichever yielded the best fit was used in downstream analysis (to avoid plateau effects and to model only exponential growth). Wells with an adjusted $R^2<0.9$ were marked as non-growing, and growth rates for the remaining wells were determined by subtracting the row, column, and plate means for each well from the exponential parameter b, yielding a rate b' for each well. These b' parameters for each mutant strain were tested for significance against the total wild type population (excluding non-growing wells) using a Mann-Whitney U test. Significance was only considered for b' parameters indicating a slower growth rate than wild-type.

To detect colonies growing exponentially but with significant differences in fitness, smoothed maximum densities d were also calculated for all wells, consisting of the average of the optical density readings for the last five time points in each growth curve. From these, plate, row, and column averages were subtracted from each well, generating adjusted maxima d'. Mutants which did not double in optical density at least once (i.e. where d' was less than twice the baseline optical density) were considered to be non-growing. The remaining d' values for each mutant were again compared with the wild type values (excluding non-growing wells) using a Mann-Whitney U test. Significance was only considered for d' parameters indicating a lower saturation density than wild-type. Combined with the exponential rate tests, this assigned each mutant phenotypes in rich media and in glycerol of no effect, no growth, or significant sickness.

In either assay, mutants with inconsistent results (disagreement among more than one of the six replicates) were deemed inconclusive and marked as "mixed." Phenotypes were never assigned based on such mixed phenotypes. For a mutant to be classified as having a respiratory growth defect, that defect was required to be specific to the glycerol media (i.e. no phenotype in glucose).

If the mutant grew slowly in both glycerol and rich media, then it was not considered to have a defect in respiration.

Immunofluorescence

Yeast immunofluorescence was carried out using standard methods (Amberg, Burke et al. 2005). Briefly, strains were grown to exponential phase in synthetic complete medium, and fixed in freshly prepared formaldehyde for 1 hour at 30C. (Mutant strains were isolated from the Magic Marker deletion set as described above; FY4 was used as a wild type strain for comparison.) Cells were washed, digested with Zymolyase and attached to polyethyleneimine-coated coverslips. Cells were blocked with BSA, and exposed to an anti-porin antibody (Invitrogen, A-6449) and a guinea pig anti-yeast actin antibody (Mulholland, Preuss et al. 1994). Secondary antibodies were Alexa 488-conjugated goat anti-guinea pig (Invitrogen, A-11073) and Alexa 555-conjugated goat anti-mouse (Invitrogen A-31621). Coverslips were mounted in PBS/glycerol/phenlyenediamene. Microscopy was performed on a Perkin Elmer RS3 spinning disk confocal microscope with a 100x objective. Exposures were 1ms per slice, and Z-stacks were taken with a 0.15 um spacing, and images were deconvoluted using and assembled into 3D volumes using Volocity (Improvision).

F-actin staining

Phalloidin staining was performed according to Methods in Yeast Genetics (Amberg, Burke et al. 2005). Briefly, strains were growth to exponential phase in synthetic complete medium, and fixed in formaldehyde (EMS 15712-5) for one hour. F-actin was stained using Alexa 488 conjugated phalloidin (Invitrogen, A12379). Cells were deposited on polyethyleneimine-coated coverslips and mounted in PBS/glycerol/phenlyenediamene. Slides were imaged and processed as for immunofluorescence.

Integration of mitochondrial GFP into deletion strains

The NatMX cassette was cut from pAG25 (Goldstein and McCusker 1999) using NotI and liagated into the EagI site of pYX122-mtGFP, which expresses a mitochondrially targeted GFP (directed by the Su9 peptide) under the control of the triose phosphate isomerase promoter (Westermann and Neupert 2000). This construct was used as a template to PCR amplify the NatMX-mtGFP cassette using primers with 40bp homology to target the cassette for integration at the dubious ORF, YDL242W. This integration was performed in the strain Y5563 to create ACY50. ACY50 was then mated to the Magic Marker yeast deletion set (Pan, Yuan et al. 2004) and selected for haploid deletion mutants carrying the cassette as described (Tong and Boone 2006).

Mitochondrial tracking microscopy

Exponential phase cultures of *S. cerevisiae* in Yeast Synthetic Complete media (YSC) were plated onto glass slides with an agarose bed growth chamber made of low melt agarose and YSC media. The slides were covered with a cover slip and sealed using VALAP (Swayne, Gay et al. 2007). Cells were then imaged using a Perkin Elmer RS3 spinning disk confocal microscope with a 100x objective. Images of mitochondrial GFP fusions were taken using a laser emitting at 488 nm at 100% power with an exposure of 1s. Phase contrast images were taken using an exposure of 3ms. For all images, 2x2 binning was used and gain was set to 255. For each field of view, both an initial Z-stack of images and a time course were taken. Each Z-stack was taken at intervals of 0.2μm through the entire depth of the cells. The time course was taken in a single focal plane for two minutes at 1 frame per second.

To determine the frequency of sustained mitochondrial movement resulting from Brownian motion or other passive processes (Doyle and Botstein 1996), sustained mitochondrial movement

was also measured in the presence of the metabolic inhibitors sodium azide and sodium fluoride. These inhibitors were added to the YSC agarose used for imaging; 10mM concentrations of these inhibitors were compared to a control of 10mM NaCl.

Image processing: frequency of sustained mitochondrial movement

To measure mitochondrial motility in vivo, individual mitochondrial tips were tracked through each frame of a 2m time course using the Manual_Tracking plugin for ImageJ. Image files were randomly coded with numbers so that the identity of each imaged strain was not known to the investigator performing image tracking. Mitochondria tips for tracking were identified using the Z-stacks (which avoids selection of tubules that appear to be tips because other sections are out of plane). These Z-stacks were assembled into a Z-projection and merged with the phase contrast image using ImageJ to permit identification of budded cells. The selection criteria for mitochondrial tips were that the tip is initially present in the mother cell of a budded cell. In cases where both termini of a mitochondrion were available for tracking, the tip closer to the daughter cell was selected. The position of the bud neck was set as a reference and the position of a mitochondrial tip in the mother cell was plotted for each of 120 frames. The distance from the mitochondrial tip to the bud neck was calculated at each frame (in our imaging hardware, each pixel corresponded to 0.15μm). Mitochondrial movement in each frame was then calculated by subtracting the distances to the bud neck in two consecutive frames and dividing by the time interval of 1s. The tracking data was saved as a table. Sustained mitochondrial movement events consisting of 3 consecutive frames of motion towards (anterograde) or 3 consecutive frames of motion away (retrograde) from the bud were identified using a custom Perl script (Garcia-Rodriguez, Gay et al. 2007). The number of these sustained movement events per minute was calculated.

Ensemble predictions for additional biological processes

In addition to the computational predictions used to study mitochondrial biogenesis and inheritance in this study, we have applied the same prediction techniques to many additional biological processes in *S. cerevisiae*. We used each of our three computational methods to predict gene functions for 387 additional biological processes in the same manner as described above. In order to demonstrate that these methods are able to capture information about these processes, we have also calculated the average precision (AP) of the cross-validated results as:

$$AP_G = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{i}{rank_i}$$

where G is a group of genes known to be involved in a process, and $rank_i$ is the rank order of gene i in the prediction results.

Human ortholog identification

Orthology between yeast and human genes was based on orthologous clusters in the Homologene, Inparanoid (Remm, Storm et al. 2001), and OrthoMCL (Li, Stoeckert et al. 2003) databases as of June 2007. Each of these uses a published algorithm for determining clusters of orthologous genes, i.e. groups of genes thought to be conserved and perform near-identical functions in different organisms. We first took the union of these databases as applied to a core set of diverse organisms (yeast, human, mouse, fly, and worm), considering a gene pair to be orthologous if declared so by any of the three databases. This resulted in a set of unified orthologous clusters, from which we eliminated any cluster containing more than 50 genes. This resulted in 14,528 clusters spanning 61,702 genes in the five organisms, and from this set we report here the human orthologs of *S. cerevisiae* genes in our study.

Human ortholog disease-related gene identification

Disease related human orthologs were determined based on the manual curation of the Online Mendelian Inheritance in Man (OMIM) resource (Hamosh, Scott et al. 2005), and the automated text mining available through GeneCards (Rebhan, Chalifa-Caspi et al. 1997). We considered all of the OMIM curations valid, while we required at least 2 independent publication citations in GeneCards for a disease relation to be valid.

Localization determination

Both mitochondrial and actin localization was based on the Gene Ontology cellular component curation. For mitochondrial localization curation to the term GO:0005739: *mitochondrion* was used. In addition, six genes were marked as computationally predicted to the mitochondrion based on the study by Prokisch et al. (Prokisch, Scharfe et al. 2004). For actin localization curation to the term GO:0015629: *actin cytoskeleton* was used.

Methodological Ramifications: Optimizing Computational Techniques

Machine learning and data mining techniques have been applied to a wealth of genome-scale data to produce meaningful predictions of gene/protein involvement in biological processes and pathways (Pavlidis, Weston et al. 2002; Jansen, Yu et al. 2003; Owen, Stuart et al. 2003; Troyanskaya, Dolinski et al. 2003; Lanckriet, Deng et al. 2004; Lee, Date et al. 2004; Nabieva, Jim et al. 2005; Barutcuoglu, Schapire et al. 2006; Jaimovich, Elidan et al. 2006). As biologists have pursued novel findings in a wide range of organisms with finite experimental resources, these approaches have promised to direct experimental efforts toward the most likely targets, with the hope of greatly accelerating the discovery process (Kitano 2002; Hughes, Robinson et al. 2004).

However, surprisingly few large-scale experimental studies of gene function have been performed on the basis of computational predictions, despite their great potential to inform and guide such investigations. Perhaps as a result, data continue to be generated at a rate that outpaces the characterization of gene functions (Pena-Castillo and Hughes 2007).

This disparity between the computational and experimental aspects of gene function discovery may be due to a lack of clear demonstrations of the effectiveness of computation in directing laboratory efforts. The few experiments that have been directed by computational systems have generally been limited to confirming individual predictions of the functions of single proteins ((Jansen, Yu et al. 2003; Owen, Stuart et al. 2003) and work from our laboratory (Myers, Robson et al. 2005; Huttenhower, Hibbs et al. 2006; Hibbs, Hess et al. 2007)). No large-scale studies have been performed to fully explore the ability of computational methods to accurately assign functions to sizeable sets of uncharacterized proteins. Without such comprehensive evaluations, it remains unclear how computational methods can best be employed to guide experimental efforts in discovering novel biology.

To explore the biological considerations important for computational function prediction and to demonstrate the general power of computationally driving experimentation, we have performed a large, systematic study of computational predictions for proteins involved in mitochondrial organization and biogenesis in *S. cerevisiae*. Mitochondrial defects are implicated in a variety of human diseases (Foury 1997; Steinmetz, Scharfe et al. 2002), including neurodegenerative disorders (Babcock, de Silva et al. 1997; Koutnikova, Campuzano et al. 1997) and muscular diseases (DiMauro and Schon 1998), making them an interesting and relevant target for such a study. The biological mechanisms of mitochondrial biogenesis are largely conserved from yeast through humans (60% of mitochondrial yeast genes have a human ortholog), and as many as one

in five mitochondrial proteins are known to be involved in human disease (DiMauro and Schon 1998; Andreoli, Prokisch et al. 2004). Mitochondrial biology is understood well enough to provide a sufficient number of training examples for computational prediction methods, but it is also thought that at least a quarter of the proteins involved have not yet been identified (Sickmann, Reinders et al. 2003; Prokisch, Scharfe et al. 2004). Mitochondrial organization and biogenesis is thus an important and tractable area where computational methods can demonstrate their utility.

In this study, we have examined the biological nature of the predictions made by an ensemble of three computational methods, including supervised and unsupervised techniques that analyze a variety of underlying data. Above, we show and describe our biological results using these predictions to direct a suite of experimental tests, including our discovery of 99 additional proteins involved in mitochondrial inheritance. Here, we present detailed analysis of the computational methods and their predictions in order to explore the utility and effectiveness of computational function prediction methods. In particular, we demonstrate several novel observations and conclusions that can greatly impact the use of computational approaches for targeting laboratory experimentation.

First, our results demonstrate that while ~75% of yeast genes are already known to participate in at least one biological process or pathway, many of these genes may have multiple functions that have not yet been characterized. This further refutes the notion of "one gene, one function," and demonstrates that both characterized and uncharacterized genes are fruitful for further experimental investigation. Second, by comparison to a new experimental screen of 48 randomly selected genes, we show that using computational predictions to guide laboratory experiments can greatly increase discovery rates. Third, we demonstrate that the specific predictions made by computational approaches are highly dependent on both the algorithmic foundation and

underlying biological data utilized by those methods. As such, we show that using an ensemble of diverse computational approaches can increase the biological breadth of scope of predictions. Lastly, we demonstrate that by iterating phases of computational prediction and laboratory experimentation, we can greatly expand our knowledge of gene functions.

Results

Our study employed an ensemble of three diverse computational methods (bioPIXIE (Myers, Robson et al. 2005; Myers and Troyanskaya 2007), MEFIT (Huttenhower, Hibbs et al. 2006), and SPELL (Hibbs, Hess et al. 2007)) to predict novel genes/proteins involved in the process of mitochondrial organization and biogenesis. Each of these methods integrated high-throughput data sources and utilized existing biological knowledge from the Gene Ontology (GO) (Ashburner, Ball et al. 2000) and Saccharomyces Genome Database (SGD) (Cherry, Adler et al. 1998) to identify candidates for involvement. Briefly, bioPIXIE performs context-specific Bayesian integration of a diverse set of genomic data to predict pairwise functional relationships between genes. MEFIT also performs Bayesian integration, but is targeted to utilize gene expression microarray data. SPELL uses the same compendium of microarray data, but uses a similarity search algorithm to identify groups of related genes. The results of all three approaches were combined based on the estimated precision of each method to produce our ensemble predictions of gene function (further details are in the Methods section). Predictions for genes involved in mitochondrial organization and biogenesis were validated using a quantitative laboratory assay indicative of involvement in mitochondrial biogenesis and inheritance. The first round of prediction and evaluation used only existing GO annotations as a training set. We then performed a second iteration of this process after updating our training set to include gene

predictions confirmed in the first iteration. A schematic view of our system for prediction, verification, and iteration is shown in Figure 10.

The full biological results of this study are presented above. The next two paragraphs contain a brief summary of the results important to the further analyses and conclusions presented here. When the study was undertaken, 106 genes were annotated by SGD to the *mitochondrion organization and biogenesis* GO term (GO:0007005 as of 4/15/2007). These genes were used as input to the computational methods during the first iteration of testing. We initially evaluated our 184 most confident computational predictions, and 122 (66%) were validated as exhibiting a significant phenotype indicative of involvement in mitochondrial biogenesis. Upon further inspection of these confirmed predictions, we found existing literature evidence for 40 of these genes. By following this literature, we found evidence for 2 more of our tested genes and identified an additional 93 genes with strong evidence for mitochondrial function that had not yet been annotated as such by SGD. Many of these genes were annotated to specific categories related to mitochondrial organization (e.g. *integral to mitochondrial membrane*), but were not yet cross-annotated to the *mitochondrion organization and biogenesis* process. In all, we identified a total of 135 genes with existing literature evidence that were "under-annotated." We have presented this list to SGD and they are evaluating these observations using their established curatorial procedures; as of now, nearly half of these genes have been added to the annotations.

Our second iteration of prediction and validation used a set of 323 genes as input to the computational methods (106 original annotations, 82 newly confirmed genes with no prior literature evidence, and 135 under-annotated genes). We evaluated the 52 most confident predictions that were not previously tested, and 17 (33%) were validated. While this confirmation rate is still high, the reduction suggests that we may be nearing the edge of genes that can be

confidently identified using our assays (details below). Altogether, our study identified 234 new annotations to the process of mitochondrial organization and biogenesis, which more than triples the number of genes previously annotated to this area (Figure 11A). A summary of these results is shown in Table 3. While these biological results are striking and important, they also have significant ramifications in the application of computational techniques as a whole and in their integration with experimental biology, which we discuss in detail below.

Many genes with known functions also play additional cellular roles

A common metric for the level of characterization of an organism is the percentage of genes with at least one experimentally confirmed function (Hughes, Robinson et al. 2004; Pena-Castillo and Hughes 2007). By this metric, one might be led to believe that our functional characterization of some model organisms is nearing completion. For instance, in *S. cerevisiae*, we now have established functions for approximately three-fourths of the genome. However, we find evidence that suggests our current understanding is much more limited than these numbers suggest. Among our 194 tested predictions without existing literature evidence for involvement in mitochondrial biogenesis, 76 (39%) are known to be involved in at least one other process, while the remaining 118 (61%) have no previously known function. The verification rate for each of these classes was the same, as 39 of 76 (51%) genes with other known functions and 60 of 118 (51%) genes with no known function were confirmed to be involved in mitochondrial biogenesis. The notion of "one gene, one function" is clearly not consistent with these findings, and we suspect that both uncharacterized genes and genes with previously known functions are fruitful areas for exploration. This issue is even more important when considering higher eukaryotes, where protein variants encoded by the same gene may participate in multiple, diverse functions (Kochetov, Sarai et al. 2005; Blencowe 2006).

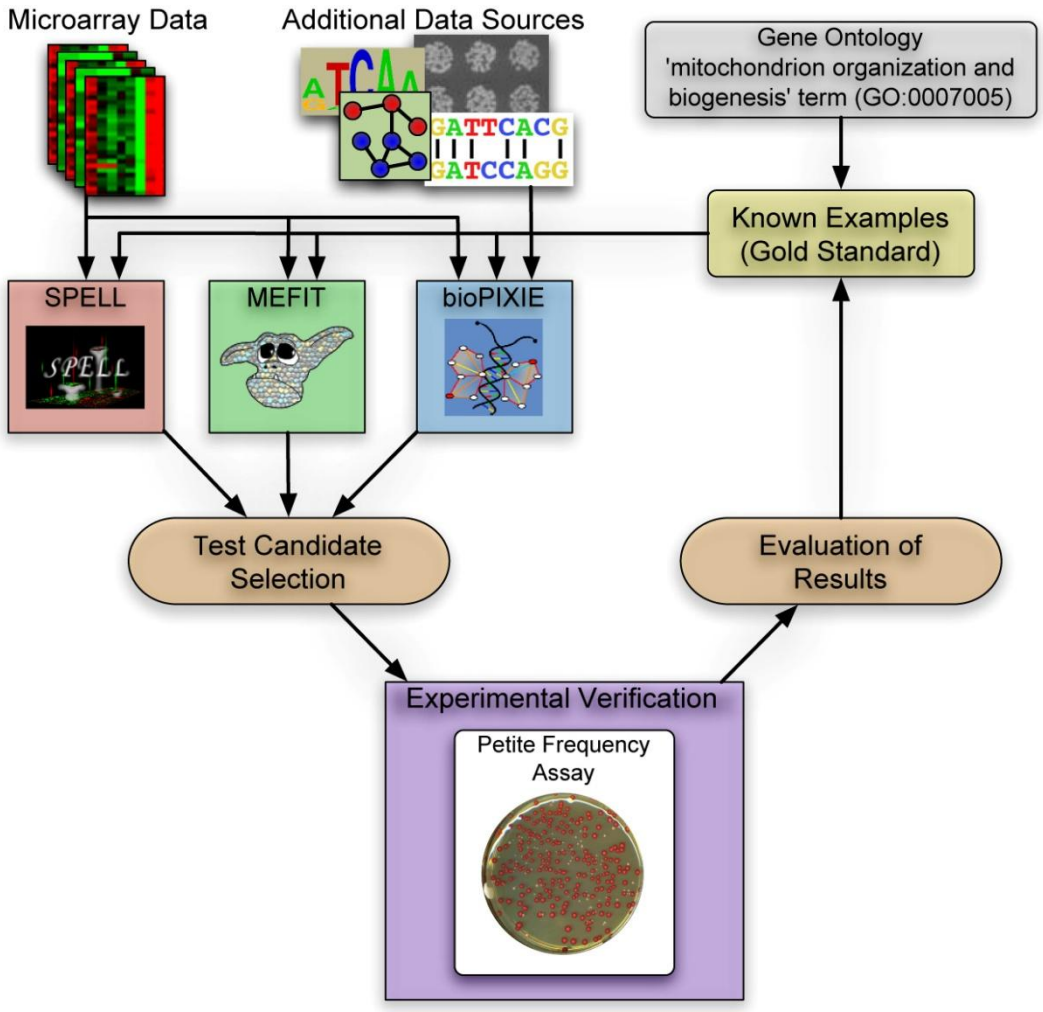


Figure 10: An overview of our iterative approach integrating computational and experimental methodologies. Our study uses an ensemble of computational gene function prediction methods (bioPIXIE, MEFIT, and SPELL) trained and evaluated on known biology to predict novel annotations to the GO term *mitochondrial organization and biogenesis*. We selected test candidates based on these computational predictions and validated these novel predictions experimentally using a quantitative, statistically verifiable biological assay. Upon obtaining the results of these tests, the set of known examples was augmented with the validated predictions, and the process was repeated to further explore this biological process.

Interestingly, there is a strong enrichment for components of the actin cortical patch among the 39 genes newly characterized in mitochondrial biogenesis that also have previously known functions (9 of 39 genes, hypergeometric $p<10^{-10}$). Most genes with known functions specifically related to mitochondrial biogenesis are not included in this number, since they were explicitly reported as under-annotations and treated as positive controls for our study. Though the actin cytoskeleton is known to be involved in mitochondrial motility in *Saccharomyces*, the precise mechanism of attachment and movement has remained elusive (Boldogh and Pon 2007). The enrichment of actin cortical patch components is particularly notable since the actin cortical patch has no explicit role in mitochondrial inheritance, but these nine genes are associated with cellular machinery known to move other membrane-bound organelles to daughter cells (Moseley and Goode 2006). Our predictions thus provide evidence that the same machinery may be employed during mitochondrial inheritance in a context similar to, but independent from, their cortical patch roles. By elucidating additional novel functions for previously characterized genes, we not only gain a greater understanding of each protein's individual responsibilities within the cell, we also form a more complete picture of higher-level interactions between cooperating pathways and processes.

These results are particularly striking within the historical context of the rates at which gene functions have been characterized. Since the full sequence of *S. cerevisiae* was published in 1996 (Goffeau, Barrell et al. 1996), nearly 3,000 genes have had their first known function characterized, while only ~1,700 genes have had a second function characterized (Figure 12). It remains unknown how many genes are truly involved in multiple processes, but it is clear that even if single functions were known for all yeast genes, we would still be far from a complete understanding of the complex network that supports most cellular processes. This further

underscores the importance of developing approaches for fast and accurate discovery of protein function.

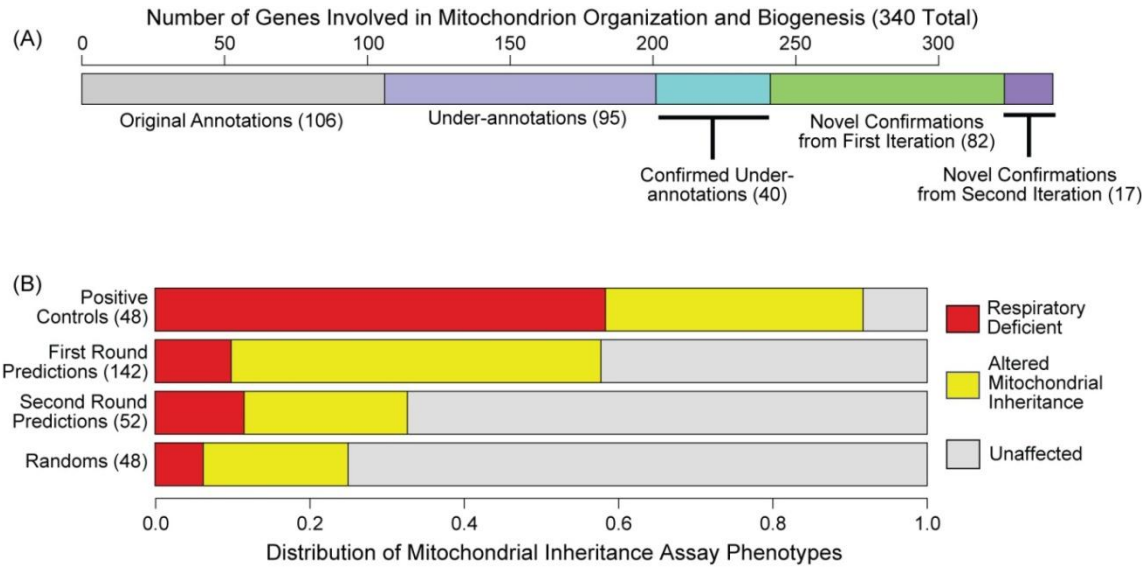


Figure 11: Annotations and phenotypic results for mitochondrion organization and biogenesis. Our study began with the 106 genes annotated to the GO term *mitochondrion organization and biogenesis*. In the first round of our iterative computational prediction and laboratory experimentation, we confirmed 122 additional genes. 40 of these confirmations had previously existing literature evidence for involvement in mitochondrial biogenesis, leaving 82 entirely novel discoveries from the first iteration. Based on further literature searches, we found an additional 95 genes with evidence for inclusion in this term (including 2 tested genes that did not exhibit a significant phenotype). During our second iteration of testing, we confirmed an additional 17 predictions. A) The number of genes involved in *mitochondrial organization and biogenesis* after each stage of this study. B) The results of our petite frequency assay for genes with previous literature evidence (positive controls), our novel first iteration predictions, novel second iteration predictions, and a random selection of genes. Note that the majority of novel confirmations exhibited the more modest phenotype of "altered mitochondrial inheritance," whereas the majority of previously known genes are "respiratory deficient," a more extreme phenotype more easily discovered by high-throughput screens.

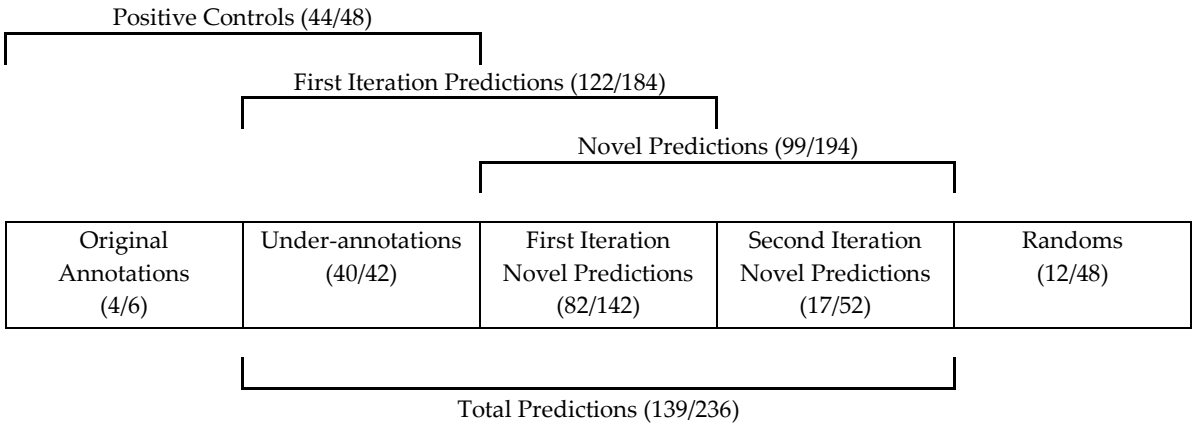


Table 3: Numbers of genes tested and verified in this study. This table shows a breakdown of the groups of genes tested in this study, with the numbers in parenthesis showing the number of verified genes over the number of tested genes. Initially, 184 first iteration prediction genes were selected from our computational ensemble for testing. We found existing literature evidence for involvement in mitochondrial biogenesis for 42 of these genes, and thus included these in the positive control set along with 6 genes that were originally annotated to the *mitochondrion organization and biogenesis* GO term. In our second iteration, we selected an additional 52 candidate genes, none of which had prior literature evidence for involvement. We also selected 48 genes at random from the genome for testing to establish the background genomic rate for our assay.

Guiding laboratory experiments with computation greatly increases discovery rates

Among our 236 experimentally evaluated computational predictions, 139 were verified, resulting in an overall true positive rate of 59%. This result is a striking confirmation that computational predictions can successfully direct laboratory experiments; nearly two out of three predictions were successfully confirmed, which would make even low-throughput follow-up experiments worth pursuing. To quantify our improvement in rate of discovery over the background rate of observing the same phenotypic classes, we chose 48 genes at random to establish baseline rates of phenotypes. Of these 48 genes, only 12 (25%) exhibited a phenotype consistent with involvement

in mitochondrial inheritance. Based on these results, the use of computational methods to guide our investigation increased our discovery rate by 236%.

In addition to a greatly increased discovery rate, we have evidence that our confirmed computational predictions are more integral to mitochondrial biogenesis than the rare positives resulting from our random screen. As mitochondria are vital for cellular respiration, our assays focused on discovering respiratory defects in single gene knockouts, which is a strong indicator that the tested gene plays a role in mitochondrial processes (Ogur and St John 1956; Ogur, St. John et al. 1957). However, it is possible for secondary effects of non-mitochondrial mutations to result in similar phenotypes. For example, one of the randomly selected genes tested, HTA1, is a histone whose deletion is known to cause pleiotropic effects on transcriptional regulation of carbon metabolism (Grunstein 1990). Consequently, our testing of an *hta1*Δ knockout strain resulted in a phenotype indicating involvement in mitochondrial organization and biogenesis, even though the true cause of this phenotype is likely a secondary effect due to a gross perturbation of carbon metabolism.

Given the possibility that secondary effects could occasionally manifest as positive phenotypes, we cross-referenced our results with known localization information from SGD. We would expect many of the genes involved in mitochondrial organization to localize either to the mitochondrion itself or to the actin cytoskeleton, as mitochondria associate with actin cables for proper inheritance of the organelle during cell division (Boldogh, Vojtov et al. 1998). Among phenotypically positive genes where localization data is available, 72% of our computational predictions are localized either to the mitochondrion or to actin, while only 36% of the 12 phenotypically positive genes from the random screen are similarly localized. The large

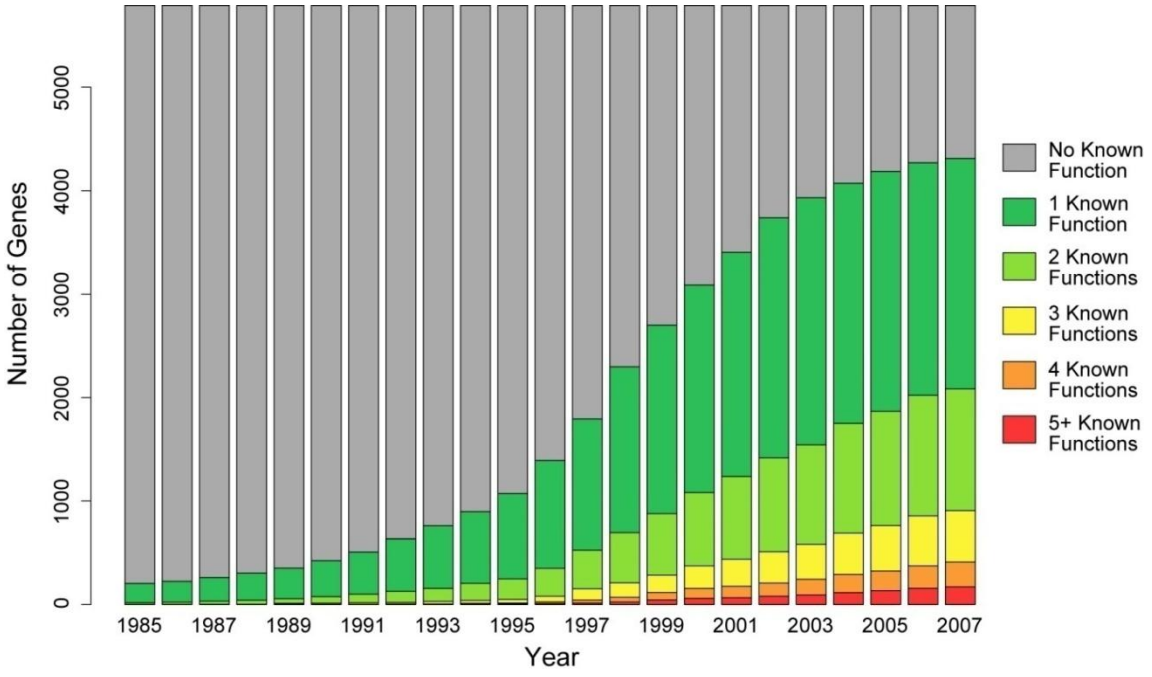


Figure 12: Historical progression of gene function discovery. We examined the historical context of SGD annotations to GO based on the dates of publications used to assign genes to biological processes. Here we define a "known function" as an annotation to a GO term within the GO functional slim mapping (Myers, Barrett et al. 2006) for *S. cerevisiae*. Function annotation accelerated after the publication of the yeast genome in 1996, but annotation of multiple functions did not accelerate accordingly.

discrepancy in localization among phenotypic positives from predictions and from the random screen indicates that positive mitochondrial phenotypes in some of the genes in the random screen may be due to secondary effects.

While enrichment for localization to the mitochondria is a strong indicator that our computational predictions are directly involved in mitochondrial maintenance, it is important to note that such localization is not a precondition for involvement. Among all of our novel tested computational predictions, 45% are known to localize to the mitochondrion or actin cytoskeleton, and of these, 59% were confirmed. However, our confirmation accuracy is also high (45%) among

the predictions not known to localize to these areas. Thus, if our study examined only genes known to localize to the mitochondrion, it would fail to discover nearly half of the verified genes that resulted from our use of computational predictions. Since computational data integration can leverage a variety of heterogeneous data sources in an unbiased manner, it can successfully direct experimental efforts to targets that might otherwise remain undiscovered.

Diverse, accurate predictions are made by different computational approaches

In addition to demonstrating the accuracy of computational function prediction approaches, our results also emphasize the importance of considering the specific biological nature of predictions. Specifically, our results show that different computational approaches can produce equally accurate, but distinct predictions depending on the algorithmic foundation and underlying data of each method. Although we did not attempt a comprehensive study of all types of computational function prediction methods, the three methods used in this study included both supervised and unsupervised approaches utilizing different data sources, and our observations are likely to be generally applicable. To demonstrate this generality, we have also analyzed additional canonical computational function prediction approaches (a Support Vector Machine (SVM) trained using only microarray data, an SVM trained using diverse data, and unsupervised correlation across microarray data). This additional analysis supports the results and conclusions presented below. Each of the three function prediction methods employed in this study achieved similarly high rates of phenotypic positives (Figure 13A). However, there was a relatively small overlap between the 40 most confident predictions of each method, as only 8% of the 88 total candidates selected from an individual method were common to all three (Figure 13B). True positive rates were similar among genes predicted confidently by only one method or by multiple methods, indicating that each computational approach was accurately predicting

disparate aspects of mitochondrial organization and biogenesis. This variation can be accounted for both by differences in the underlying data and by algorithmic diversity among the computational approaches. As discussed below, such differences among methods should be carefully considered when developing new prediction techniques or applying them in a biological setting.

Underlying data affects the specific biological nature of predictions

Of the three function prediction methods, two are based on detailed analyses of microarray data (MEFIT (Huttenhower, Hibbs et al. 2006) and SPELL (Hibbs, Hess et al. 2007)), while the third (bioPIXIE (Myers, Robson et al. 2005; Myers and Troyanskaya 2007)) focuses on integration of heterogeneous data sources such as affinity precipitation results, two-hybrid screens, sequence information, synthetic genetic interactions, etc. As stated, there was relatively little overlap between the three methods' predictions, although all three achieved similar true positive rates during laboratory validation. However, the microarray-based predictions from SPELL and MEFIT did show slightly more correlation with each other than with predictions from bioPIXIE (Figure 13B).

We characterized the importance of underlying data by examining the cross-validated results for the predictions of each method on more specific sub-processes of mitochondrial organization (Figure 14, see Methods for details). The microarray-based approaches (MEFIT and SPELL) clearly best capture information regarding *mitochondrial ribosome and translation*, which is consistent with other studies that have observed a strong ribosomal bias among microarray data (Myers, Barrett et al. 2006). The method based on diverse data (bioPIXIE) best captured information about *mitochondrial distribution* and *mitochondrial fission and fusion*. This is likely due

to the use of physical interaction data, which enables this method to more easily discover proteins involved in mitochondrial structure and motility.

Another significant difference occurs in the area of *mitochondrial respiratory complex assembly*, where the microarray-based methods are more successful than the method based on diverse data. Many of the proteins involved in this process are integral membrane proteins, making them technologically difficult to assay by common sources of physical interaction data (e.g. yeast two-hybrid, affinity precipitation). However, because the number of mitochondria in a cell (and thus the amount of membrane and membrane-bound complexes) depends on environmental conditions, these proteins can be strongly transcriptionally regulated when conditions change. This co-expression is captured by microarray data, providing evidence for our microarray-based predictors of functional relationships.

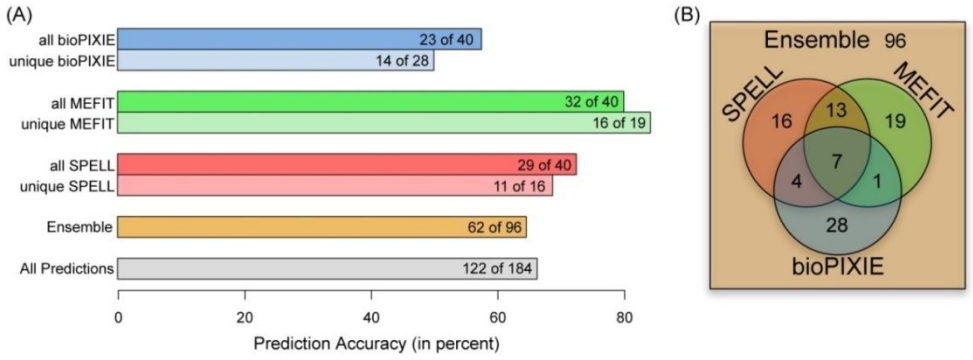


Figure 13: Individual method accuracy and overlap. Three computational methods and an ensemble of those methods were used to select candidates for experimental evaluation. Of the 184 predictions evaluated in our first iteration, 88 were chosen from the top 40 results of at least one individual method, while the remaining 96 were selected from the ensemble of all three. A) The accuracy of the predictions chosen from each method, from genes selected by the ensemble, and the overall accuracy for all candidates tested in our first iteration. B) Overlap between candidates selected from the individual methods. Each individual method performs with similar accuracy but predicts unique genes.

We have also examined the cellular localization of the predictions made by each of the computational methods and those made by the ensemble of all three (Figure 15A). While the majority of the predictions made by the microarray-based methods are known to localize to the mitochondrion, predictions from the method based on diverse data also contained a significant number of proteins known to localize to the actin cytoskeleton. This is consistent with the functional enrichments of the prediction methods, as mitochondria interact with actin for distribution, fission, and fusion. Interestingly, the verification rate was over 50% among genes localized to the mitochondrion and to actin. Precision was even higher (nearly 70%) among predicted genes with no known localization (Figure 15B). Additional analysis demonstrating the impact of underlying data, including training SVMs with different underlying data, produces similar results.

Algorithmic differences affect specific computational predictions

Even among methods based on the same underlying data, analyses by different computational approaches can produce very different function predictions. Only 20 of the top 40 predictions made by each of this study's two microarray-based methods (MEFIT and SPELL) overlapped (Figure 13B). However, each method achieved similarly high levels of biological accuracy (Figure 13A), and the functional and localization enrichments of the predictions made by these methods are similar (Figure 14 and Figure 15). These findings can be explained by the fact that these two methods employ very different analytical approaches when generating gene function predictions from microarray data.

One important difference is that MEFIT employs a supervised learning process, while SPELL is unsupervised. MEFIT relies on supervised Bayesian learning to up- or down-weight datasets, using prior knowledge of functional relationships. SPELL performs a query-driven similarity

search to identify significant patterns of expression within datasets that are determined to be informative for each query. When performing function prediction, MEFIT infers a complete functional interaction network, which is mined using "guilt by association" for genes predicted to be involved in mitochondrial organization. Conversely, SPELL averages a collection of searches, each querying an individual subset of known mitochondrial genes. As the underlying data collection and normalization procedures were the same for both methods, these algorithmic differences account for the diversity of specific predicted genes. This highlights the potential impact of specific algorithms, as well as underlying data, when predicting gene functions.

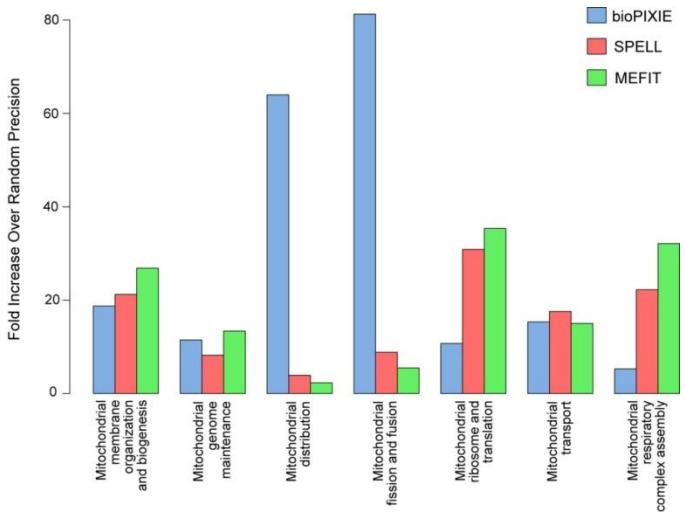


Figure 14: Biological differences between the three computational prediction methods. We evaluated which aspects of mitochondrial biology were targeted by each computational function prediction method. Even though all three methods learned and were evaluated using the same set of training genes, the methods differ in the sub-groups of mitochondrial biology on which they focused. SPELL and MEFIT are both based solely on gene expression microarray data, which explains their strong coverage of the mitochondrial ribosome and translation sub-group. bioPIXIE is based on diverse data, including physical binding data, which explains its strong coverage of sub-groups involving mitochondrial motility and physical interactions.

An ensemble of diverse prediction methods increases breadth of results

By employing multiple, complementary functional prediction techniques, we substantially expanded the breadth of our experimentally assayed genes. As described above, the three methods used in this study produced diverse, yet uniformly accurate, predictions spanning many biological aspects of mitochondrial organization and biogenesis. In addition to testing the top 40 predictions of each method individually, we also produced an ensemble prediction set by combining the results of each method based on estimated precision (see Methods for details). From this list, we selected 96 additional candidates for experimental validation.

Thus, approximately half of the novel predictions tested in this study did not occur among the top 40 predictions of any individual method, but were selected based on the ensemble of all three methods. The accuracy of these ensemble predictions is roughly the same (65%) as the predictions made by any of the individual methods (Figure 13). Similarly, the localization and functional enrichments of the ensemble predictions were distinct from those of any one prediction set (Figure 15). By harnessing the diversity and complementarity of our computational prediction methods, we were able to expand the biological scope of our investigation.

Iterative approaches converge on comprehensive prediction sets

To identify further promising mitochondria-related proteins, we performed a second prediction and validation iteration where confirmed predictions were fed back into the gold standard used in the computational prediction process. Initially, we selected 184 gene candidates to test, 122 of which were verified as likely involved in mitochondrial organization and biogenesis. In addition, we found that 40 of our verified candidates had strong existing support in the literature, which led us to identify 95 further genes with previously published literature evidence for inclusion in this process. After this first round of testing, we created a new training standard of 323 genes

including the original annotated genes, the genes with strong literature support, and the experimentally verified genes. Using this updated training set with our ensemble classifier, we selected an additional 52 novel testing candidates, 17 (33%) of which demonstrated a significant phenotype in the lab, resulting in our total of 139 gene function associations (99 entirely novel, 40 with previous literature support). Beyond simply providing additional genes verified to function in mitochondrial biogenesis, this iteration process led us to several important observations.

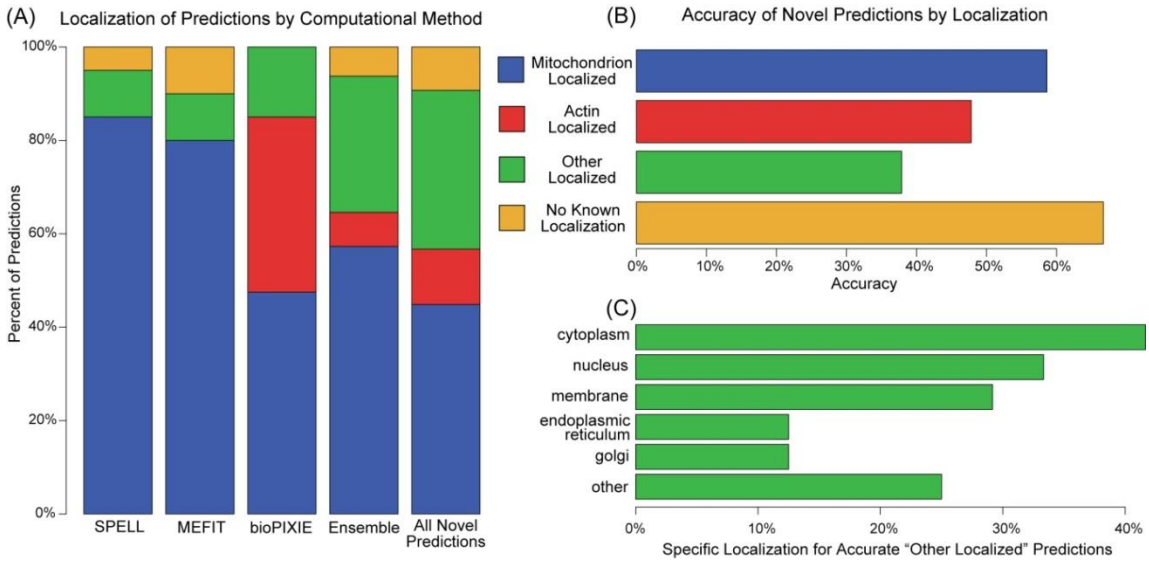


Figure 15: Localization of predictions from computational methods. The known localization of genes predicted by our computational methods differed greatly between the microarray based predictions (SPELL and MEFIT) and the predictions based on diverse data (bioPIXIE). A) Localization breakdown of the predictions made by each method, by the ensemble, and for all of our novel predictions. B) Accuracy of our novel predictions by localization. C) Breakdown of localization for those predictions in areas other than the mitochondrion or actin cytoskeleton. Accurate predictions are not confined to mitochondrially localized genes, suggesting that computation can discover more diverse gene functions than a screen based only on localization data.

While the predictions from our second iteration were verified at a rate higher than that of the random set, the discovery rate decreased relative to our first iteration. This suggests that we may be nearing the limit of predictions that can be verified using the single gene knockout assay employed in this study. Among our predictions that remain unconfirmed, some may still be involved in mitochondrial inheritance without exhibiting a significant phenotype using these assays. One of our top unconfirmed predictions, RMD9, was recently shown to synthetically interact with the known mitochondrial insertase, Oxa1 (Nouet, Bourens et al. 2007). Our remaining confident, but unconfirmed, predictions may thus be further characterized by double-knockout or over-expression studies.

Additionally, while the prediction methods differ with regard to which aspects of mitochondrial biology they best capture (Figure 14 and Figure 15), the methods begin to converge on similar predictions after just one round of re-training. Upon iteration, the correlation between the predictions of each method increased greatly (Figure 16). This convergence indicates that we have expanded our knowledge of this area to a level of biologically reasonable generality, since very different computational approaches can now arrive at similar conclusions. It also suggests that we have successfully avoided bias toward any one functional aspect of the mitochondria caused by over-reliance on individual methods.

These aspects of iterative learning - breadth and convergence - are especially important as the field moves to less well-studied areas of biology and to less well-understood organisms. Iterative applications of computational analysis and directed experimentation provide a means to refine the set of novel predictions and to increase the amount of information used for training. Even when beginning with relatively little information, this process can enable the accurate annotation of a significant number of novel participants in a biological process of interest.

Conclusion

In order to fulfill the broad promise of computational functional genomics, we must undertake large-scale, iterative efforts to predict, evaluate, and experimentally verify novel gene functions. Our study demonstrates the utility of these types of approaches, and we have made several observations potentially relevant to any computationally directed experimental setting. We find that both characterized and uncharacterized proteins can be fruitful candidates for laboratory investigation. Our results demonstrate that different computational methods can generate accurate but unique predictions, with characteristics dependent on both their underlying data and algorithmic basis. As such, utilizing an ensemble of diverse methods increased the biological breadth of our newly characterized genes. Further, the iterative use of an ensemble with rigorous laboratory experiments allowed us to confirm roles for additional genes and to converge on a refined prediction set.

An important aspect of this study discussed more thoroughly above is the enrichment of our novel discoveries for subtle phenotypes. Among the novel predictions examined in this study, subtly (but significantly) altered mitochondrial inheritance rates comprised 80% of the confirmed phenotypes; the remainder exhibited the more extreme respiratory deficient phenotype. Of the genes with prior literature evidence, only 36% exhibited altered inheritance rates, while the majority were respiratory deficient. Biologically, this is relevant in the study of the molecular mechanisms of human disease, since genetic disorders are often caused by mutations that only partially impair protein function (Perocchi, Mancera et al. 2008). From a computational perspective, it represents an opportunity to explore an untapped reservoir of novel biology. Many extreme phenotypes have already been discovered by high-throughput screens. Conversely, experimental assays sensitive and quantitative enough to detect these more subtle

phenotypes can be more difficult and time-consuming, and they can thus benefit greatly from computational direction.

Computational methods are critical in a field where the collection of functional genomics data is outpacing the characterization of novel biological knowledge from these experiments. While we used three specific computational approaches to study a particular biological process in yeast, our results demonstrate the broader applicability of combining functional prediction methods with experimental efforts. By directing laboratory investigations to more promising candidates, we can reduce the amount of time and effort required to discover new biology. This includes the characterization of multiple functions for individual proteins, an area still largely unexplored. Through the careful combination and iteration of computational and experimental biology, the rate and breadth of discovery can be enhanced in a variety of conditions, processes, and organisms.

Methods

A high level overview of our iterative prediction/experimentation/validation approach is shown in Figure 10. This section briefly details each of the steps involved in this process.

Computational prediction methodologies

We utilized three complementary computational gene function prediction methods in this study (bioPIXIE (Myers, Robson et al. 2005; Myers and Troyanskaya 2007), MEFIT (Huttenhower, Hibbs et al. 2006), and SPELL (Hibbs, Hess et al. 2007)). Each of the methods generated predictions of genes involved in the GO biological process *mitochondrial organization and biogenesis* (GO:0007005). All methods were initially trained and/or evaluated through cross-validation using the 106 annotations to this process as of April 15th, 2007. Full details of these methods can be found in

their respective publications. Here we present a brief summary of each approach and a description of how each method was used to produce computational function predictions.

bioPIXIE utilizes a suite of context-specific Bayesian networks to predict pairwise functional relationships between genes, which are then used to create fully-connected graphs weighted by confidence of functional interaction, $w(i, j)$:

$$w(i,j) = P\left(FR_{ij} \middle| D_{ij}^1, D_{ij}^2, \ldots, D_{ij}^k, B_{ij}\right) = \alpha P\left(FR_{ij} \middle| B_{ij}\right) \prod_{n=1}^{k} P\left(D_{ij}^n \middle| FR_{ij}, B_{ij}\right)$$

where FR_{ij} refers to the presence or absence of a functional relationship between proteins i and j, D^n_{ij} refers to the observed association in dataset n between the proteins i and j, B_{ij} is the biological context of the pair, and α is a normalization constant. This method integrates a wide variety of data sources, including physical interaction data (e.g. yeast two-hybrid, affinity precipitation, etc.), genetic interaction data (e.g. synthetic lethality, SLAM, etc.), gene expression data, and sequence data (e.g. coding and regulatory sequence similarity). The Bayesian classifier was trained within the biological process of interest, in this case using the genes annotated to *mitochondrial organization and biogenesis*. Predicted annotations to this term were derived from the resulting weighted interaction network by finding the significance of each gene's connectivity to known mitochondrial genes:

$$c_M = \left\{ \sum_{i \in M} \sum_{j \in G} w(i,j) \right\}, \quad c_G = \left\{ \sum_{i \in G} \sum_{j \in G} w(i,j) \right\}$$

$$c_i = -\log\left[1 - HG\left(\left\{ \sum_{j \in M} w(i,j) \right\}, \left\{ \sum_{j \in G} w(i,j) \right\}, c_M, c_G \right) \right]$$

where c_i is gene i's confidence of mitochondrial function, M is the set of genes known to be involved in mitochondrial organization, G is the set of all genes in the genome, $w(i, j)$ is the predicted probability of functional relationship between genes i and j, $HG(w, x, y, z)$ denotes the hypergeometric cumulative distribution function (CDF), and $\{x\}$ indicates that x is rounded to the nearest integer.

MEFIT also predicts pairwise functional relationships using a GO-trained naive Bayesian classifier; however, it is based entirely on gene expression data. Both MEFIT and SPELL (below) integrate roughly 2,400 microarray conditions that are grouped into ~120 datasets by publication and further subdivided by biological process examined. A ranked list of predictions was derived from the mitochondrial organization and biogenesis-specific network by calculating each gene's ratio of connectivity to known mitochondrial genes:

$$c_i = \frac{|G| \sum_{j \in M} w(i, j)}{|M| \sum_{j \in |G|} w(i, j)}$$

where c_i, M, G, and $w(i, j)$ are as above.

SPELL utilizes the same gene expression microarray data as MEFIT, but uses a query-driven search algorithm to identify novel players. While SPELL is not trained in a supervised fashion, it assigns a reliability weight to each dataset based on the co-regulation of a specified set of query genes and then orders the rest of the genome based on their weighted co-expression with the query set. SPELL generated predictions by using all possible subset pairs of known mitochondrial organization and biogenesis genes as queries ("leave two in" cross-validation), and then averaged these rank orders together to produce a final prediction list.

Each of these methods generated a ranked list of all genes in order of confidence of involvement in mitochondrial organization and biogenesis. We assigned an estimated precision level (EP) to each gene, *g*, in each list by calculating the fraction of genes with a higher confidence level that were already annotated to this GO term (disregarding genes with no biological process annotation or with annotations to the mitochondrial ribosome due to unusually strong expression co-regulation):

$$EP(g) = \frac{\#\ \text{of annotated proteins with rank}\ \leq\ rank_g}{rank_g}$$

We created an ensemble of the three methods by averaging these estimated precision levels for each gene. In this way, each prediction method contributed to the ensemble based on its reliability to recapitulate known biology. Further, this ensemble allows a gene with moderate confidence from multiple methods to rise in the overall rankings.

Identification of under-annotated genes

Our initial evaluation of the computational predictions led us to discover that 40 of our experimentally confirmed predictions were "under-annotated," meaning that they already had strong literature evidence for their involvement in *mitochondrial organization and biogenesis* but were not yet annotated to the corresponding GO term. In most of these cases, the information was already curated by SGD in the form of annotations to other GO terms, such as *integral to the mitochondrial membrane* or *mitochondrial protein import*. However, due to the structure of the GO hierarchy, these terms are not directly linked to our process of interest, *mitochondrial organization and biogenesis*. Beginning with these 40 genes, we identified an additional 95 genes that we believe have enough literature evidence to warrant their inclusion in this process without further

laboratory testing (including 2 genes tested that did not exhibit a significant phenotype). We have notified SGD of all 135 of these genes, and they are in the process of restructuring the GO hierarchy and making additional annotations. As of submission of this manuscript, SGD has already updated the annotations for more than half of these genes.

Selection of candidates for experimental testing

Novel candidates for laboratory evaluation were systematically chosen on the basis of both the three individual computational approaches as well as the ensemble of their predictions. As our experimental methodology (described below) is based on assessing phenotypes exhibited by single gene knockout mutants, we limited ourselves to consider only those genes with viable knockouts available in the heterozygous deletion collection. Additionally, we aimed to evaluate both genes with no previously known association to a biological process as well as genes known to be involved in an area other than mitochondrial organization and biogenesis. Thus, we divided the predictions into genes of entirely unknown function and genes with existing biological process annotations.

We selected the 20 most confident genes of unknown function and the 20 most confident genes with existing annotations from each of the three individual methods for testing. Due to overlaps between the methods, this resulted in the selection of 88 genes as novel candidates (the overlap between methods is shown in Figure 13B). We then chose an additional 96 genes from the ensemble list of predictions (38 from genes of unknown function and 58 from genes with known non-mitochondrial function) to arrive at our total of 184 test candidates in our first round of laboratory evaluation. In this way we could evaluate the performance of each individual method as well as the ensemble as a whole.

Of these predictions chosen for testing, we identified 42 as under-annotated, whereas the remaining 142 predictions have no previous literature evidence for involvement in mitochondrial maintenance. We selected 6 additional test candidates from the existing annotations to mitochondrial organization and biogenesis, resulting in a total of 48 genes with prior literature evidence for involvement in this process. We also chose 48 genes at random from the set of all viable single gene knockouts in order to establish baseline rates of phenotypic positives. It should also be noted that by chance we would expect some overlap between our random selection of genes and our novel candidates; in our case, 3 genes are in common between these two groups.

Experimental methodologies and evaluation of results

We utilized two experimental approaches to assess a gene's involvement in mitochondrial organization and biogenesis. Both of these methods quantitatively measure a single gene knockout phenotype in comparison to the same phenotype for matched wild type strains. Also, these methods were performed in replicate for each candidate examined such that robust statistical analysis could be performed on the results.

Strain preparation

For all of the genes examined, six independent isolates of complete ORF deletions were obtained from freshly sporulated strains from the yeast heterozygous deletion collection (Tong, Evangelista et al. 2001; Amberg, Burke et al. 2005).

Petite frequency assay

Yeast is able to grow and proliferate even without functional mitochondria on fermentable carbon sources. As such, yeast cells occasionally fail to pass aerobic respiration competent mitochondria on to daughter cells, but these cells can continue to proliferate. Cells lacking

functional mitochondria are called petite cells. In this assay, we assessed the rate at which single gene knockout strains produced petite offspring. A significantly altered petite formation frequency is indicative of a defect in mitochondrial biogenesis and inheritance (Ogur and St John 1956; Ogur, St. John et al. 1957).

For each mutant strain tested, we grew several replicates of the strain for 48 hours using glycerol as a carbon source. Strains severely deficient in their ability to maintain functional mitochondrial cannot grow on glycerol and were classified as respiration deficient in this first stage. Strains able to grow on glycerol were diluted and plated for single colonies on rich media (Amberg, Burke et al. 2005), which releases the requirement for functional mitochondria. Thus, as colonies formed, cells without functional mitochondria were generated. When the colony is fully formed, it is a mixture cells with functional mitochondria and cells without functional mitochondria. We measured this ratio by re-suspending a colony and plating a dilution of this re-suspension such that 100-300 colonies are formed on a plate. By overlaying with soft agar containing tetrazolium, cells with functional mitochondria were stained red, while cells without functional mitochondria remained white. The ratio of white cells to total cells gives the petite frequency. Eight independent petite frequencies were measured for each strain tested. The distribution of these frequencies was compared to the frequency of petite generation in wild-type yeast. Strains identified as having the altered mitochondrial inheritance phenotype in this assay exhibit at least a 20% change in petite frequency from wild type, and have a p-value of less than 0.05 when comparing the petite frequency distributions of that strain to the wild-type petite frequency distribution, using a Mann-Whitney U test.

Assessing the comparative accuracy of the computational methods

In order to compare which aspect of mitochondrial biology was best captured by each of the computational methods, we created a breakdown of known mitochondrial biology into several sub-groups. Based on the 106 original annotations and the literature evidence for the 135 under-annotations we created 7 more specific sub-groups of mitochondrial biogenesis genes shown in Figure 14. Given the prediction ordering of each computational method from our first iteration (i.e. using the original 106 genes as the training set) we calculated the average precision for each of the 7 more specific groups for each of the three computational approaches. The average precision was calculated for each sub-group, *G*, as

$$AP_G = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{i}{rank_i}$$

where $rank_i$ is the rank order of the ith gene appearing from the sub-group in the ordered prediction list. For display in Figure 13, the average precisions were normalized by the expected average precision if the genome were ordered randomly, which corresponds to the number of genes in each sub-group divided by the number of genes in the genome.

Iterative re-training, prediction, and verification

After our first round of testing, 122 of the 184 predictions were found to have a significant phenotype strongly indicating involvement in mitochondrial organization and biogenesis. Combined with the original 106 annotated genes and the 95 genes identified as under-annotated, this results in a total of 323 genes. Each of the three computational methods was re-applied using this updated training set of 323 genes and the same procedure was used to form an updated ensemble list of predictions. We selected 52 of the genes with the highest confidence from the

updated results that were not previously tested for a second round of laboratory investigation. The same experimental assays and evaluation procedures were used, and an additional 17 genes demonstrated a significant phenotype, resulting in a total of 139 out of 234 total predictions indicating involvement.

Performance Evaluation: Incomplete Knowledge Impedes Comparative Evaluations

Methods for gene function prediction and inference of biological networks have recently been of interest due to the growing availability of highly informative genomic data. Many different learning models have been applied to the problem, including kernel methods (Lanckriet, Deng et al. 2004; Barutcuoglu, Schapire et al. 2006), Bayesian networks (Jansen, Yu et al. 2003; Troyanskaya, Dolinski et al. 2003; Sachs, Perez et al. 2005), and graph-based approaches (Karaoz, Murali et al. 2004; Lee, Date et al. 2004). In these methods (and in this manuscript), "gene function prediction" is the task of associating genes with specific biological processes at the cellular level. Many of the methods used for this problem are similar to those used in predicting the biochemical function(s) of a gene (e.g. kinase activity), but here we focus specifically on the problem of predicting involvement in biological processes rather than molecular functions.

All of these approaches fall into the broad category of supervised machine learning classifiers. As such, each method requires trusted sets of examples from the classes it is learning about (e.g. a set of known DNA repair genes for learning about the response to DNA damage, etc.) These gold standard sets of genes are typically derived from repositories of gene annotations such as the Gene Ontology (Ashburner, Ball et al. 2000), KEGG (Kanehisa, Araki et al. 2008), or MIPS (Ruepp,

Zollner et al. 2004) databases. Given such a standard and a collection of training data, classifiers can be learned from the data using the algorithm of interest. The same gold standard is typically used for both learning (training) and assessing classifier performance (testing), usually through such techniques as hold-out testing, cross-validation, or jack-knifing (Mitchell 1997). New and improved function prediction algorithms are often then justified based on their performance relative to existing methods in such evaluations.

Clearly, these gene annotation databases play a central role in the successful application of machine learning techniques to gene function prediction. In fact, this is one of the major reasons why so many published methods have been developed and applied only in well-characterized model organisms. Gene annotations are generally considered to be more complete for such organisms, largely because the biological systems themselves are better understood and because of specific annotation efforts in model organism communities (e.g. SGD for yeast (Hong, Balakrishnan et al. 2008)). However, even in *S. cerevisiae*, one of the most extensively curated organisms, approximately 20% (~1,100 of ~5,800) genes have no annotations within the biological process Gene Ontology. Furthermore, the majority of the remaining genes (~60%) have only a single GO term annotation, which likely fails to capture the multiple cellular roles many genes are expected to play. The situation in higher eukaryotes such as mouse and human reflects an even greater degree of incompleteness.

The incomplete state of current gene annotations immediately raises at least two questions: how does this affect our ability to develop effective machine learning approaches, and how can we accurately estimate their performance when much of the ground truth is yet to be established? We address these issues with a comprehensive experimental validation of gene function predictions related to mitochondrion organization and biogenesis in *S. cerevisiae*. We employed

three previously published approaches for predicting gene function from large collections of microarray (Huttenhower, Hibbs et al. 2006; Hibbs, Hess et al. 2007) and other genomic data (Myers and Troyanskaya 2007). The most confident predictions from all three methods were tested, along with a collection of positive controls (genes known to play a role in mitochondrial function). In all, we tested 241 unique genes for association with mitochondrial function, and this experimentally confirmed set serves as the basis for answering fundamental questions about classifier performance. Coupled with our initial training data and GO-based standard, we provide this data as a benchmark for the experimentally validated evaluation of function prediction methods.

In our evaluation, we find that machine learning approaches can learn effectively even from current, limited functional annotations; our classification accuracy as confirmed through laboratory experiments is much higher than estimated for all three methods (an average of ~40% higher precision at 50% recall). However, we also observe substantial discrepancies in the estimated and actual relative performance of different prediction methods, even those based on exactly the same training data. These discrepancies have serious implications in comparative prediction evaluation, which we discuss below.

The organization of this paper is as follows: we first describe the details of our experimental validation, including a brief summary of prediction methods and the experimental assays used to test mitochondrial function. Second, we present a comparison of estimated classifier performance (based on cross-validation) with actual classification accuracy (based on experimental results). Finally, we conclude with a discussion of these results and their implications for the general task of predicting gene function.

Methods

To successfully combine computational gene function prediction with medium-throughput experimental validation, we employed a pipeline summarized in Figure 16. The system was bootstrapped by generating predictions from three computational methods (detailed below) that consume information from the Gene Ontology (Ashburner, Ball et al. 2000). These three methods generated ranked lists of genes to be assigned to the *mitochondrion organization and biogenesis* (*MOB*) term, which were then combined into a master list of testable predictions. The first evaluation was performed on these predictions using only information currently in GO.

Genes predicted to function in the mitochondrion were then further validated using medium-throughput laboratory experiments: assays covering several hundred genes over the course of 4-6 person-months with the accuracy of low-throughput techniques. In the case of *MOB*, this consisted of a petite frequency assay (detailed below) supplemented with semi-automated liquid growth rate measurements, both yielding statistically rigorous results. Genes verified in this manner to function in the mitochondrion were added to the GO-derived positive standard, augmenting the information available to the computational methods and allowing more accurate predictions to be generated. This allowed a second evaluation of our predictions to be performed incorporating the results of our laboratory experiments.

Taking advantage of these experimental results allowed the generation of new, more accurate lists of genes predicted to function in mitochondrial biogenesis. These lists were recombined, and genes newly predicted to have mitochondrial function were again experimentally validated. We found that the accuracy of both the individual prediction methods and of the combined predictions was greatly underestimated by the initial GO-derived standard. This implies that while GO provides enough knowledge to enable predictive machine learning, GO annotations

alone are insufficient (at least for the MOB term) to fully describe a biological process or to allow comparative method evaluation.

Computational predictions

The three systems employed to generate computational function predictions were bioPIXIE (Myers, Robson et al. 2005; Myers and Troyanskaya 2007), MEFIT (Huttenhower, Hibbs et al. 2006), and SPELL (Hibbs, Hess et al. 2007). The systems' implementation details are provided in their respective publications; in brief, bioPIXIE predicts pairwise functional relationships using a Bayesian framework consuming diverse genomic experimental data. This framework includes one Bayesian classifier per biological context of interest, where in this case, each context was an individual Gene Ontology term. A positive standard generated from GO was used to learn conditional probability tables specific to *MOB*. Predicted annotations to this term were derived from the resulting weighted interaction network by finding the significance of each gene's connectivity to known mitochondrial genes:

$$c_M = \left\{ \sum_{i \in M} \sum_{j \in G} w(i, j) \right\}, c_G = \left\{ \sum_{i \in G} \sum_{j \in G} w(i, j) \right\}$$

$$c_i = -\log HG\left(\left\{ \sum_{j \in M} w(i, j) \right\}, \left\{ \sum_{j \in G} w(i, j) \right\}, c_M, c_G \right)$$

where c_i is gene i's confidence of mitochondrial function, M is the set of 106 genes annotated to *MOB*, G is the genome, $w(i, j)$ is the predicted probability of functional relationship between genes i and j, $HG(w, x, y, z)$ denotes the hypergeometric probability distribution, and $\{x\}$ indicates that x is rounded to the nearest integer.

MEFIT also predicts pairwise functional relationships using a collection of GO-trained naive Bayesian classifiers. It consumes gene expression data drawn from ~2,500 microarray conditions drawn mainly from GEO (Barrett, Suzek et al. 2005), SMD (Demeter, Beauheim et al. 2007), and ArrayExpress (Parkinson, Kapushesky et al. 2007). A ranked list of mitochondrial function predictions was derived from the *MOB*-specific network by calculating each gene's ratio of connectivity to known mitochondrial genes:

$$c_i = \frac{|G| \sum_{j \in M} w(i, j)}{|M| \sum_{j \in G} w(i, j)}$$

where c_i, M, G, and $w(i, j)$ are as above.

SPELL is a query-driven system that also consumes these ~2,500 microarray conditions. When provided with a set of query genes, SPELL preprocesses each microarray dataset using SVD and weights them based on the correlations among the query genes in that data. Using these weights, the remainder of the genome is ranked by weighted average correlation with the query genes. To generate a set of predicted mitochondrial genes, the 106 genes annotated to *MOB* were used as a query. In all cases, these systems were initially trained and evaluated on the GO structure and annotations from April 15, 2007.

These three prediction methods were also used to produce an ensemble prediction set using estimated precision. A unified standard was formed by considering the 106 genes annotated to *MOB* to be positive examples, withholding the 80 genes of the mitochondrial ribosome (due to inordinately strong coexpression; see (Myers, Barrett et al. 2006)), and considering the remaining 4,824 annotated genes in the genome to be negative examples. This allowed the assignment of

standard precision/recall scores to each gene in each method's ranked list of predictions (see Figure 17). These precisions were thus comparable across methods (unlike the method-specific connectivities and weights), and the final combined prediction list was generated by ranking each gene by its average precision across the three methods.

Laboratory experiments

The primary assay used to validate our mitochondrial predictions was a measurement of petite colony frequency, supplemented with a measurement of growth rate in liquid medium. Detailed methods for these assays can be found above; in summary, we performed all assays on haploid deletion mutants drawn from the *S. cerevisiae* heterozygous deletion collection (Tong and Boone 2006). Knockout strains corresponding to genes with predicted mitochondrial function were drawn from the collection, sporulated, selected for haploids, and assayed as follows.

"Petite" yeast colonies form from yeast lacking functioning mitochondria (specifically, mitochondrial DNA). The mitochondrial genome is naturally somewhat unstable, and wild type *S. cerevisiae* forms petites in our assay with a base frequency of ~23%. To compare this base rate with that of each deletion mutant, we sporulated the heterozygous deletion collection, isolated six independent deletion mutants for each gene tested, and grew these strains in media requiring aerobic respiration. The resulting plates were stained with tetrazolium, turning respiring colonies red and leaving petite colonies white (Ogur, St. John et al. 1957). This allowed colony types to be counted manually; these counts were converted into percentages, which were then compared against wild type for significance using the Mann-Whitney U test.

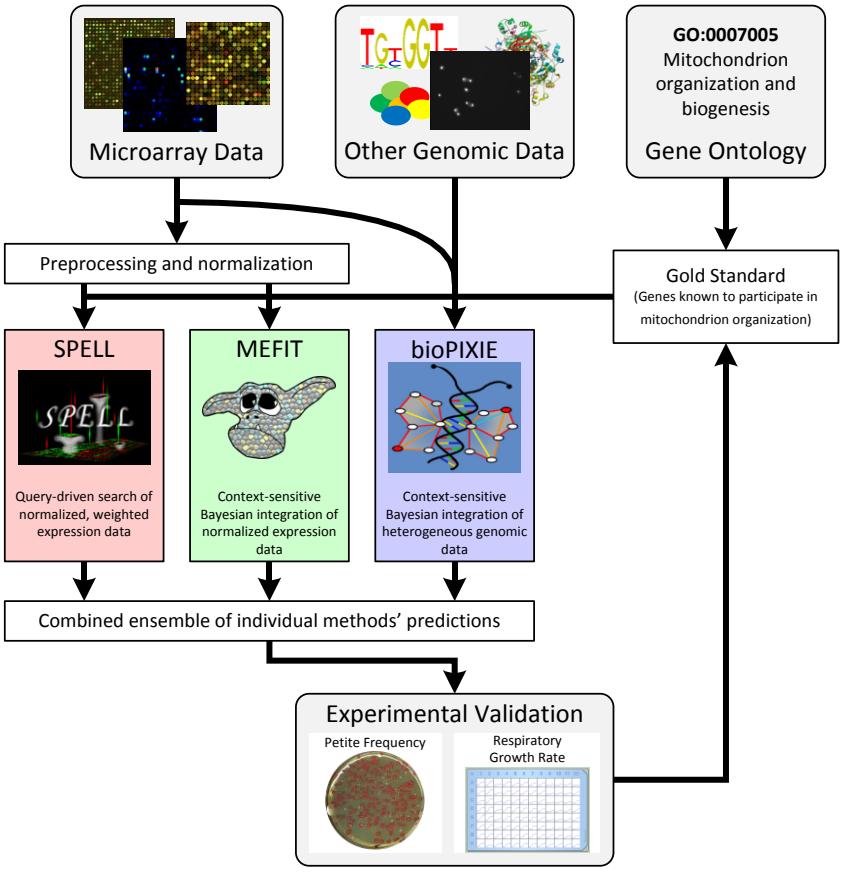


Figure 16: Overview of the system employed for computational function prediction and medium-throughput experimental validation. We used three computational data integration systems to predict *S. cerevisiae* genes functioning in the area of mitochondrion organization. An initial gold standard was generated from the Gene Ontology and used to train two of the machine learning systems: MEFIT, which integrates microarray data, and bioPIXIE, which integrates other diverse genomic data. SPELL was queried using mitochondrial genes from the same gold standard. Genes predicted to function in mitochondrion organization after training or as the result of queries were combined and used to select candidate genes for experimental validation. Genes that significantly perturbed mitochondrion organization when deleted were added to the gold standard, the three prediction methods were retrained, and a second round of experimental validation was performed. By augmenting the gold standard with experimentally validated and "under-annotated" genes, we increased the collection of mitochondrion organization and biogenesis genes by 220%.

Growth curves in liquid media (a measurement of optical colony density over time) were determined using a Tecan GENios plate reader and incubator to record colony densities in 96-well plates every 15 minutes over 42 hours. Each plate contained 12 mutants with six replicates each plus 24 wild type replicates. Growth rates were derived from these curves by using Matlab (MathWorks, Natick, MA) to fit an exponential model:

$$y = a + b2^{cx}$$

This model was fit over each whole curve, the first 2/3, of the first half, whichever yielded the best fit (to avoid plateau effects and to model only exponential growth). Wells with an adjusted $R^2<0.9$ were marked as non-growing, and growth rates for the remaining wells were determined by subtracting the row, column, and plate means for each well from the exponential parameter c. This yielded a rate c' for each well, and each knockout's c's were tested for significance against the wild type population using a Mann-Whitney U test.

To detect colonies growing exponentially but with significant differences in fitness, smoothed maximum densities d were calculated for all wells deemed exponential. Wells in which the maximum density was less than twice the minimum were marked as non-growing. From the remainder, plate, row, and column averages were subtracted from each well, generating adjusted maxima d'. Each mutant's d's were again compared with the wild type values using a Mann-Whitney U test. In both exponential growth and maximum saturation measurements, mutants with more than one outlier were deemed inconclusive and excluded from the results.

All defects specific to respiratory growth (i.e. significant in glycerol but not glucose) were considered. Mutants that failed to grow by both exponential growth and maximum density measurements were assigned a severe phenotype; mutants that failed to grow by one

measurement or were significantly defective in both were assigned a moderate phenotype. Mutants with a significant defect by only one measurement were assigned a weak phenotype, and all other mutants received no phenotype.

Validation methods and criteria

Each stage of our experimental validation relied on a combination of controls and replicates to ensure statistical rigor. Several categories of mutants were tested, beginning with independently isolated wild type control colonies. We chose positive controls for the various experimental assays from among the 106 genes annotated to *MOB*. Finally, three types of predictions were tested: under-annotated genes with literature support for mitochondrial function (but not annotated to *MOB*; these were treated as positive controls), known genes with some GO annotation outside of *MOB* (and no current literature support for mitochondrial function), and unknown genes with no current GO annotation. See above for a complete list of the 48 positive controls (six from *MOB*, 42 under-annotated), 76 knowns, and 117 unknowns tested in our assays.

The results of experimental assays were deemed significant enough to validate a gene's involvement in *MOB* only after passing stringent statistical requirements. In the case of the petite frequency assay, any mutant differing from the wild type controls with effect size >20% and $p<0.05$ was deemed to be verified to *MOB*. These genes were added to the augmented standard used for retraining and in Figure 17. The growth rate assay was used to explore more specific subprocesses of the general *MOB* term, e.g. respiratory growth as discussed below.

Results

It is striking that even in *S. cerevisiae*, one of the organisms most thoroughly annotated in current functional catalogs, publicly available experimental data provide a wealth of gene function information not captured by the GO mitochondrial organization and biogenesis term. Figure 17 contrasts the estimated accuracy of our three function prediction systems (and of the combined consensus predictions) before and after multiple rounds of experimental validation. Our initial predictions were generated using only preexisting experimental data and GO annotations; scoring these against GO (without holdout data) yields Figure 17A. Figure 17B evaluates the same predictions using an answer set augmented with the results of our first round of experimental validation. Figure 17C and D show the equivalent difference after the prediction methods are retrained on this augmented standard and after the standard is augmented again by a second round of experimental validation.

Of particular note is the difference in performance between Figure 17A's GO-based standard and Figure 17B's experimentally verified standard, also captured in the expected versus actual phenotype counts of Figure 18. Figure 17A and B's precision/recall curves are generated using the same set of computational predictions made using only existing high-throughput data and the Gene Ontology - but evaluation using GO alone vastly underestimates their accuracy. One may conclude from this that, at least in certain functional areas, functional catalogs such as GO currently possess sufficient depth to direct accurate machine learning in large datasets, but they do not have sufficient breadth to fully characterize novel predictions generated in this way from experimental data. We stress that this is no fault of GO in particular or of genomic curators in general; it is simply a product of the large amount of biology left to discover even in time-honored model organisms.

Comparative evaluation of predictions can be misleading

This variation in functional coverage within a gold standard, when combined with similar functional variations in prediction methods, can substantially misrepresent both global (Figure 17A and Figure 18) and function-specific prediction accuracy. As indicated by the experimentally validated standards of Figure 17B and D and the experimental phenotypes in Figure 18, our three prediction systems perform with roughly equivalent overall accuracies. However, there is sufficient diversity in the three prediction sets that they overlap quite differentially with the existing 106 GO *MOB* annotations. Prior to experimental validation, for example, bioPIXIE ranks genes such as ARP2 and ARP3 very highly; these are present in the original MOB term and thus raise bioPIXIE's precision in Figure 17A. However, genes such as YMR157C and YMR098C were ranked highly by MEFIT and SPELL but not initially annotated to *MOB*. Our experimental assays found that these genes do indeed function in the mitochondria, revealing in Figure 17B that all three prediction methods were performing quite well despite their initially low apparent precision.

This differential masking of performance by incomplete standards has clear implications in comparative evaluation of biological function predictions. Due to the highly complex nature of systems biology and the amount of knowledge still missing from even the best-curated functional catalogues, it becomes possible - even likely - to learn "real biology" that is not reflected in a gold standard and thus degrades, rather than improves, apparent performance. Conversely, it is equally possible to overfit a standard, improve performance on computational evaluations, and produce fewer experimentally verifiable predictions. It is thus essential that, until a greater understanding of the breadth of systems biology allows the construction of more complete

functional catalogues, computational predictions are validated using appropriately designed and scaled laboratory experiments.

Many validated predictions assign mitochondrial function to genes also annotated elsewhere

Of 142 mutants validated to MOB by our petite frequency assay, 99 were novel predictions with no prior literature support of mitochondrial function. 39 of these novel predictions, almost 40%, already possessed annotations to non-mitochondrial functions within the Gene Ontology. This underscores a biological detail not generally reflected in current functional catalogs: many genes participate in multiple biological processes. While GO and other functional catalogs are specifically designed to encode such characteristics, their significance and commonality has perhaps not been fully appreciated. This underrepresentation of functional plurality in current standards can, like the lack of coverage discussed above, obscure or bias comparative evaluation of gene function predictions.

Medium-throughput experiments validate predictions

The fact that these experimental validations were done in medium throughput is key to achieving both coverage and reliability in our augmented standards. In addition to the global evaluative power of our petite frequency assay demonstrated in Figure 17 and Figure 18, directed medium-throughput experiments such as the growth rate assay can indicate specific sub-functions for particular genes. For example, the myosin MYO3 and the functionally uncharacterized AIM8 both show much higher than expected petite frequencies (150%, $p<10^{-3}$ and 136%, $p<10^{-3}$). However, a *myo3*Δ mutant shows no significant growth defect in liquid media, while *aim8*Δ is significantly impaired (achieving neither exponential growth nor a single doubling of culture density). While the petite frequency assay alone could not differentiate these genes' activities within *MOB*, the more specific growth rate assay suggests that AIM8 may function specifically

within respiratory growth. Additionally, this *myo3*Δ phenotype is in interesting contrast to MYO5, which leaves petite frequency essentially unchanged when deleted (108%, p>0.2); to our knowledge, these two myosins have not previously been shown to act differentially in mitochondrial inheritance (Moseley and Goode 2006).

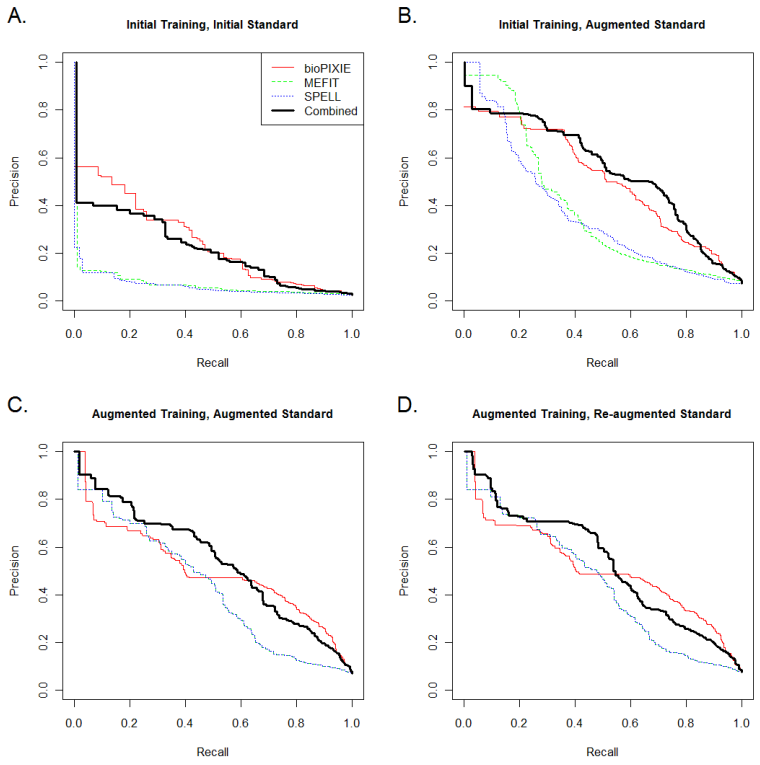


Figure 17: Prediction accuracy as estimated by prior knowledge, one round of laboratory validation, and a second iteration of experimental validation. A) Performance of three function prediction methods and their ensemble as evaluated by a GO-based gold standard. B) Accuracy of the same predictions as evaluated by a standard augmented with the results of one round of experimental validations (189 tests). C) New predictions (generated by the same three methods) evaluated using the augmented standard of B. D) Accuracy of these predictions evaluated using additional information from a second round of laboratory experiments (52 additional tests). Actual predictive accuracy as evaluated by experimental results is very different than would be expected from a GO-based evaluation.

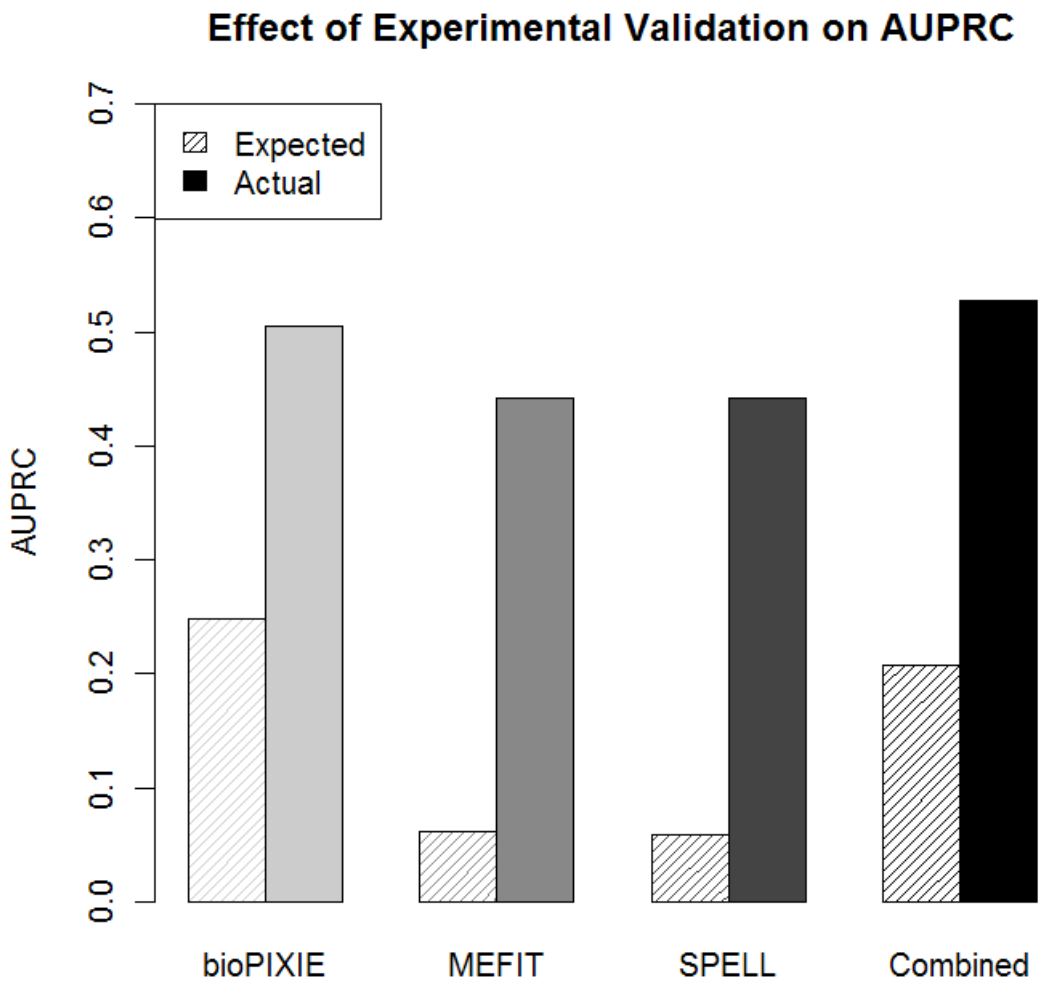


Figure 18: Comparison of phenotype frequencies expected from a computational gold standard versus experimentally validated frequency. Expected mitochondrial phenotype AUPRCs were calculated from Figure 17A using only the Gene Ontology; experimentally validated AUPRCs use the augmented standard of Figure 17D. Phenotype frequencies and accuracy of computational predictions are much higher in all cases than anticipated by preexisting functional catalogs.

Conclusions

We have described the results of a large-scale experimental validation of gene function predictions, focusing on the impact of an initially incomplete standard on machine learning behavior and evaluation. While we used mitochondrial organization in yeast as a model system, there are important global lessons we can derive from the results. We have demonstrated that, while a variety of machine learning methods can discover novel biology based on incomplete gold standards, a lack of functional coverage in these standards can seriously bias subsequent evaluations of these learning methods. This difficulty in evaluation emphasizes the importance of rigorous experimental validation of computational predictions.

One of the most striking observations throughout our validation process was the lack of existing annotations for yeast mitochondria. Our work began in April 2007, at which point there were 106 *S. cerevisiae* genes associated with the GO term mitochondrion organization and biogenesis. Manual examination of each method's top predictions revealed another 135 genes that had ample evidence in the literature for mitochondrial function but had no existing annotation to the MOB GO term. Of the 193 additional novel predictions we tested experimentally, we confirmed mitochondrial impairment phenotypes for 99 proteins (51%), bringing the total number of new *MOB* annotations to 339. Thus, we have effectively increased the number of annotations to this GO term by 220% with only a few months of computationally directed experiments and literature review.

We expect that the fraction of unannotated genes in other yeast processes is similar to that of *MOB*, with the exception of intensively studied processes such as RAS signaling or transcription. Mitochondrial processes are highly conserved across eukaryotes, and yeast mitochondria have been heavily used as a model system. Thus, annotated knowledge in this area should be

representative of general biological processes. This is not a shortcoming of annotation efforts, but simply an effect of the relative novelty of genome scale biology, particularly for organisms more complex than yeast. This incompleteness should be kept in mind as we use current functional catalogues as gold standards and develop new functional prediction techniques.

A second key observation about our results is that gene function prediction methods are much more reliable in this context than anticipated from a purely computational evaluation. For instance, using only the Gene Ontology, we estimated that 5-25% of the genes we tested would be true positives (the range of average precisions of the three individual methods). However, we confirmed mitochondrial phenotypes for 51%, an increase of two- to ten-fold over expected. These confirmed phenotypes include both genes of previously unknown function as well as genes with known involvement in other processes. Moreover, these confirmations are not just peripherally related to mitochondrial function, but some appear to play crucial roles in core mitochondrial activities (e.g. respiration or mitochondrial inheritance). These results indicate that our prediction methods were able to correctly find novel biological function for 99 proteins and to assign under-annotated function to 135 additional proteins.

Not only does this result speak well of the particular methods employed here, but more importantly, it demonstrates the promise of computational approaches to function prediction as a whole. While current gold standards remain incomplete, this does not necessarily impair the machine learning process; the quality and quantity of available genomic data are sufficient to convey rich functional information to existing methods. The limitation to such approaches on a large scale is instead the scope and availability of experimental follow-up and validation.

A less optimistic conclusion of this study is the difficulty of relative performance comparisons between function prediction methods using the currently available incomplete gold standards. We evaluated three different methods using the April 2007 GO annotations and compared this to a standard augmented with our experimental validations. We found that the methods' relative performance across these two evaluations was dramatically different, even for two methods based on identical training data. A comparative evaluation based solely on existing annotations was misleading due to incomplete knowledge.

Unfortunately, functional annotation repositories are one of the only sources of comparative evaluations of prediction methods short of more resource-intensive experimental validation; computational groups are often left with no other recourse when justifying publication of new methods. We certainly do not argue that such evaluations should not be done. However, our findings draw into question the field's ability to accurately resolve performance differences between competing approaches. This observation suggests that the application of existing gene function prediction methods in a laboratory setting can produce more tangible biological results than can the incremental refinement and development of new computational approaches.

When possible, integrative collaboration with experimental groups offers an initial solution to this problem. This is, of course, nontrivial; experimental studies based on computational predictions require special attention from computational groups and, obviously, substantial commitment from experimental labs. However, such collaborations can be highly rewarding from both computational and experimental perspectives: our study resulted in hundreds of novel candidates for detailed biological follow-up experiments, all of which were identified solely through computation.

Here, we have experimentally validated 99 computational predictions of novel mitochondrial gene function in yeast; in the process, we have demonstrated that the current incompleteness of gene annotation repositories does not necessarily impair computational function prediction, but it does hamper comparative performance evaluation of different techniques. Through the application of medium-throughput experimental validation, we rapidly expanded the annotation of the GO mitochondrial organization and biogenesis term by ~220%, and we have begun the process of as-signing more specific function to several of these genes. We anticipate that this combination of computational effort with rapid laboratory validation can be applied to a variety of other biological processes (e.g. DNA repair) to generate more complete, area-specific functional catalogues. These would in turn provide more accurate bases for the comparative evaluation of computational techniques, although this is still not a substitute for the depth, precision, and scientific potential of collaborative computational and laboratory investigation.

Efficiency and Effectiveness: Software for Biologists and Bioinformaticians

Since the earliest applications of specialized software to biological problems (Altschul, Gish et al. 1990; Eisen, Spellman et al. 1998; Ewing, Hillier et al. 1998), it has been necessary to finely balance a variety of computational and biological considerations. From the standpoint of computer science, much of the research import and challenge of the field lies in the algorithms, often from a theoretical perspective: sequence alignment can be solved by dynamic programming (Needleman and Wunsch 1970; Smith and Waterman 1981), DNA crossovers can be framed as a sorting problem (Gates and Papadimitriou 1979), and Hidden Markov Models can be used to identify genes within genomes and domains within proteins (Baum, Petrie et al. 1970). As in any field involving applied computation, the divide between algorithmic theory and practical software is always present, at times larger than others.

From a biological perspective, the challenges to be overcome are quite different: again, as in any applied field, the engineering issues of efficiency, usability, and reliability often come to the fore. Some concerns are specific to biology; for example, when investigating a particular disease or pathway, it is often more important to know a few key molecular participants than it is to obtain a complete list of every peripherally related gene. This in turn can place a computational emphasis on precision at the expense of recall. Likewise, in light of the tremendous complexity and sparse prior knowledge regarding many molecular systems, it can be desirable to retain, visualize, or learn from all data, even records that might seem too noisy or insignificant to represent useful examples. To quote from a seminal paper on practical computational analysis of microarrays by Eisen et al:

"It is not the purpose of this paper to survey the various methods available to cluster genes on the basis of their expression patterns, but rather to illustrate how such methods can be useful to biologists in the analysis of gene expression data. We aim to use these methods to organize, but not to alter, tables containing primary data; we have thus used methods that can be reduced, in the end, to a reordering of lists of genes... we need to develop our ability to 'see' the information in the massive tables of quantitative measurements that these approaches produce."

- Eisen, Spellman, Brown, and Botstein, PNAS 1998

In this spirit, this chapter presents four examples of computational tools developed to address specific biological questions. First, the Sleipnir library brings several of the most critical aspects of applied computer science to where they are most needed in modern biology: it provides a complete C++ framework for analyzing and integrating genome-scale data. This includes the manipulation of microarrays and many other high-throughput data types, the analysis of biological sequences, convenient and uniform access to functional catalogs such as GO (Ashburner, Ball et al. 2000) and KEGG (Kanehisa, Araki et al. 2008), and machine learning tools including Bayesian networks (Druzdzel 1999) and Support Vector Machines (Joachims 1999). Efficient manipulation and learning from large collections of genomic data can be difficult or impossible without proper computational tools, which this library attempts to provide to the biological and bioinformatic communities.

Second, we describe COALESCE, a comprehensive algorithm for regulatory network discovery from large data collections. COALESCE integrates gene expression measurements, genome sequence and transcription factor binding information, evolutionary conservation, nucleosome positioning, and any other appropriate data in a Bayesian framework to discover coregulated

biclusters, i.e. genes coexpressed under specific conditions, and the putative binding sites potentially responsible for their coregulation. Again, this process requires both careful algorithmic and software engineering development to perform rigorously and efficiently while also taking advantage of the depth and breadth of currently available data. Third, we discuss the Nearest Neighbor Networks (NNN) algorithm, designed to cluster microarray data with a specific biological goal of finding clusters enriched for functional similarity, as opposed to strictly tight coregulation. Finally, the Graphle tool provides a web-based interface for collaborative sharing and interactive exploration of large biological networks, ranging from protein interaction networks to predicted functional relationships to ontologies such as GO. All of these methods attempt to balance algorithmic novelty with biological applicability, while also bringing the best of both computational and biological worlds to their implementation.

We would like to thank K. Tsheko Mutungu and Sajid Mehmood for their tremendous efforts on the COALESCE and Graphle systems, respectively, as well as Hilary A. Coller for her always-insightful and helpful biological collaboration.

Sleipnir: A Software Library for Computational Functional Genomics

Whole-genome assays have now become pervasive, and the resulting wealth of data represents a new opportunity for biological discovery. A single genome-scale dataset can capture a snapshot of cellular function; integrative analysis of hundreds or thousands of genome-scale datasets can provide even more extensive systems-level insights regarding gene interactions under diverse conditions (Troyanskaya 2005). Integrated approaches have already resulted in important

biological discoveries (Myers and Troyanskaya 2007; Hong, Balakrishnan et al. 2008), and the breadth and depth of possible analyses will only increase as additional experimental data is collected.

As the amount of data to be analyzed continues to increase, computational efficiency becomes a greater concern. Specialized resources exist to enable very high-throughput computing for specific applications (Swindells, Rae et al. 2002; Pekurovsky, Shindyalov et al. 2004), but few computational options exist allowing researchers to quickly mine large collections of genome-scale datasets.

To address this need, we have created the Sleipnir library for computational functional genomics. The library contains algorithms and data types for efficiently manipulating and mining very large biological data collections. The core C++ library can be integrated into computational systems to provide rapid analysis of functional genomic data. Additionally, a variety of tools are provided that use the library to perform common tasks: microarray processing, Bayesian and Support Vector Machine (SVM) learning, and so forth. Even when analyzing individual datasets, Sleipnir often out-performs existing utilities in processing time, memory usage, or both (Table 4). Tools provided with Sleipnir address common data manipulation requirements, in many cases processing hundreds of datasets on a standard desktop computer. Additionally, the core Sleipnir library can be easily employed to efficiently apply new algorithms to complex biological data.

Implementation	Peak RAM (KB)	Time (s)
Bayesian learning (500 genes, 15 datasets)		
Sleipnir	1376	<1
GeNIe	6832	4
BNT	593180	15
Bayesian inference (500 genes, 15 datasets)		
Sleipnir	1216	1
BNT	273992	>600
Missing value estimation (10% missing, k=10)		
Sleipnir	27232	195
knnimpute	115708	368
Hierarchical clustering		
Sleipnir	83188	156
Cluster 3.0	176836	154
MeV	198292	361
K-means clustering (k=100)		
Sleipnir	8780	114
Cluster 3.0	28544	102
MeV	198292	361

Table 4: Memory usage and runtimes for Sleipnir and a number of other common tools for Bayesian analysis and biological data manipulation (Druzdzel 1999; Murphy 2001; Troyanskaya, Cantor et al. 2001; Saeed, Sharov et al. 2003; de Hoon, Imoto et al. 2004). All microarray operations were performed on the 300 conditions and 6,153 genes of (Hughes, Marton et al. 2000) using Euclidean distance. Bayesian operations were performed on simulated data using a binary gold standard and five randomly distributed values per dataset. Tests were run in a single thread on a 2GHz Intel Core 2 Duo. In every case, Sleipnir demonstrates a substantial advantage in speed, memory usage, or both.

Methods

The Sleipnir library contains a wide variety of tools for consuming standard biological data formats, manipulating and normalizing data, and performing machine learning and prediction. These are discussed extensively in the user and developer documentation included with the library (http://function.princeton.edu/sleipnir) and are presented here in summary.

Sleipnir provides C++ classes to parse pairwise interaction data and standard microarray file formats. Microarray data can be converted into pairwise similarity/distance scores using a variety of measures, discretized, normalized, randomized for bootstrapping or synthetic data production, split or merged, imputed, or clustered.

To facilitate functional enrichment analysis, gene function prediction, and gold standard generation from known gene functions and relationships, Sleipnir provides a uniform interface to several organism-independent function annotation catalogs. Information from organism-specific annotations can be merged with these functional annotations. Sleipnir also includes collections of data structures for dealing with common biological entities: gene identifiers, sets of genes, groups of related files, etc. Other utility classes include resources for multithreading, a ready-made network client/server class, and a variety of mathematical and statistical tools.

Sleipnir provides several tools for rapid machine learning and data mining. The SMILE Bayesian network library (Druzdzel 1999) and the SVM Light (Joachims 1999) library are used to learn and evaluate Bayesian or SVM models from very large collections of biological data. Arbitrary Bayesian structures are allowed, with parameters learned either discriminatively or generatively (Greiner and Zhou 2005) from data in a context-specific manner (Huttenhower, Hibbs et al. 2006);

extremely fast customized learning and evaluation implementations are used for naive structures.

Results

While Sleipnir's efficiency in integrating and mining biological datasets is most critical for very large data collections, it is also practical for single dataset tasks and smaller analyses (Table 4). When compared to several common tools for microarray manipulation or Bayesian learning, Sleipnir consistently demonstrates a substantial advantage in runtime, memory usage, or both. These improvements arise from a variety of optimizations but are broadly attributable to the flexibility allowed by C++ in manipulating large quantities of individual data (microarray values, interaction pairs, etc.) What Sleipnir trades off in generality (e.g. with respect to BNT) or robustness to malformed input (e,g. with respect to MeV), it gains in speed, memory management, and overall scalability, allowing it to efficiently manipulate large data collections.

The Sleipnir library is particularly useful for large integration tasks involving hundreds of diverse biological datasets; example applications of Sleipnir in such settings include (Huttenhower, Hibbs et al. 2006) and (Myers and Troyanskaya 2007). A schematic of such a task is shown in Figure 19, where Sleipnir was used to learn 200 context-specific Bayesian classifiers each integrating 186 *S. cerevisiae* datasets. Conditional probability tables were learned for each dataset within each context, entailing ~75,000 probability distributions. The resulting Bayesian classifiers were used to infer context-specific functional relationship networks, each consuming 90MB of disk space and calculated in 16.3 minutes. Sleipnir also supports an online mode for functional relationship inference in which no additional disk space is consumed and individual context-specific functional relationships can be produced in as little as 100ns. Parallelization on four processor cores reduces the total learning and evaluation time by an optimal 4-fold speedup

(~13h each for Bayesian learning and inference). Every stage of this complex data integration and machine learning task was performed using Sleipnir and its associated tools.

Discussion

The Sleipnir library for computational functional genomics provides a wide range of data processing and machine learning algorithms optimized for integrating very large collections of heterogeneous biological data. These include algorithms for data integration, machine learning by Bayesian networks or SVMs, and data types for manipulating microarrays, gene identifiers, functional annotations, and other common biological entities. Several tools are provided with the core library to perform common tasks, and most algorithms are multithreaded or parallelizable for distributed computing. The Sleipnir library enables computational biologists to efficiently integrate thousands of genomic datasets and to rapidly mine them for biological knowledge.

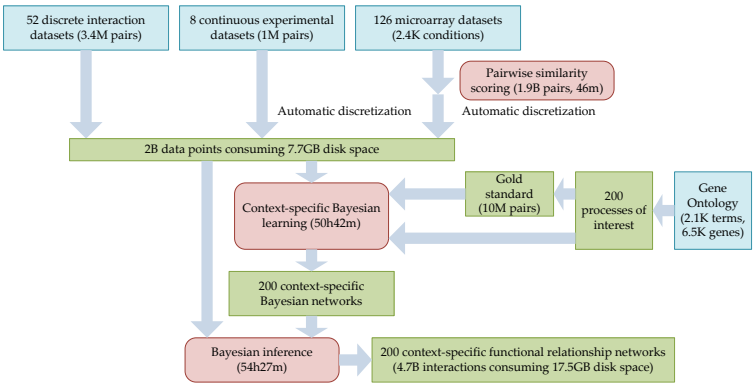


Figure 19: Sample application of the Sleipnir library to integrate 186 heterogeneous genomic datasets in *S. cerevisiae* within 200 biological contexts. Blue boxes indicate externally generated data, green boxes data generated by Sleipnir, arrows processing performed by Sleipnir, and red bubbles highlight time-consuming tasks. Times were generated on a 2GHz Intel Xeon CPU; peak RAM usage was ~200MB. Sleipnir is extensively parallelizable, and running these tasks on four cores reduces processing time by an optimal 4-fold to ~13h each for Bayesian learning and inference.

COALESCE: Data Integration for Biclustering and Regulatory Network Discovery

While the genome sequence of an organism describes its complement of potential proteins, it is the controlled expression, translation, and modification of these proteins that allows cells to survive and grow. At the level of transcription and mRNA stability, a complex regulatory network of transcription factors, RNA binding proteins, and microRNAs governs the interactions between components of a cell's internal state and its external environment. Understanding the elements of this regulatory network and the stimuli to which it responds in higher organisms has been of increasing recent interest (Kloster, Tang et al. 2005; Reiss, Baliga et al. 2006; Elemento, Slonim et al. 2007) as a key to metazoan systems biology, particularly as genetic misregulation is a major cause of human disease.

Here, we describe a Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction (COALESCE) allowing the discovery of regulatory motifs and modules from large collections of genomic data. COALESCE takes advantage of Bayesian integration of multiple data types on a large scale to predict coregulated gene modules, the conditions under which they are coregulated, and the consensus binding motifs responsible for their regulation. Through a novel synthesis of gene expression biclustering, motif prediction, and data integration (including expression, DNA sequence, nucleosome positioning, and evolutionary conservation), COALESCE can successfully find coregulated modules for organisms ranging from *E. coli* to human beings and from data collections as large as 15,000 experimental conditions.

We present the results of applying COALESCE to data from a wide range of organisms, including *H. sapiens*, *M. musculus*, *C. elegans*, *S. cerevisiae*, *H. pylori*, and *E. coli*. Using ~2,200 yeast

expression conditions, we recapitulate many known regulatory interactions (e.g. *AFT2* in iron transport, *STE12* activating mating genes) and highlight the importance of PUF family 3' UTR binding in a wide variety of targets, often ribosomal. In an analysis of ~15,000 human gene expression conditions, we extract a wide variety of putative upstream binding sites and potential 3' miRNA sites. On synthetic data comprising 5,000 genes and 100 conditions with 10 "activators" and "repressors" generated from a randomized model, COALESCE successfully recovered 60-90% of the affected genes, conditions, and binding motifs. In five sets of synthetic data containing no such regulators, COALESCE generated zero false positives. We are currently in the process of testing several novel transcriptional regulators of quiescence in human fibroblasts as predicted by COALESCE, as its ability to probabilistically leverage large collections of heterogeneous data is particularly suited to unraveling complex metazoan regulatory networks.

NNN: Nearest Neighbor Networks for Functionally Informative Clustering

The availability of DNA microarrays has made it possible to observe the transcript levels of every mRNA in an entire genome simultaneously. This has allowed researchers to monitor global changes in gene expression that occur in response to a cellular perturbation or the gene expression profiles characteristic of a particular state, such as a tissue type or a disease state. A major goal of integrative genomics is to interpret these gene expression patterns in order to define underlying signaling networks.

As the bulk of publicly available coexpression data has grown, a variety of successful techniques have been proposed for its analysis. In broad terms, these include normalization and meta-

analysis (Choi, Yu et al. 2003; Moreau, Aerts et al. 2003; Griffith, Pleasance et al. 2005; Hu, Greenwood et al. 2005), detection of differential expression (Ideker, Thorsson et al. 2000; Baggerly, Coombes et al. 2001; Cui and Churchill 2003), several forms of clustering (Eisen, Spellman et al. 1998; Heyer, Kruglyak et al. 1999; Cheng and Church 2000; Allison, Cui et al. 2006), and many others. However, each time a new microarray dataset is produced, it is ultimately in the hands of the generating biologist(s) to inspect the data and to determine what biological insights it might provide. This initial inspection is often aided by classical clustering algorithms such as K-means (MacQueen 1967; Tavazoie, Hughes et al. 1999) or hierarchical clustering (Sokal and Michener 1958; Eisen, Spellman et al. 1998), both of which are intended to present an intuitive, accessible view of genes whose coexpression might indicate similar regulation or biological functionality.

While these traditional algorithms can serve as a convenient first tool for microarray analysis, they can also be confounded by certain characteristics of biological data. K-means clustering, for example, requires prior knowledge of the number of clusters to find, and it will find that number of clusters even in random data (Dougherty, Barrera et al. 2002). Similarly, hierarchical clustering is incapable of leaving any genes unclustered, and its results can be driven by strong features in a small number of initially clustered genes (Quackenbush 2001). Many more recent clustering algorithms have been proposed to overcome these limitations, with Aerie (Gasch and Eisen 2002), CAST (Ben-Dor, Shamir et al. 1999), CLICK (Sharan, Maron-Katz et al. 2003), GenClust (Di Gesu, Giancarlo et al. 2005), Quality Threshold Clustering (QTC) (Heyer, Kruglyak et al. 1999), and SAMBA (Tanay, Sharan et al. 2004) representing a small cross-section of the tools available for the purpose of coexpression-based gene clustering.

These newer algorithms have overcome the drawbacks of traditional clustering in a number of ways. SAMBA, for example, represents a family of biclustering algorithms capable of excluding conditions as well as genes from a cluster; CLICK and QTC allow genes to remain unclustered, and Aerie and other fuzzy clustering algorithms permit genes to inhabit multiple clusters probabilistically. However, it is unclear how these algorithms perform with respect to their original purpose: providing biologists with a view of coexpressed biological processes within microarray datasets. Given a new dataset containing a collection of active biological pathways or functions, do these clustering algorithms accurately group functionally related genes?

We report below a clustering algorithm based on shared nearest neighbors called Nearest Neighbor Networks (NNN) intended to serve as a useful tool for biologists when discovering functional activity in coexpression datasets. NNN is unique in its focus on groups of genes sharing a mutual nearest neighborhood (based on some distance or similarity measure) rather than on groups of genes that are tightly correlated in some absolute measure. We present the results of a functional evaluation (Huttenhower, Hibbs et al. 2006; Huttenhower and Troyanskaya 2006) demonstrating NNN's ability to retrieve precise clusters that represent the diverse biological activity present in six qualitatively different microarray datasets. This evaluation also examines the behavior of the eight clustering algorithms discussed above to determine their accuracy in producing related gene clusters from many types of coexpression data and within many biological processes. Additionally, we compare the behavior of these clustering algorithms when presented with random data and when extracting clusters from integrated data (i.e. from a merged collection of all six microarray datasets). We believe that NNN represents an intuitive, simple tool providing biologists with a way to rapidly obtain and visualize a comprehensive collection of the processes coexpressed in a microarray dataset.

Implementation

NNN Algorithm

In designing a clustering algorithm that would allow us to make highly coherent clusters, we were inspired by the approach taken by Stuart and colleagues to define the homologues of a specific gene in multiple species (Stuart, Segal et al. 2003). In Stuart et al, a metagene was defined as a set of genes across multiple organisms whose protein sequences are one another's best reciprocal BLAST hits. We used a similar approach to find clusters of coexpressed genes by first identifying small cliques of genes in which each member is within the n nearest neighbors of each other. We then group together cliques that overlap to form larger clusters of genes.

NNN receives as input a set of genes of size m, a similarity measure $d(g_1, g_2)$ (such as Pearson correlation or Euclidean distance between the two genes' expression vectors), a clique size g, and a neighborhood size n. Its output is an assignment of each gene to zero or more clusters.

For each gene g_i, the n nearest neighbors $N(g_i) = \{g_{i,1}, ..., g_{i,n}\}$ are calculated based on the similarity measure d. If genes are considered to be vertices in a graph, this results in a directed graph in which each node is of out degree n (Figure 20A). An undirected graph is then constructed by connecting any two genes g_i and g_j such that $g_i \in N(g_j)$ and $g_j \in N(g_i)$, i.e. the two genes are mutual nearest neighbors (Figure 20B). All cliques (complete subgraphs) of size g within this graph are identified, and overlapping cliques are merged to produce preliminary networks representing potential clusters of related genes (Figure 20C).

A small number of genes in any genome often serve as interaction hubs connecting a large collection of minimally related partners (Tong, Lesage et al. 2004), and these genes can cause NNN to merge cliques to an undesirable extent. To address this issue, NNN uses a well-

established algorithm to remove cut-vertices in its preliminary networks (Tarjan 1972; Thulasiraman and Swamy 1992; Gross and Yellen 1999). A cut-vertex is a node whose removal results in an additional disconnected component in a graph; in our preliminary networks, such nodes represent genes connecting clusters that share no other interactions and are thus likely to be functionally irrelevant interaction hubs. Each of our preliminary networks is divided at its cut-vertices into multiple final networks, and the cut-vertices are included in each of the two networks which they induce (Figure 20D). Finally, to further ensure that cliques are not merged undesirably, any network (at most one) containing more than half of the input genes is removed.

NNN runtimes are generally below five minutes with reasonable parameter settings on a modern computer; with g = 5 and n = 25, the Hughes dataset (the largest used in our analysis) is fully clustered in approximately three minutes running in a single thread on a 2GHz Intel Core 2 Duo processor. Clustering with a worst-case g = 5 and n = 40 takes approximately 11.5 minutes, and the lower bound g = 3 and n = 10 runs in under 2.5 minutes. In the latter case, most of this time is spent calculating gene pair correlations.

Microarray Data Processing

To evaluate the abilities of NNN and other clustering algorithms to accurately cluster functionally related genes across a range of biological processes, we ran them on six *Saccharomyces cerevisiae* microarray datasets (Spellman, Sherlock et al. 1998; Gasch, Spellman et al. 2000; Hughes, Marton et al. 2000; Primig, Williams et al. 2000; Haugen, Kelley et al. 2004; Brem and Kruglyak 2005). The datasets range from seven to 300 conditions, include Agilent, Affymetrix, and custom cDNA arrays, include both time course and isolated measurements, and span a wide variety of biological perturbations and conditions.

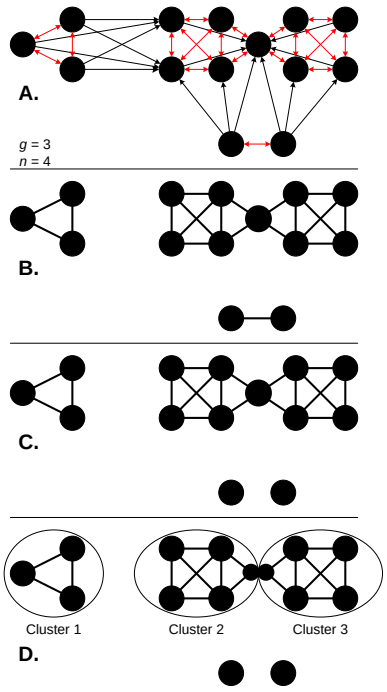


Figure 20: An example of the Nearest Neighbor Networks operating on 14 genes with clique size g = 3 and neighborhood size n = 4. A) A directed graph is generated in which each gene is connected to its n nearest neighbors. B) An undirected graph is constructed from bidirectional connections. C) Overlapping cliques of size g are merged to produce preliminary networks. D) Preliminary networks containing cut-vertices are split into final networks, with copies of the cut-vertices occupying both networks.

In all cases save Haugen et al (who provide data that has already been preprocessed), the datasets were filtered to remove genes with more than 50% missing data. Any remaining missing values were imputed using KNNImpute (Troyanskaya, Cantor et al. 2001) with k = 10, and replicated genes were averaged to ensure that each dataset contained at most one expression vector per open reading frame. For single channel data, expression values less than two were considered to be missing, and all single channel values were logarithmically transformed as a final preprocessing step. The two replicates in Brem et al were averaged together.

In order to construct a merged dataset consisting of conditions from all six individual microarray datasets, a data matrix was constructed containing each gene present in any of the datasets. Genes were assigned missing values for datasets in which they were not present. This merged data matrix was filtered to remove genes missing data for 50% or more of the resulting 664 conditions, and any remaining missing values were imputed using KNNImpute with $k = 10$. This left 6,160 genes, each represented by an expression vector of length 664 containing no missing values.

Random Data Generation

Randomized synthetic data was generated to characterize the behavior of NNN and other clustering algorithms when presented with data containing clusters present only by chance. Two sets of randomized data were generated, both containing 6,000 "genes" and 10 conditions. In the uniform case, each data value was drawn uniformly from the range [-1, 1]. In the normally distributed datasets, each value was drawn from $N(0, 1)$. Five datasets of each type were generated and used for the evaluations discussed below.

Evaluation Methods

In order to determine the accuracy and coverage of the functional relationships predicted by these clustering methods, we employed an evaluation method similar to that described in (Myers, Barrett et al. 2006). Specifically, we used the same 200 functions drawn from the Gene Ontology (Ashburner, Ball et al. 2000) as sets of "known" related genes; genes coannotated below these terms were considered to be functionally related. To generate negative examples, any gene pairs not coannotated below some GO term including at least 10% of the *S. cerevisiae* genome (roughly 645 genes) were considered to be unrelated. This resulted in an answer set of 620,854 related and 8,531,975 unrelated pairs.

Each clustering method was evaluated by considering any gene pair sharing a cluster to be related and any gene pair clustered separately to be unrelated; unclustered genes (when applicable) were neither related nor unrelated. This process transforms any clustering result into a set of related and unrelated gene pairs from which we calculated precision, recall, and/or area under a ROC curve (AUC) relative to the answer set. When performing per-biological function evaluations, these measures were calculated over subsets of the global answer set relevant to each function of interest; specifically, a gene pair was considered relevant to some function if i) it represented a positive relationship and both genes were included in the function or ii) it represented a negative relationship and one gene was included in the function (Huttenhower, Hibbs et al. 2006). All AUCs were calculated analytically using the Wilcoxon Rank Sum formula (Lehmann 1975).

Evaluation Parameters

Where possible, we evaluated each clustering algorithm over a range of parameters, e.g. K-means for values of k ranging from two to 30. By recording the most restrictive parameter setting at which each gene pair clustered together, we were able to generate full precision/recall curves for most clustering methods. In cases where this was not possible, a single clustering was generated per dataset, resulting in a point rather than a curve (but not affecting AUC calculations). All applicable clustering algorithms used Pearson correlation as a similarity measure.

Nearest Neighbor Networks was evaluated using our own Java implementation with the neighborhood size parameter n ranging from one to 30 in increments of three. The maximum neighborhood size used with the concatenated dataset for the per-function evaluation (Figure 23) was increased to 40 in order to provide coverage of a greater number of Gene Ontology terms. In all functional evaluations, the clique size g was fixed at five. The effects of varying g can be seen

in Supplemental Figure 1, with larger values slightly increasing precision while becoming more computationally expensive (Sipser 2005).

The K-means, CLICK, and SAMBA algorithms were evaluated using the implementation provided by the Expander tool (Sharan, Maron-Katz et al. 2003). For K-means, *k* was varied from two to 30 by increments of two. The CLICK and SAMBA algorithms were run with the default parameters provided by Expander, resulting in a single clustering. The predicted cluster confidences produced by SAMBA were used in lieu of a parameter setting to determine cluster specificity, with a higher confidence indicating a more specific cluster.

TIGR MeV (Saeed, Sharov et al. 2003) was used to execute the CAST algorithm, with the threshold parameter varied from 0.5 to 0.9 by increments of 0.05. Our own C++ implementation of Quality Threshold Clustering was used with a minimum cluster size of five and diameters ranging from 0.05 to 0.8 by increments of 0.05. QTC was unable to evaluate the concatenated dataset due to its reliance on the computationally intensive jackknife distance measure (Heyer, Kruglyak et al. 1999). Our own implementation of Pearson correlation was used as a representation of hierarchical clustering, with the raw pairwise correlation value itself behaving as a parameter over which precision and recall were calculated.

Implementations of GenClust and Aerie were provided by (Di Gesu, Giancarlo et al. 2005) and (Gasch and Eisen 2002), respectively. GenClust was run for 1,000 iterations with cluster counts *k* ranging from two to 30 by increments of two. GenClust failed to produce any output for the Hughes or concatenated datasets, apparently due to their high condition counts. Aerie was executed with *k* ranging from 10 to 40 by increments of two, as it failed to produce results for any *k* below 10. Aerie would not operate on the Primig dataset regardless of parameter settings, and

produced output for the Haugen dataset only for *k* up to 22. Since Aerie's *k* does not correspond to a final cluster count, each gene was assigned a vector of centroid distances corresponding to different initial *k*s, and gene pair similarities were calculated as correlations between these vectors.

Results and Discussion

As shown below, NNN succeeds in producing small, precise clusters from coexpression data, and these clusters generally span a wider variety of biological processes than those produced by the other clustering algorithms evaluated. While NNN's recall is lower than that of clustering algorithms in which all genes are always clustered, the capability to leave genes unclustered allows NNN to present an analyst with results consisting of only the high precision results of biological interest. This is evidenced, for example, in NNN's behavior when run on random data, which is left unclustered (Table 5). Furthermore, Figure 23 demonstrates the functional diversity of the clusters obtained from NNN; particularly on larger datasets, NNN detects activity in processes such as *conjugation* and *phosphorus metabolism* not captured by other clustering algorithms.

Nearest Neighbor Networks

NNN is intended to be an accessible and convenient tool for rapidly producing functionally coherent clusters from coexpression data, and visualization is therefore an important aspect of its results. Figure 21 demonstrates a sample of the default NNN output format as visualized by Java TreeView (Saldanha 2004). Here, each colored subtree represents a cluster found by NNN; these have been internally hierarchically clustered using standard correlation and average linkage for visual coherence, and the clusters centroids have in turn been clustered to produce a full tree.

Our NNN implementation also provides a tabular output format assigning genes to numbered clusters for further computational processing.

Global Evaluation of Clustering Algorithms

A global evaluation of NNN and eight other clustering algorithms on each of the six microarray datasets appears in Figure 22. As recommended in (Myers, Barrett et al. 2006), we have excluded the Gene Ontology term *ribosome biogenesis and assembly* during these evaluations so as not to bias the outcome towards this function. Myers et al discusses the problems raised in coexpression analysis by ribosomal genes, in particular their tendency to correlate so strongly even across conditions unrelated to ribosomal functions that they can obscure other biological activity. Especially in datasets eliciting strong stress responses (e.g. Figure 22B), this has a substantial impact on many of the clustering methods, accounting for a portion of their low performance and indicating that they may be clustering more easily discovered ribosomal genes at the expense of genes coexpressed for other biological reasons.

Although no one clustering algorithm is appropriate for every situation, Nearest Neighbor Networks demonstrates a clear advantage in precision in many of these datasets. In particular, the Gasch, Haugen, and Spellman datasets are perhaps best analyzed by NNN, demonstrating a robustness to functional bias (Myers, Barrett et al. 2006), low condition count, and periodicity, respectively. NNN performs approximately equivalently to QTC and Pearson correlation on the Brem dataset, and the Aerie, CAST, and SAMBA algorithms fall slightly beneath these due mainly to precision issues at low recall. CLICK is difficult to evaluate in this context due to its insensitivity to homogeneity parameter changes, leaving no way to trade off between precision and sensitivity. Thus, in a variety of contexts, NNN is best able to extract functionally relevant clusters from coexpression data with high precision.

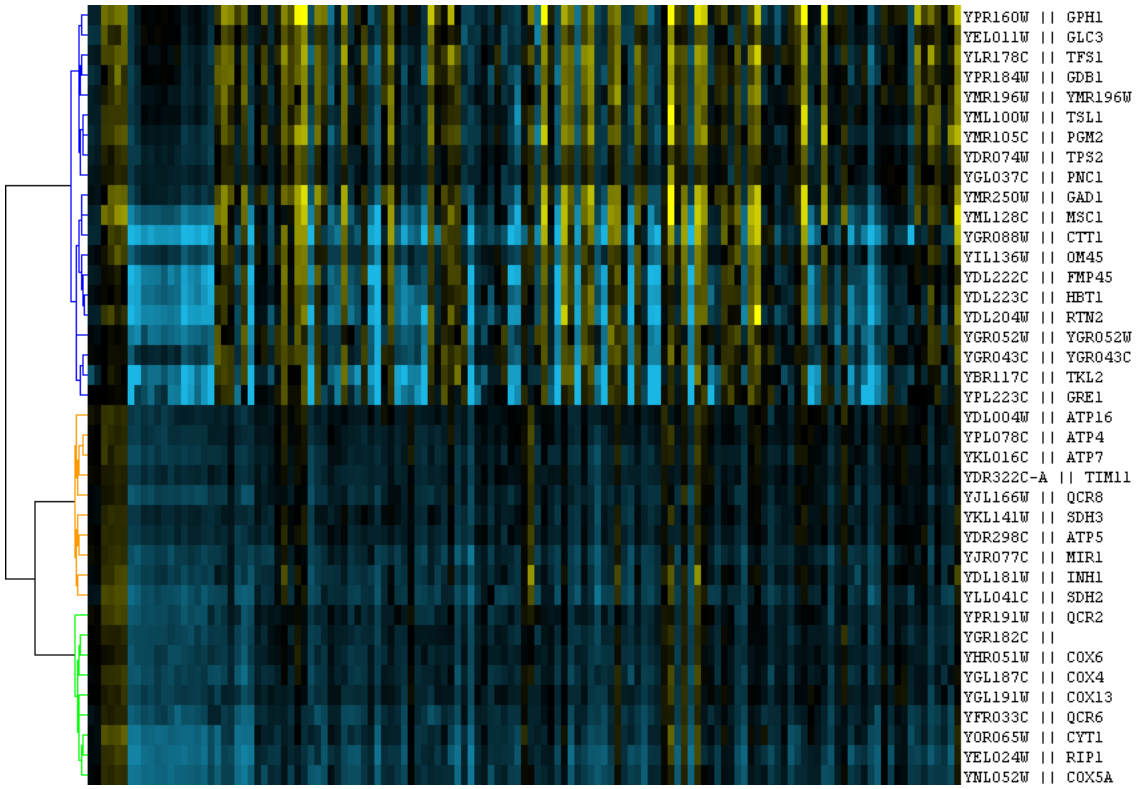


Figure 21: A subset of the Nearest Neighbor Networks clusters produced from the (Brem and Kruglyak 2005) dataset using the parameters g = 5 and n = 10, visualized using Java TreeView (Saldanha 2004). NNN clusters have been colored, internally hierarchically clustered, and the cluster centroids have in turn been hierarchically clustered to provide an easily interpretable tree.

NNN falls slightly short of QTC and, to a lesser extent, Pearson correlation in the Primig dataset, and QTC and SAMBA are both strong performers on the Hughes data. This latter effect might be attributable to the unordered nature of the Hughes data (a deletion study rather than a time course) from which SAMBA is able to bicluster correlated conditions as well as genes, and the large condition count likely benefits both SAMBA and QTC. NNN's performance in the high precision/low recall region of the Primig dataset is impaired by the fact that the Gene Ontology annotates MATALPHA1 and HMLALPHA under the *development* term, STE14 under the *protein processing* term, and STE3 and MF(ALPHA)1 under the *reproduction* term. This results in our

answer set considering their pairwise combinations (e.g. MATALPHA1 with STE14, STE14 with STE3, and so forth) to be unrelated, while NNN predicts them to be tightly clustered together.

While NNN is never more than slightly below the best performing algorithms, certain specific issues with other methods become apparent from this type of functional analysis. For example, SAMBA has some difficulty with the extremely small Haugen dataset (Figure 22C) and the periodic Spellman cell cycle data (Figure 22F).

Table 5 provides summary statistics describing the output of NNN using default parameters of g = 5 and n = 25 on the six datasets evaluated more fully below, on the concatenation of those six datasets, and on random synthetic data. For purposes of comparison, similar statistics have been provided from other clustering algorithms (where applicable) using their default parameter settings. NNN, QTC, and SAMBA are capable of leaving genes unclustered, and NNN and SAMBA succeed in taking advantage of this to recognize and ignore random data. With default parameter settings, NNN tends to be conservative, generally producing fewer, smaller, and (as evaluated above) more precise clusters than SAMBA.

Note that the default parameters may not be appropriate for all analyses; they are used here for comparison purposes. For example, more clusters can be obtained from the Haugen or concatenated datasets (if desired) by increasing n. The global evaluation above and functional evaluations below cover a wide range of parameter settings for all clustering methods and show results largely independent of specific parameter values.

Behavior on Random Data

It is of interest to note that only Nearest Neighbor Networks and SAMBA succeed in excluding randomized data from their clustering output. SAMBA achieves this by computing the statistical

significance of bicluster weights and retaining only those unlikely to occur by chance (Tanay, Sharan et al. 2002). NNN instead takes advantage of the fact that random data of this form tends to over-cluster, i.e. for an appropriate neighborhood size, all or nearly all genes cluster together. Since substantially overlarge clusters are eliminated by NNN, this results in the removal of randomized data from the functional clusters provided to the user.

Behavior on Concatenated Data

Only NNN and Pearson correlation succeed in extracting functional relationships from the concatenated datasets, with NNN achieving somewhat better recall. As discussed in (Huttenhower, Hibbs et al. 2006), algorithms relying solely on correlation measured over a long expression vector can be easily misled. This can be caused by differences in normalization between the datasets making up the concatenated vector or by overriding "global" signals providing high correlation among only a small set of ubiquitously coexpressed genes (e.g. the ribosomal genes discussed above). This has the effect of producing a small number of very highly correlated genes and relegating most of the correlations of functional interest to near-background levels. NNN avoids this problem by regarding both tight and diffuse clusters as equally valid, so long as cliques of mutual nearest neighbors are present.

For example, consider a group of ribosomal proteins coexpressed across all conditions with a mutual correlation of 0.9. A group of meiotic genes only activated under specific circumstances might achieve a correlation of 0.3 when tested across many conditions, since they will not usually be coregulated. If functionally unrelated genes tend to correlate at a level of 0.2, the ribosomal cluster will be far easier to discover. However, NNN will not distinguish between absolute correlation levels so long as the genes in each group are within each others' nearest neighborhoods - which will likely be the case, since their mutual correlations remain above

background. Meta-analytic normalization techniques provide another solution to this problem; correlations combined by z-scoring substantially outperform raw correlations, and these z-scores are in turn outperformed by NNN clustering using z-scores in place of Pearson correlation as input (data not shown).

Functional Evaluation of Clustering Algorithms

A global evaluation such as the one described above does not reveal the functional diversity of the predicted interactions; even with ribosomal interactions removed, it is possible for an algorithm to perform well by accurately predicting only a few biological processes. A complementary functional evaluation demonstrates that Nearest Neighbor Networks not only performs approximately as well or better than other clustering methods in global evaluations, it produces clusters that capture a wider array of biological functions. The heat map in Figure 23 indicates AUC scores for a variety of Gene Ontology terms within each dataset. NNN succeeds in accurately predicting clusters for several terms poorly analyzed by other algorithms, particularly within the Brem and Gasch datasets.

The high predictive power of Nearest Neighbor Networks in the Brem dataset likely reflects the unique nature of these microarray conditions. This dataset includes gene expression profiles from the segregants of a cross between two different strains of yeast. As opposed to most datasets, in which haploid yeast of one mating type are profiled, segregants with both the MATA and the MATALPHA phenotypes were present in the Brem data, making it possible to identify other genes correlated with mating type. In addition, there is a polymorphism between the parental strains in the pheromone response G protein GPA1, which is expected to result in differences in expression of effector genes among the segregants. Further, an interaction between the mating-type locus MAT and the pheromone response gene GPA1 has been detected (Brem and Kruglyak

2005). The expression profiles of genes in the *response to pheromone*, *sexual reproduction*, and *conjugation* functions are consequently related in this dataset and provide an opportunity for identifying high precision networks of genes with these Gene Ontology annotations.

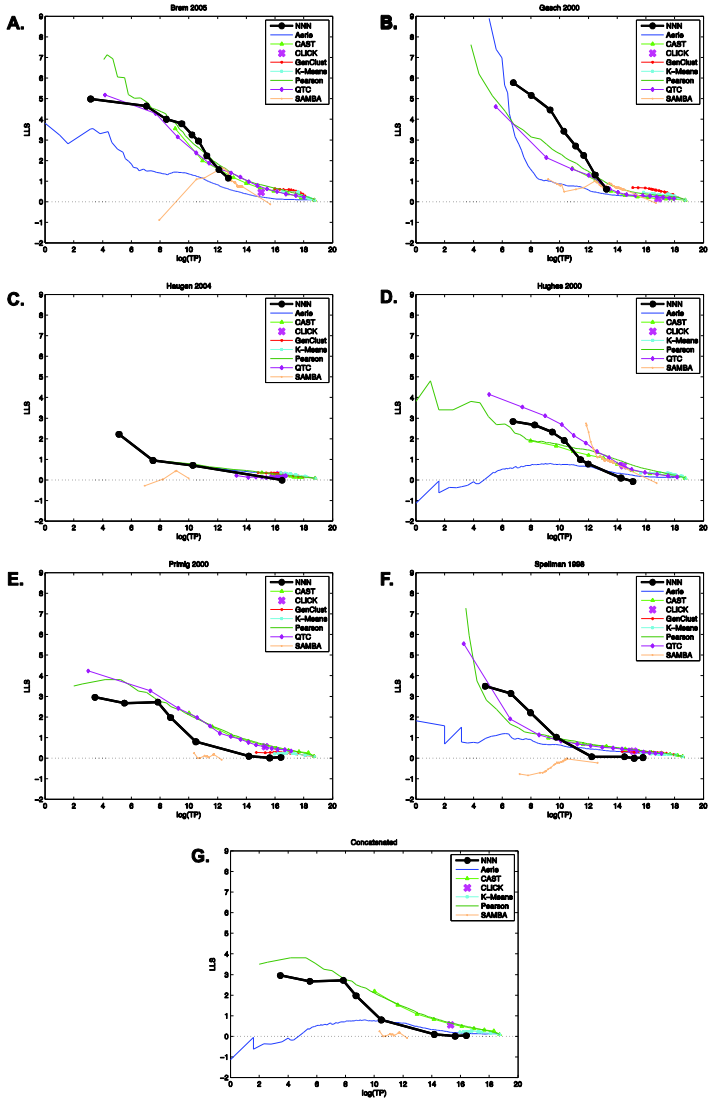


Figure 22: Evaluation results for eight clustering algorithms and six microarray datasets based on the global answer set (employing 200 GO terms of functional interest and discarding *ribosome biogenesis and assembly* (Myers, Barrett et al. 2006)). Performance has been measured using $\log_2(TP)$ on the horizontal axis and log-likelihood score $LLS=\log_2((TP/FP)/(P/N))$ for P total positive pairs, N total negative pairs, and TP and FP the number of true and false positives at a particular recall threshold. A) Brem 2005. B) Gasch 2000. C) Haugen 2004. D) Hughes 2000. E) Primig 2000. F) Spellman 1998. G) All six datasets concatenated.

	NNN g=5, n=25	CAST t=0.8	CLICK h=μτ	QTC d=0.5, n=5	SAMBA
			Brem 2005, 6162 genes, 131 conditions		
Genes	1527	6162	6162	6137	2284
Clusters	54	3552	82	127	113
Mean Size	28.4	1.73	75.1	48.3	102
Size Dev.	49.2	8.14	161	93.3	70.3
			Gasch 2000, 6115 genes, 173 conditions		
Genes	1142	6115	6115	6092	3120
Clusters	38	2702	9	69	128
Mean Size	30.1	2.26	679	88.3	130
Size Dev.	62.5	17.8	787	220	101
			Haugen 2004, 6256 genes, 7 conditions		
Genes	64	6256	6256	6236	280
Clusters	11	50	16	56	5
Mean Size	5.82	125	391	11.4	88.4
Size Dev.	1.19	332	474	258	36.5
			Hughes 2000, 6153 genes, 300 conditions		
Genes	1996	6153	6153	6121	3375
Clusters	29	4093	75	177	325
Mean Size	68.9	1.50	82.0	34.6	45.9
Size Dev.	245.4	4.46	107	57.8	44.1
			Primig 2000, 6005 genes, 24 conditions		
Genes	2247	6005	6005	5970	778
Clusters	27	872	46	110	25
Mean Size	83.2	6.89	131	54.3	139
Size Dev.	390	17.4	187	80.4	96.3
			Spellman 1998, 5701 genes, 25 conditions		
Genes	2050	5701	5701	5669	777
Clusters	28	782	47	100	32
Mean Size	73.3	7.29	121	56.7	69.0
Size Dev.	324	26.9	206	114	37.3
			Concatenated Data, 6160 genes, 660 conditions		
Genes	694	6160	6160	-	4892
Clusters	29	12	5	-	609
Mean Size	23.9	513	1232	-	63.7
Size Dev.	34.7	1691	1768	-	82.0
			Uniformly Distributed Random Data, 6000 genes, 10 conditions		
Genes	0 (±0)	6000 (±0)	3600 (±3286)	5964 (±28.8)	0 (±0)
Clusters	0 (±0)	228 (±3.05)	9.8 (±9.81)	109 (±4.72)	0 (±0)
Mean Size	0 (±0)	26.3 (±0.353)	190 (±175)	53.0 (±1.39)	0 (±0)
Size Dev.	0 (±0)	22.1 (±0.225)	48.8 (±45.7)	35.2 (±0.791)	0 (±0)
			Normally Distributed Random Data, 6000 genes, 10 conditions		
Genes	0 (±0)	6000 (±0)	6000 (±0)	5975 (±4.77)	0 (±0)
Clusters	0 (±0)	246 (±3.74)	28.8 (±11.9)	124 (±1.30)	0 (±0)
Mean Size	0 (±0)	24.4 (±0.371)	235 (±82.6)	48.3 (±0.482)	0 (±0)
Size Dev.	0 (±0)	18.5 (±0.0860)	64.8 (±46.3)	30.9 (±0.374)	0 (±0)

Table 5: Summary statistics detailing Nearest Neighbor Networks clusters formed from the datasets employed in this study, from their concatenation, and from two synthetic random datasets using default parameters (g = 5, n = 25). Results from other clustering algorithms with appropriate output formats (CAST, CLICK, QTC, and SAMBA) have been included, also utilizing default parameter settings provided by the algorithms' implementations. Random values are shown with standard deviations over five different seeds.

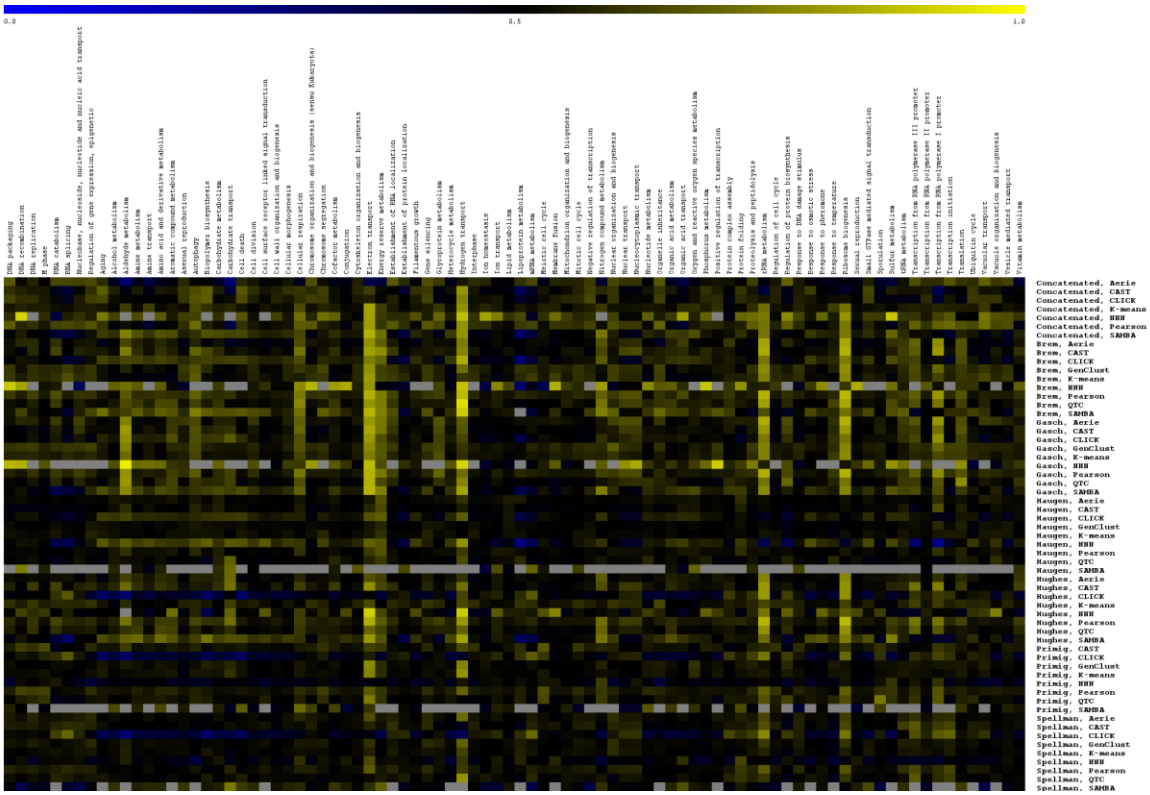


Figure 23: Function-specific evaluation results for each clustering method on a per dataset and GO term basis. Each cell represents an AUC score calculated analytically using the Wilcoxon Rank Sum formula; below baseline performance appears in blue, and yellow indicates higher performance. Dataset and term combinations for which ten or fewer pairs were able to be evaluated are excluded and appear as gray missing values; functions for which less than 10% of methods were available due to gene exclusion by NNN, QTC, or SAMBA were removed. Visualization provided by TIGR MeV (Saeed, Sharov et al. 2003).

NNN clusters tend to describe a broader array of biological processes than those of previous methods, and they often relate functional information that might otherwise remain undetected. Figure 24 summarizes each clustering algorithm's maximum performance for each biological function across all six datasets. Of the 88 functions evaluated in this manner, 40 are predicted at biologically uninformative levels (AUC <0.65) by previous methods. NNN improves 18 of these functions to an AUC greater than 0.64 (as high as 0.9 in several cases). It further improves

performance in an additional 21 functions also predicted well (AUC>0.65) by other algorithms. In the concatenated data, NNN improved the best AUC above 0.65 in 14 functions and was the best predictor of an additional 10 beyond those. Particularly since clustering has become an essential part of microarray analysis, it is critical to provide a method such as NNN that will extract the most precise and functionally diverse clusters from a dataset.

Conclusion

We present the Nearest Neighbor Networks clustering algorithm as an efficient and convenient tool for extracting precise, functionally diverse clusters from coexpression data. NNN leaves less active genes unclustered and focuses on networks of potential interaction rather than on minimizing distances; this results in smaller clusters with a high degree of functional relationship as measured by known annotations in the Gene Ontology. Particularly in complex datasets for organisms without comprehensive reference data readily available, NNN's more precise clusters should be beneficial in coexpression analysis. Moreover, these clusters span a wider range of biological processes than those typically extracted from microarray datasets by other clustering algorithms. We hope that these features will allow NNN to serve as a useful method for biologists to obtain an overview of the genes and processes active in new datasets.

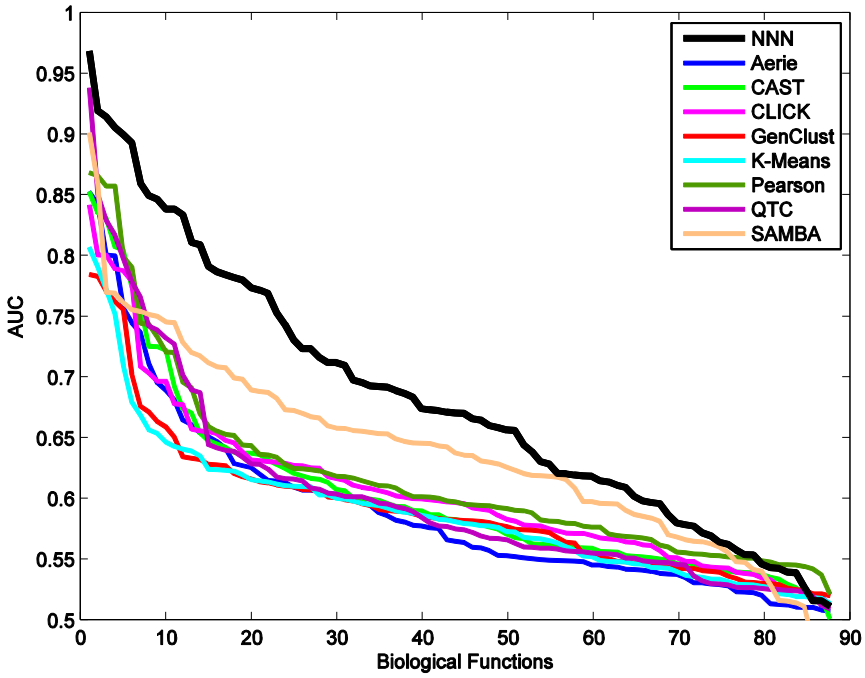


Figure 24: An evaluation of each clustering algorithm's ability to detect the 88 biological processes for which data was available in our analysis. For each algorithm, the maximum AUC across all six datasets was determined, and the resulting AUCs are presented here in descending order per algorithm. NNN correctly clusters genes from substantially more biological processes relative to previous methods.

Graphle: Interactive Exploration of Large, Dense Graphs

As the breadth, depth, and quantity of biological data has continued to grow, this data has increasingly been represented as graphs for the purposes of analysis and visualization. Historically, biological networks have been used to represent the organization of metabolic pathways (Kanehisa, Araki et al. 2008), protein complexes (Schwikowski, Uetz et al. 2000; Iragne, Nikolski et al. 2005), and regulatory networks (Kohn 1999; Baker, Carpendale et al. 2002), often based on painstaking laboratory work carried out before the advent of high-throughput technologies. With the introduction of genome-scale data, datasets from protein-protein

interaction networks (PPIs, (Breitkreutz, Stark et al. 2003; Prieto and De Las Rivas 2006)) to microarray correlations (Chung, Park et al. 2005; Freeman, Goldovsky et al. 2007) have all been represented as graphs. Even computational predictions of regulatory networks (Qian, Lin et al. 2003; Sachs, Perez et al. 2005) or functional relationships (Lee, Date et al. 2004; Myers, Robson et al. 2005) are generally presented as network structures. Most commonly, each vertex indicates a gene and each edge a biological relationship, weighted or unweighted (e.g. expression correlation versus PPIs) and undirected or directed (e.g. PPIs versus regulator/target interactions). Not only do graph structures represent a well-understood computational platform for the analysis of these networks on a whole-genome scale (Milo, Shen-Orr et al. 2002), they offer a rich visual representation of the varied molecular interactions underpinning systems biology.

The visualization of biological networks has inspired substantial research and tool development, ranging from the detailed organization of small, sparse networks as pathways (Gansner and North 2000; Baitaluk, Sedova et al. 2006; Cline, Smoot et al. 2007) to visual overviews of entire genomes (Adai, Date et al. 2004). Unfortunately, many biological networks of interest fall between these two extremes. Genomic data is often large (most organisms of interest have tens of thousands of genes), but not so large that it falls into the class of "huge" network visualization (e.g. maps of the Internet, with some half a billion current hosts). Similarly, while many types of biological networks have a small-world-like property (Middendorf, Ziv et al. 2005) and are thus relatively sparse, other graphs are dense or even fully connected (e.g. microarray correlations); standard visualizations of such graphs usually degenerate into uninformative "hairballs" (Suderman and Hallett 2007). Moreover, regardless of network size, useful biological graph visualizations must allow for wide variation in scale and detail: most biologists, when presented

with a biological network, want to see both the big picture and the specific interactions surrounding their gene(s) of interest.

We have created Graphle as a tool to address these issues and to provide biologists with a tool for exploring large biological networks. As shown in Figure 25, Graphle consists of two parts, the main one being a Java-based client that runs in a user's web browser to display interactive, controllable portions of large biological networks (as well as associated data on genes, protein functions, and experimental datasets). This client allows a user to navigate within a biological network either horizontally, by focusing different sets of one or more query genes and viewing their network neighborhood, or vertically, by including more or less heavily weighted edges and vertices. For example, if edge weights represent microarray correlations, this allows a user to view only the most correlated pairs of genes. Underlying the Graphle client is a server that can run in a centralized location to manage up to hundreds of biological networks, possibly representing several hundred gigabytes of data. Communication between the server and client is optimized so that only the small, focused portions of the underlying networks surrounding a user's query are communicated to the client, which in turn fine-tunes the visualization of this subgraph. Graphle thus allows a user to flexibly explore any biological network and to interactively scale between very general and very detailed visualizations of specific genes of interest. An implementation of Graphle is available online at http:///function.princeton.edu/graphle, showing functional relationship networks predicted for *S. cerevisiae* by the bioPIXIE system (Myers and Troyanskaya 2007) and for human beings by the HEFalMp system (Huttenhower, Haley et al. 2009); a downloadable Java implementation with source code and documentation are also available at this address.

Methods

Graphle is implemented in Java using a client/server architecture to modularize the two main components of the system: a graph server that manages a (potentially very large) collection of weighted graphs and associated metadata, and a user interface client that provides an interactive visualization of portions of this data. This partitions the system to allow hundreds of gigabytes of biological network data to be managed on the server while still providing a focused, responsive user experience. The responsibilities of the graph server include accessing large amounts of graph data on disk in a query-driven manner, caching this data to improve performance, executing graph query algorithms based on client input, and providing information on genes (vertices) and underlying data (edges) as needed. The graph client must run in a web browser and provide rapid, interactive access to all data managed by the server in an informative visualization. Fundamentally, just as Google acts as a query-driven server to intelligently filter the content of the web into a client browser, the Graphle server acts in a query-driven manner to filter the content of biological networks into its interactive client.

Graph server

The Graphle server is based on a Java port of portions of the Sleipnir C++ library for computational genomics (Huttenhower, Schroeder et al. 2008) that allow it to efficiently manage multiple large biological networks. Subgraphs are retrieved from these networks using any graph query algorithm (currently the bioPIXIE (Myers, Robson et al. 2005) and HEFalMp (Huttenhower, Haley et al. 2009) algorithms are implemented) and communicated to the client using a standard socket connection. The graph data organized by the server can include continuous or discrete experimental results (e.g. pairwise correlations from microarray data or protein-protein

interaction networks), predicted interaction networks, ontological structures such as the Gene Ontology (Ashburner, Ball et al. 2000), or any undirected weighted (or unweighted) graphs.

Graph data is stored using the Sleipnir CDat interface, and can thus be interconverted between human-readable text (referred to as the DAT format) and a compact binary (DAB) format. Graphs stored as DABs are automatically indexed and memory mapped; due to memory mapping restrictions on many platforms, an LRU cache is used to maintain a subset of currently mapped graphs. Retiring a graph from this cache, loading a new ~25,000 gene graph, and performing a complete graph query takes at most ~20s on a modern server, most of which time is spent in disk access.

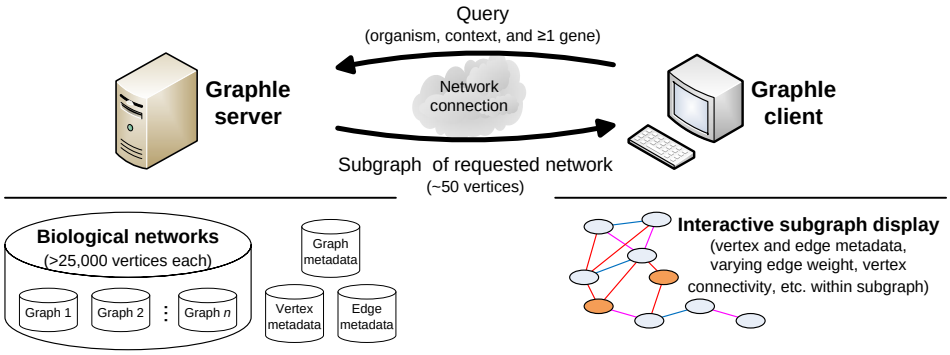


Figure 25: Overview of the Graphle system architecture. The Graphle server manages up to hundreds of gigabytes of weighted undirected graphs; while any graph data can be used, Graphle is specifically designed for biological networks in which vertices represent genes and edges represent experimental results (microarray correlations, protein-protein interactions, etc.) or computational predictions (e.g. probabilities of functional interactions). The server also associates metadata with graphs (such as what organism or biological context they are drawn from), vertices (gene identifiers, aliases, known cellular functions, etc.), and edges (e.g. what experiments or data contributed to that edge). The Graphle client communicates user-provided queries to the server consisting of one or more genes of interest, receives an appropriate subgraph, and displays it interactively for the user in a web browser. The user can then change the focused genes or the stringency cutoff for vertex or edge weights and can access the associated metadata to interactively explore tractable portions of the large underlying graphs.

The graph server also maintains metadata describing graphs, vertices, and edges. Each graph is assigned to a particular organism (or other broad category) and to a "context" within that organism, where a context can be a biological process, tissue type, or other specific subcategory. Vertices are described by a unique identifier (e.g. ORF IDs for yeast genes, HGNC (Eyre, Ducluzeau et al. 2006) symbols for human genes, etc.) and zero or more synonymous aliases; they may also possess zero or more categories of metadata, with each category consisting of an arbitrary dictionary of key/value descriptors (e.g. textual descriptions, Gene Ontology annotations, etc.) Similarly, edges may also be decorated with arbitrary category dictionaries of metadata; this is particularly useful in the case of graphs representing predicted biological networks, as it provides a convenient way to indicate what experimental data was integrated to produce each predicted interaction (Myers, Robson et al. 2005).

User interface client

The Graphle client is a Java applet designed to interactively visualize configurable subgraphs of biological networks (or other graph data) in a web browser. The client uses the Prefuse library (http://prefuse.org) for graph layout, supplementing it with an interface for selecting organisms and contexts, displaying vertex/edge metadata, exporting image or text representations of the current graph, and performing graph queries. These queries consists of a user-provided set of genes (or other vertex identifiers) sent to the Graphle server, which performs a configurable graph query algorithm to return the most relevant portion of the selected (potentially very large) complete graph. In addition to controlling which genes make up the current query, the client also provides realtime filters for vertex and edge inclusion (based on the weight of the graph's edges and the confidence with which the server indicates that nodes are included in the graph query

results). The combination of these three features allows a user to fluidly and tractably navigate through large, dense, weighted graphs.

Results

Graphle provides a web-based system for interactively browsing large biological networks. These graphs can represent experimental results (e.g. protein-protein interaction networks, microarray correlations, etc.), computational predictions (e.g. probabilities of functional interaction), or any other undirected, weighted graphs. Each underlying graph can be very large (tens of thousands of vertices, billions of edges, gigabytes of data), and the Graphle server can manage hundreds of such graphs along with associated metadata (organism, biological context, gene, and dataset descriptors). The Graphle client executes in a user's web browser and retrieves subgraphs focused on a specific set of query genes. This query and the displayed subgraph can be interactively modified in realtime, allowing a user to conveniently explore targeted subgraphs of interest extracted from the large body of underlying data.

Graph queries and exploration

A Graphle query consists of two components: a particular underlying graph specified by an organism and biological context (Figure 26D), and one or more gene identifiers specific to that organism (Figure 26B and C). For example, a Graphle server may have access to several graphs, each covering a specific context in yeast, human, mouse, or another organism's data; contexts represent variables such as biological processes (such as the cell cycle, apoptosis, glucose metabolism, etc.), tissue or cell types, or developmental stages. A user of the Graphle client selects an organism and context from the server-provided list and queries on one or more of the organism's genes. These genes are sent to the server, which uses a graph query algorithm (Myers, Robson et al. 2005; Huttenhower, Haley et al. 2009) to select the subgraph of the requested

network most relevant to the query genes (Figure 26A). This subgraph is of sufficiently small size (~50 fully connected vertices and the associated edge weights) that it can be sent to the client in full; the client then provides a configurable visualization of the subgraph for the user.

Edge weights in biological networks often represent the strength of or confidence in an experimental outcome: greater sequence similarity, higher correlation between gene expression values, or larger probabilities of functional interactions, for example. Similarly, using the concept of guilt by association, most graph query algorithms assume that vertices more strongly connected to the query set in the aggregate are in turn more biologically related to those query genes. Correspondingly, the Graphle client allows a user to fine-tune the visualization of a queried subgraph by filtering edges by weight and vertices by score (Figure 26E); filter changes automatically rerun the graph layout algorithm, which is animated to maintain visual context. A biologist can thus easily visualize both strong and diffuse clusters in the data, expand from the most related genes to more distant neighbors, and easily track the relationship(s) of the original query genes to these neighbors.

Multiple organisms and biological contexts

The Graphle server organizes its collection of graphs using two biologically motivated levels of abstraction: each graph is assigned to exactly one organism and one biological context (Figure 26D). A graph's organism dictates what unique gene identifiers (and non-unique gene aliases) are used to label its vertices, since the server maintains sets of known genes specific to each organism. A context, practically speaking, can be any unique identifier of a particular graph; in practice, a context is often the experiment that generated the graph data, the computational algorithm that generated a set of predictions, a specific biological system (cell/tissue type, pathway or process, subcellular compartment, etc.), or a combination of these. For example, the

Graphle system running at http://function.princeton.edu/graphle offers graphs generated by bioPIXIE (Myers and Troyanskaya 2007) in yeast or HEFalMp for human data (Huttenhower, Haley et al. 2009), with contexts representing different biological processes on which the two algorithms focused.

Gene (vertex) and data (edge) information

Graphle maintains arbitrary metadata optionally describing each vertex (gene) and edge in its graphs (Figure 26G). For genes, this metadata is most often useful for conveying standard knowledge associated with genes: synonymous gene identifiers, chromosomal location, known functions cataloged in the Gene Ontology (Ashburner, Ball et al. 2000) or elsewhere, etc. For edges, this metadata can provide information on the experimental data underlying the graph visualization. This is most important in graphs representing computational data integrations, since each edge might then summarize information from many experimental results - the specifics of which can be provided in the appropriate edge metadata.

Exporting graph images and data

Graphle provides the opportunity for users to export the current subgraph as an image (e.g. for publication) or as raw textual data (e.g. for further analysis, Figure 26F). Data exported in this manner is provided as a simple edge list linking unique vertex identifiers (i.e. gene names) with the weight of the edge joining them (the semantics of which are dependent on the source of the underlying graph). The currently visible, filtered subgraph can be exported as an image of quality suitable for publication.

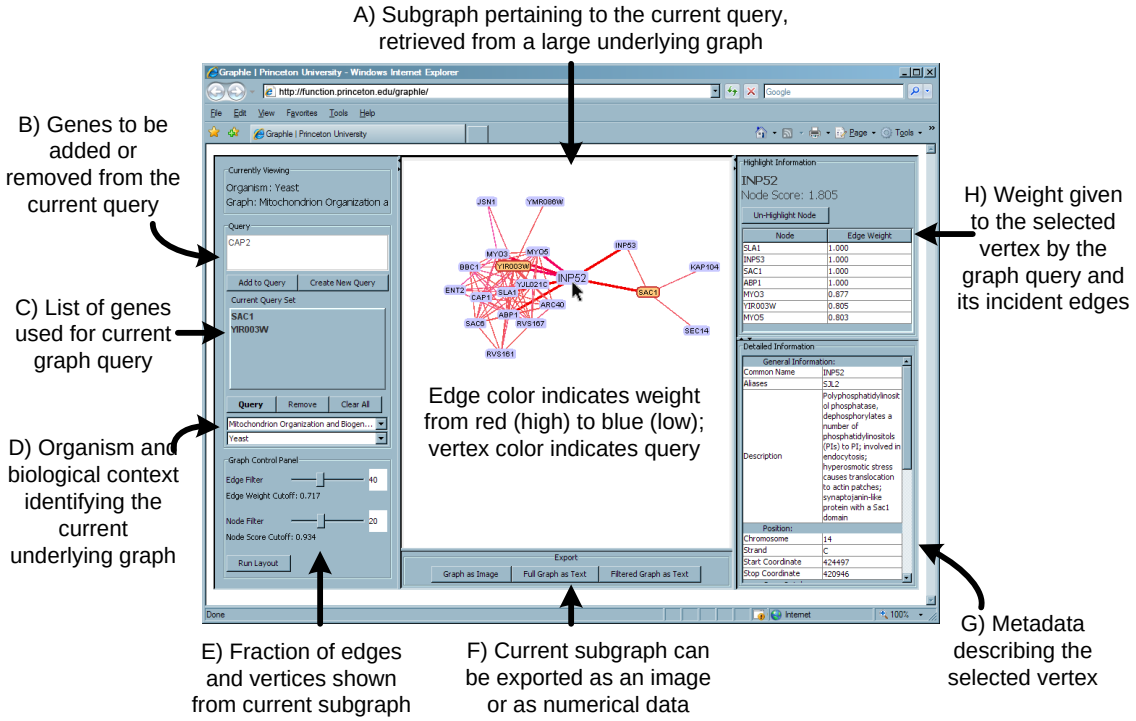


Figure 26: The Graphle client user interface. A user can specify one or more genes that are sent as a query to the server. This information allows the server to execute a graph query in the underlying large biological network specified by the requested organism and biological context. A subgraph comprising ~50 vertices total is returned to the client, which then lays out and displays in real time the most informative portion of this subgraph. The visible subgraph can be controlled by modifying the edge and vertex cutoffs. Detailed information on the numerical scores of the selected node and its incident edges are shown on the right. The current subgraph can be exported as an image (e.g. for publication) or as raw data (e.g. for further analysis).

Conclusion

We present Graphle, a system for interactively exploring large, densely connected biological networks. This task has been particularly challenging in the past due to the impracticalities of storing these graphs (which can each be several gigabytes in size) and visualizing them in an informative manner (as they can be fully connected, but with edge weights varying over a potentially wide range). Graphle allows collections of dense, weighted graphs to be stored on a

server and accessed through focused queries by a web-based client. The data comprised by such graphs can range from experimental results to computationally predicted interaction networks, and Graphle allows each vertex (i.e. gene) and edge to be annotated with arbitrary descriptive metadata. A web-based client sends sets of query genes from a user to the server and interactively displays the resulting focused subgraphs, which can be manipulated in realtime and exported as data for analysis or as images for publication. The Graphle source code, documentation, and a demonstration client can be found at http://function.princeton.edu/graphle. Graphle thus provides a complete solution for storing, sharing, and exploring biological networks.

Meaningful Modeling: Biologically Grounded Statistics of High-Throughput Data

One of the most straightforward ways in which computational tools can be brought to bear on biological problems is by treating them as a target for applied mathematics. Microarrays are perhaps the best example of this phenomenon (Quackenbush 2002). In their raw form as an image taken straight from a scanner, they are simply a tremendous matrix of pixel intensity values - small integral numbers. This image must be gridded (i.e. broken down into discrete spots), tested for quality, normalized globally, locally, and between channels, and probe sets spanning multiple spots per gene must be resolved. In the end, millions or billions of pixels are summarized as a single number per gene. Each of these steps represents transformations between matrices, vectors, and one- or two-dimensional distributions of values; the biological content of the data becomes almost completely irrelevant, and microarray normalization is largely an exercise in linear algebra and applied statistics.

Likewise, interaction networks have been variously analyzed using techniques from graph theory and statistics independently of their biological meaning (Sharan and Ideker 2006). Many biological networks are unweighted and undirected (protein-protein interaction networks, synthetic lethal interactions, etc.) and thus represent prime targets for statistical analyses: degree distribution (Barabasi and Albert 1999), motif frequency (Milo, Shen-Orr et al. 2002), and construction mechanisms (Middendorf, Ziv et al. 2005). Other phenomena are better modeled with directed graphs (Hasty, McMillen et al. 2001) (e.g. regulatory networks or even journal article cocitation), providing additional options for Bayesian analysis and learning (Sachs, Perez et al. 2005). When weights are added to these networks, a host of probabilistic (Segal, Taskar et al.

2001) and differential equation (El-Samad, Prajna et al. 2006) techniques become applicable as well. Again, all of these analysis techniques allow mathematical tools to be applied to data that happens to be of biological origin - often with great success.

Of course, the richness and accuracy of computational modeling of biological data can only be improved by taking the field's extensive prior knowledge of biology into account. An opportunity in this area that is often overlooked is the ability of directed statistical modeling to answer specific biological questions. That is, rather than using general mathematical tools to obtain a bird's eye view of some biological dataset, those same tools can often be applied in a more nuanced manner to discover new biology; even more promising is the ability of such analyses to work in tandem with experimental design. For example, given some collection of gene expression data, a broad statistical analysis can easily retrieve information on genes with high or low variability or with differential expression. More biologically focused methodology can answer questions about the activity of known pathways, protein complexes, or cellular processes. But by integrating such techniques into the experimental design process, one can, for example, collect a small number of microarrays (or other laboratory results) to precisely and quantitatively delineate the activity of a specific process of interest.

This chapter demonstrates three instances of such studies: numerical analyses coupled with directed biological experimentation in such a way as to benefit both analytical aspects. First, we present a statistical linear model of gene expression as regulated by cellular growth rate (defined as change in biomass over time). This model is inspired by and learned from a set of 36 microarrays drawn from continuous *S. cerevisiae* cultures at known growth rates and under known nutrient limitations. Due to the focused nature of this dataset, not only is the model accurately descriptive of the yeast genome's regulatory response to changes in growth rate, it is

predictive of other cultures' growth rates. Given new gene expression data, the model predicts the relative instantaneous growth rate of the originating culture, robust to expression measurement technology, the culture's growth environment, and even some degree of evolutionary distance (as the model has also been successfully applied to other unicellular fungi).

We continue by developing a similar model of the effects of aneuploidy on gene expression in yeast. Aneuploid cells possessing an incorrect number of copies of one or more chromosomes have long been known to be common in many tumors (Boveri 1902), and incorrect chromosomal copy number is directly responsible for a variety of other genetic disorders (e.g. Down syndrome (Epstein 2006)). By constructing yeast mutants monosomic (one copy in a diploid cell), disomic (two copies in a haploid cell), or trisomic (three copies in a diploid cell) for individual chromosomes, we can model the resulting gene expression changes universally present in these aneuploid states and not due merely to changes in a gene's copy number. This demonstrates a variety of yeast-specific biological responses (e.g. a marked downregulation of mating pathways and upregulation of mitotic cell division in trisomes) as well as potential markers of a general aneuploidy response (primarily, as expected, decreased growth and increased transcription, translation, and protein degradation). The model thus provides insights into the molecular mechanisms of cancer, since it suggests regulatory mechanisms by which cell deal with increased protein load as chromosomal copy numbers increase. Likewise, it raises questions for future research, since the putative fitness advantage conferred on cancer cells by various aneuploidies must somehow offset the general growth defect caused by chromosomal copy number changes.

Finally, we describe an analysis of the *S. cerevisiae* phosphorylation network in an evolutionary context. That is, given a genome-scale measurement of phosphorylated proteins and specific amino acid residues, what can we observe about the conservation of these biological interactions

in other organisms, and how does this reflect on the behavior and robustness of phosphorylation as a regulatory mechanism? An immediate observation also made by others (Jeong, Mason et al. 2001) is that phosphoproteins tend to interact with each other and with other proteins significantly more than would be expected by chance, confirming the role of phosphorylation as a central regulatory mechanism (particularly in processes involving the mitotic cell cycle). Surprisingly, while phosphorylation interactions are very strongly conserved across large evolutionary distances (e.g. from yeast to human), individual phosphorylation sites are not - a finding later confirmed by others (Beltrao and Serrano 2007). This is thought to be indicative of the evolutionary benefits of maintaining some plasticity in regulatory networks, i.e. maintaining a balance between properly functioning regulatory interactions while also providing an opportunity to acquire beneficial mutations. All three of these results - growth rate, aneuploidy, and phosphorylation - are predicated on straightforward mathematical analyses applied to targeted biological data and demonstrate the potential of such an integrated approach.

We would like to thank Matthew J. Brauer, Edoardo M. Airoldi, David Gresham, and David Botstein for their collaboration in modeling cellular growth rates; Maitreya J. Dunham for her extensive experimental work in studying aneuploidy; and An Chi and Donald F. Hunt for their examination of the yeast phosphoproteome.

Transcriptional Regulation and Cellular Growth Rates

Proper regulation of growth rate is a key systems-level challenge for all cells, particularly microorganisms facing an ever-changing and often hostile environment. Cell growth, defined as an increase in cellular biomass due to biosynthetic processes, is one of the primary functions that

must be coordinated with the environment in order for cells to maintain viability and reproduce. It is of central importance to our understanding of basic biology to determine how cells integrate information from the external environment and from their internal state to mount an appropriate response: growing in the presence of nutrients, arresting growth when stressed, and resuming afterwards. From a genomic perspective, growth also raises the issue of disentangling correlated systems-level behaviors and determining causality. When the expression levels of thousands of genes change due to a growth-related stimulus, which underlying regulatory parameters are responsible?

In this paper, we identify quantitative aspects of the transcriptional regulatory mechanisms underlying cell growth in *Saccharomyces cerevisiae* and develop a model to predict instantaneous growth rates of cellular cultures based on gene expression data. This provides a mechanism for estimating growth rates under any conditions for which microarray data is available, even at time scales too brief to measure with standard experimental techniques (Amberg, Burke et al. 2005). For example, a culture undergoing continuous growth in a chemostat (Hayes, Zhang et al. 2002) can be perturbed from steady state by means of a short heat pulse, but the departure from and return to steady state growth is too brief to observe conveniently with optical density measurements. Our model allows such a decrease (and subsequent resumption) of growth rate to be quantified under a variety of conditions: batch or chemostat cultures, different microarray platforms, and under any environmental stimulus for which gene expression can be assayed. Surprisingly, this model also successfully predicts growth rates from *Saccharomyces bayanus* and *Schizosaccharomyces pombe* microarray data, the latter of which is evolutionarily diverged from *S. cerevisiae* by an estimated billion years (Hedges 2002).

By analyzing mRNA abundance data we obtained from 36 chemostat cultures (six different limiting nutrients each at six different growth rates), we found that a surprisingly large fraction (ca. 27%) of all yeast genes are expressed (as measured by relative mRNA abundance) in a way that is closely correlated (either negatively or positively) with the growth rate of the culture. We showed that the statistically well-defined functional subsets of genes whose expression is most sensitive to the growth rate are ones that have been observed previously as coherent groups of genes with coordinated behaviors in response to such disparate experimental contexts as environmental stress (e.g. (Gasch, Spellman et al. 2000)) and synchronous metabolic oscillations (e.g. (Tu, Kudlicki et al. 2005)) and as subsets of genes whose mRNA abundances appear to be substantially regulated by changes in mRNA stability (Grigull, Mnaimneh et al. 2004).

This suggests that our statistical model of cellular growth can provide a broadly applicable biological characterization of the transcriptional regulatory network underlying growth rate control. This response is functionally cohesive, with genes upregulated with increasing growth enriched for translational and ribosomal functions and downregulated genes enriched for oxidative metabolism and the peroxisome. This provides a rich environment in which to study transcripitional growth regulation; for example, production of new proteins at the ribosome is vital to cellular proliferation, and yeast devotes some ~60% of its transcriptional throughput to ribosomal RNA (Warner 1999). Similarly, growth rate regulation is highly interconnected with a variety of other cellular processes (e.g. the environmental stress response (Gasch, Spellman et al. 2000), metabolic cycling (Klevecz, Bolen et al. 2004), and the cell cycle (Pramila, Wu et al. 2006)), and we discuss potential causative regulatory signals from the Ras/PKA pathway (Wang, Pierce et al. 2004) and growth-related transcription factors.

Here, we demonstrate that the model can accurately predict relative growth rates under a variety of conditions and is robust to the conditions of the originating culture, the technological platform used to assay gene expression, and evolutionary conservation to other organisms (*S. bayanus* and *S. pombe*). This allows us to predict growth rates for published microarray collections (e.g. the stress response (Gasch, Spellman et al. 2000) or gene deletions (Hughes, Marton et al. 2000)) and for new data we have generated, providing biological insight into the growth rate response at very short time scales - minutes, rather than the hours necessary to experimentally assay doubling times. These real-world biological predictions are accompanied by an out-of-sample validation and outlier analysis to establish the model's statistical accuracy. We have made an implementation of this model available to the public at http://function.princeton.edu/growthrate.

We also apply our model to specifically study two important aspects of cell growth regulation, nutrient sensing and the cell cycle. Artificial activation of the Ras/PKA pathway has been previously observed to recapitulate approximately 85% of the expression response associated with increased growth in the presence of glucose (Zaman, Lippman et al. 2008); here, we show that the cell's regulatory state during this activation is indicative of an upregulated growth response, even in the absence of appropriate nutrient availability. This conflict between internal regulatory state and the external environment leads to rapid cell death. In contrast, analysis of growth rate regulation during metabolic cycling (Tu, Kudlicki et al. 2005) and synchronous cell cycles (Spellman, Sherlock et al. 1998; Pramila, Wu et al. 2006) indicates that growth rate regulation is not specific to cell cycle phases, but it is strongly limited to the oxidative phase of the metabolic cycle. These observations, coupled with an analysis of putative transcription factors mediating the growth response, establish a substantial foundation on which to base further experimental work on the systems-level control of cellular growth rate.

Materials and Methods

We fit a linear model to a collection of expression data drawn from *S. cerevisiae* chemostat cultures over several growth rates and nutrient limitations. This model provides estimates of parameters that characterize each gene's response to changes in growth rate, and these provide insight into the transcription factors and regulatory network responsible for yeast growth homeostasis. By applying this model to new expression data sets, we are able to predict instantaneous growth rates for any yeast culture. This inference process is robust to the biological and technical conditions of the originating gene expression data and predicts growth rates at instantaneous time scales inaccessible to standard experimental methods (e.g. optical density). We have also successfully applied the model to the related organisms *S. bayanus* and *S. pombe*. Data and tools relating to this model are made available at http://function.princeton.edu/growthrate.

Experimental design and data

Our model is based on a collection of gene expression measurements from steady state (chemostat) cultures limited across several nutrients and growth regimes (Brauer, Huttenhower et al. 2008). This experimental design provides the opportunity to discover gene expression patterns correlating with growth rate independently of nutrient-specific responses. Briefly, 36 CEN.PK derived *S. cerevisiae* chemostat cultures were grown under six nutrient limitations: Glucose (G), Nitrogen (N), Phosphate (P), Sulfur (S), Leucine (L), and Uracil (U). Six growth rates were used for each nutrient, ranging by steps of $0.05h^{-1}$ from $0.05h^{-1}$ to $0.3h^{-1}$. Agilent Yeast V2 microarrays were used to measure gene expression in the resulting 36 chemostats; for details, see (Brauer, Huttenhower et al. 2008).

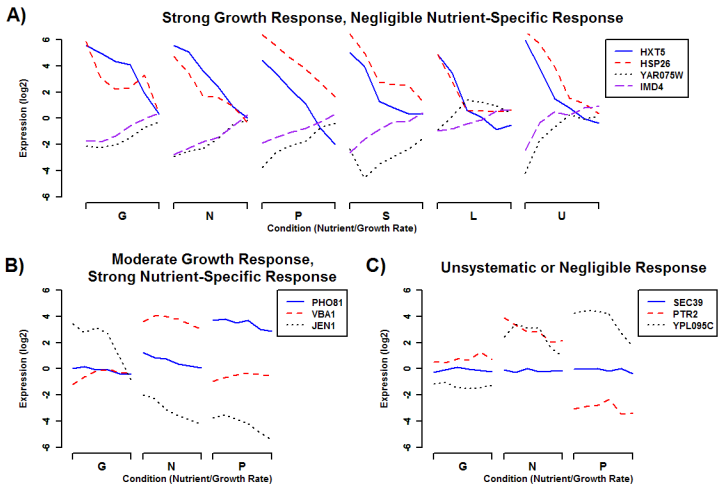


Figure 27: Representative genes responding to growth rate, specific nutrients, or unsystematically in our chemostat-derived training data. Our statistical model of growth rate regulation is based on microarray data collected from 36 chemostats at six growth rates (0.05hr^{-1} through 0.3hr^{-1}) under six nutrient limitations (Glucose, Nitrogen, Phosphate, Sulfur, Leucine, and Uracil) as described in (Brauer, Huttenhower et al. 2008). By employing the genes responding strongly, consistently, and only to changes in growth rate (and not specific nutrients) as growth-specific genes, we can apply our model to predict relative growth rates in new expression data. Gene expression in our original 36 conditions fell into three main categories as shown here. A) Genes strongly up- or down-regulated in response to changes in growth rate, independent of limiting nutrient. The most statistically significant members of this set became our growth-specific calibration genes for application of the linear model to other expression data. B) A subset of conditions highlighting genes with expression levels showing some correlation with growth rate, but with a strong nutrient-specific component. This represents a sizeable portion of the genome (~25%), with positively growth-correlated genes enriched mainly for ribosomal function and negatively correlated genes enriched for oxidative metabolism. C) A subset of conditions highlighting genes showing a non-systematic or negligible change in gene expression. Unresponsive genes were enriched for a variety of cellular processes not expected to show a strong relationship with growth, e.g. transcription, DNA metabolism and packaging, secretion, and many others.

Figure 27 highlights the sources of variability in the gene expression profiles that the experimental design aims at capturing. The resulting data contain a number of characteristic gene expression patterns, including genes with strong growth-specific transcriptional regulation and negligible nutrient-specific response (Figure 27A). Other genes include a growth-specific expression component but are also strongly up- or down-regulated under specific nutrient limitations (Figure 27B). Finally, Figure 27C displays expression profiles that show unsystematic or negligible responses under these conditions. The linear model described below summarizes the variability in the expression profiles of individual genes specifically due to changes win growth rate, which leads to a characterization of growth-specific calibration genes such as those shown in Figure 27A. This growth-specific signature enables predictions of the instantaneous growth rate of any cellular culture based on the relative expression values these growth-specific genes.

Table 6 summarizes the collections of expression data analyzed in this study. Six collections were previously published by others, one was published in our previous work (Brauer, Huttenhower et al. 2008), and four are new to this study: 1. chemostats limited for different nitrogen sources, 2. heat pulses inducing a temporary departure from steady state growth, 3. artificial activation of the Ras/PKA pathway, and 4. *S. bayanus* diauxic shift and heat shock time courses. All gene expression collections were pre-processed as in (Huttenhower, Hibbs et al. 2006).

Linear models and identification of growth-specific signature

We sought to identify a small set of genes providing a quantitative summary of cellular growth rate regulation. These 36 chemostat-derived microarrays provided us with the opportunity to determine which genes were responding consistently (i.e. linearly) and only to changes in growth rate (and not to differences in nutrient limitation). To model this statistically, we performed four

steps, beginning by using maximum likelihood to fit a linear model of each gene g's expression under all training conditions ($\mathbf{Y}_g$) based on the conditions' known growth rates ($\mathbf{X}_c$):

$$\mathbf{Y}_g = \alpha_g + \beta_g \mathbf{X}_c + \boldsymbol{\varepsilon}_g$$

Experimental Conditions	**Method**	**Platform**	**Organism**	**Publication**
Nutrient-limited growth	Chemostat	Agilent	*S. cerevisiae*	(Brauer, Huttenhower et al. 2008)
Cell cycle synchronization	Batch	Spotted	*S. cerevisiae*	(Spellman, Sherlock et al. 1998)
Cell cycle synchronization	Batch	Spotted	*S. cerevisiae*	(Pramila, Wu et al. 2006)
Metabolic cycling	Batch/Chem.	Affymetrix	*S. cerevisiae*	(Tu, Kudlicki et al. 2005)
Environmental stress	Batch	Spotted	*S. cerevisiae*	(Gasch, Spellman et al. 2000)
Gene deletion mutants	Batch	Spotted	*S. cerevisiae*	(Hughes, Marton et al. 2000)
Heat pulses	Chemostat	Agilent	*S. cerevisiae*	C. Lu
Nitrogen-limited growth	Chemostat	Agilent	*S. cerevisiae*	D. Gresham
RAS/PKA activation	Batch	Agilent	*S. cerevisiae*	J. R. Broach
Diauxic shift, heat shock	Batch	Spotted	*S. bayanus*	A. A. Caudy, M. J. Dunham
Hydroxyurea response	Batch	Spotted	*S. pombe*	(Chu, Li et al. 2007)

Table 6: Overview of expression data analyzed in this study. Of the 11 gene expression data sets for which we predict and discuss growth rates, four are previously unpublished. These data span various experimental conditions, dual- and single-channel expression array platforms, batch and steady-state growth regimes, and three species of yeast. Under these varied conditions, our growth model predicts instantaneous growth rates and provides insight into regulatory mechanisms for growth homeostasis.

This yields two learned parameters per gene, a baseline expression level α_g and a growth rate response β_g. The model is fit to minimize the residual error ε_g, which can represent either non-growth-related biological variability or technical noise. We fit this model for the yeast genome using the expression levels from our 36 chemostat conditions, recording each gene's α_g and β_g parameters and its goodness of fit (total explained variability) R^2_g.

We next used the bootstrap (i.e. randomized resampling) to assess the expected background distributions of these parameters in the absence of a growth-related biological signal (i.e. the null distributions). We constructed 100,000 randomized expression vectors of length 36 by sampling each condition from all equivalent growth across all genes and nutrient limitations. For example, the first value randomly chosen for such a vector could be drawn from any gene or nutrient limitation in our chemostat data at a flow rate of $0.05h^{-1}$, the second from any flow rate of $0.1h^{-1}$, and so forth. This sampling scheme maintains rate-specific information while normalizing for gene- and nutrient-specific signals, producing an estimate of the null distribution in the absence of growth related gene expression. This process yields null distributions for parameters α_g, β_g, and the goodness of fit R^2_g.

Third, from these null distributions, we assign false discovery rate corrected p-values (Benjamini and Hochberg 1995) to each gene's α_g, β_g, and R^2_g values. Finally, a gene was deemed to have a significant expression response to changes in growth rate if it fit this model well (R^2_g $p<0.05$) and was up- or down-regulated significantly with growth (β_g $p<0.05$); this information is available in (Brauer, Huttenhower et al. 2008). We further characterized a specific set of growth-specific calibration genes responding only and significantly to changes in growth rate (β_g $p<10^{-5}$ and R^2_g $p<10^{-5}$) that we used to infer instantaneous growth rates in new expression data (Supplemental Table 1, see also (Brauer, Huttenhower et al. 2008)).

Model-based prediction of instantaneous growth rates from expression data

The set of growth-specific calibration genes defined above represents a quantitative signature of a cellular culture's transcriptional regulation of growth rate, i.e. the speed at which its cells are proliferating. By examining these genes' expression levels in new data, we can thus predict the instantaneous growth rate of the originating cellular culture. This instantaneous growth rate is comparable to the derivative of an optical density growth curve, but it can be inferred robustly by our model on any time scale (e.g. minutes) from microarray data without the need to measure one or more full doubling times of a culture.

Given expression data for a new experimental condition, we use an iterative maximum likelihood approach to infer its growth rate using the parameters captured by our linear model. Formally, consider a vector of expression measurements for n calibration genes, $\mathbf{Z}_{1:n}$. As described above, the expression of these growth-specific genes varies primarily in response to changes in a condition's growth rate, which we model as the mean μ of a Gaussian with variance σ^2. Using our previously calculated maximum likelihood estimates of the calibration gene parameters $\boldsymbol{\alpha}_{1:n}$ and $\boldsymbol{\beta}_{1:n}$, the expected value of a gene's expression is thus:

$$E[\mathbf{Z}_i] = \alpha_i + \beta_i \mu + \delta$$

Here, δ is a condition-specific parameter capturing the condition's baseline gene expression, i.e. an average offset between a new experimental condition and our training data. In dual-channel data, this represents differences between a new condition's reference channel and our training data; for a single-channel array, it captures the absolute difference between the platform baseline and our training data. Similarly, the expected variability is:

$$V[\mathbf{Z}_i] = \beta_i^2 \sigma^2$$

The likelihood of the expression measurements $\mathbf{Z}_{1:n}$ is thus a product of Gaussians:

$$L[Z_{1:n}] = \prod_{i=1}^{n} N(\alpha_i + \beta_i \mu + \delta, \beta_i^2 \sigma^2)$$

From this, we derive the maximum likelihood estimate of the condition's growth rate μ_{ML}:

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i - \alpha_i - \delta_{ML}}{\beta_i}$$

Similarly, the maximum likelihood estimate of the condition's baseline δ_{ML} is given by:

$$\delta_{ML} = \frac{1}{n} \sum_{i=1}^{n} Z_i - \alpha_i - \beta_i \mu_{ML}$$

Note that the estimate of δ_{ML} depends on the estimate of μ_{ML}, and vice versa. To calculate these estimates, we initialize $\mu_{ML}^{(0)}$ assuming $\delta_{ML}^{(0)}=0$ and iterate subsequent computations of $\mu_{ML}^{(t+1)}$ and $\delta_{ML}^{(t+1)}$ to convergence. In practice, growth-specific calibration genes with residuals outside the inner fences of all calibration gene residuals (more than 1.5 interquartile ranges from the lower or upper quartiles (Moore and McGabe 2005)) are noted as outliers and removed from that condition's growth rate inference. This allows outlier genes responding to non-growth related stimuli (which are, in general, infrequent, e.g. six in one of our most variable conditions as discussed below) to be noted for further investigation while also decreasing the cross-validated error of predicted growth rates.

Extending predictions to additional organisms

In principle, this model of growth rate can be extended to study and predict instantaneous growth in any organism for which appropriate homology exists to our set of growth-specific calibration genes. To analyze growth rates in expression data from *S. bayanus* and *S. pombe*, the *S.*

cerevisiae calibration genes were mapped to known orthologs. This mapping was performed using the unambiguous pairings from (Kellis, Patterson et al. 2003) for *S. bayanus* and the curated orthologous groups from (Penkett, Morris et al. 2006) for *S. pombe*. This resulted in 51 growth-specific genes for *S. bayanus* and 74 for *S. pombe*, the increase being due to one-to-many mappings; see Supplemental Table 1.

Online tool availability

The data driving our model (individual genes' growth rate response parameters) and tools allowing users to predict growth rates in new data sets are available at http://function.princeton.edu/growthrate. Specifically, users can upload *S. cerevisiae* expression data (single- or dual-channel in standard PCL format) to receive estimates of relative growth rate for each condition. If a reference with known growth rate is provided, absolute rate estimates will be generated. This growth rate prediction tool has been implemented in R and is also available for offline use, allowing further customization (such as application to additional organisms).

Results

We apply our linear model of growth rate regulation to predict instantaneous growth rates for a variety of expression data. This includes new chemostat cultures used to assess prediction quality, publicly available stress response and gene deletion microarrays from batch cultures, growth differences between metabolic cycling and the cell cycle, several different microarray platforms, and an out-of-sample validation to quantify model accuracy. We also observe good predictive performance for growth rates in *S. bayanus* and *S. pombe* data sets, the latter despite up to a billion years of evolutionary divergence from our *S. cerevisiae* training data. This suggests that the growth-related transcriptional regulation captured by our model is a key feature of

unicellular homeostasis, a feature we explore by examining nutrient sensing inputs through the Ras/PKA pathway and potential growth rate transcription factors and binding sites.

Growth rate accounts for a large fraction of the signal in the gene expression pattern

Hierarchical clustering (Eisen, Spellman et al. 1998) of gene expression from the 36 chemostats is shown in Figure 28. Visual inspection of Figure 28 shows a pattern that is strikingly similar for the six different media, with large groups of genes that increase their expression with increasing growth rate, and comparably large groups decreasing their expression with increasing growth rate. In addition, there are much smaller clusters of genes that are expressed strongly in only one or two media; in general, these do not show as much relationship to growth rate. Fewer than 8% of the genes respond in a uniform, nutrient-specific manner (Figure 28). The largest nutrient-specific cluster, responding to phosphate limitation, comprises just 133 genes (2.4% of the total).

Expression of many genes is strongly influenced by growth rate to a degree characteristic for each gene

For each of the 5,537 genes in the imputed data, expression was individually modeled as a linear function of growth rate (independent of limiting nutrient). Of these, 3,049 (55.1%) have expression patterns than fit (at a bootstrapped $p<0.05$) this linear model; of these, approximately half (1,470, or 26.5% of all genes) have expression patterns that respond significantly (bootstrapped $p<0.05$) to growth rate.

For those genes whose expression is correlated to growth rate, the magnitude of the effect of growth rate on gene expression is given by the slope of the regression of expression on growth rate. The distribution of these slopes (shown in Figure 29 as a histogram) is significantly broader than the null distribution generated by bootstrap sampling (standard deviation 2.97 vs. 1.41). By

plotting the positions of a set of genes on this histogram, one can see systematic relationships between the aggregate response to growth rate and any other characteristic a query set of genes might share. At http://growthrate.princeton.edu, we provide data for all the genes, including the significance with which their expression correlates with growth rate. On the website, we also provide a simple utility that plots the distribution of growth rate slopes for any query set of genes relative to the overall distribution shown in Figure 29.

Functional roles of genes strongly correlated with growth rate

In order to identify the most prominent of the potential functional reasons for the correlation between the expression of some genes and the growth rate, we chose 1,608 genes whose expression was best linearly correlated with growth rate (see Methods for details). One subset (337 genes) had negative slopes (more than 1.5 standard deviations less than the average), another subset (291 genes) had positive slopes (more than 1.5 standard deviations more than the average) and the third had low variability (bootstrapped $p<10^{-4}$) and low slope (within 0.5 standard deviations of average), i.e. their expression was not detectably related to growth rate (see the dash-dotted blue line in Figure 29).

Each of these subsets was submitted to GO Term Finder (Boyle, Weng et al. 2004) querying all three ontologies (Process, Molecular Function, and Cellular Component). The nature of the GO hierarchy produces highly redundant results in this situation. To maintain only the biologically and statistically strongest relationships, we limited further analysis to GO terms with $p<10^{-3}$. These results were sorted by the fraction of genes hit within each GO term, and we focused on the terms in which this fraction was at least 15%. The final result (Table 7) gives a very clear picture: the processes associated with the gene subset whose expression is negatively related to growth rate are focused on energy metabolism, especially oxidative metabolism; the only

functional category that met our stringent criteria was oxidoreducase activity, and the only cellular component implicated at this level of statistical stringency was the peroxisome. Our negatively-growth-correlated subset contained 25% and 67% (designated hits/size in Table 7) of the genes annotated to peroxisomes and the peroxisomal matrix, respectively.

An equally clear picture emerged from the GO term analysis of the subset of genes whose expression is positively correlated with growth rate. About half of all the yeast genes associated with mitochondrial protein import are found in this subset, and substantial fractions of all the genes associated with translation, ribosome biogenesis, and rRNA metabolism are represented. Consistently, ribosomal constituents (mitochondrial as well as cytosolic) are very strongly represented in both the Function and Component hierarchies.

In contrast, the 980 genes whose expression is robustly independent of growth rate are annotated (with similar statistical certainty and with similar Term Fraction values) to very many (ca. 80) diverse GO terms. The many processes that underlie basic cytoplasmic cell biology or non-nucleolar nuclear biology are well represented in this subset of genes, i.e. those whose expression is unrelated to the growth rate.

Genes whose expression is correlated with growth rate are highly represented in the Environmental Stress Response

One of the questions that can be addressed using the tools developed above is the relationship between growth rate and the many genes whose expression changes regardless of the nature of the environmental stress imposed. In their dataset of 156 such stress conditions, (Gasch, Spellman et al. 2000) found two clusters of genes that were either induced or repressed together in this way, which they called the "environmental stress response" (ESR) genes.

The distributions of the 283 genes in the ESR-induced cluster (red) and the 585 genes in the ESR-repressed cluster (green) are superimposed on the histogram of expression versus growth rate slopes in Figure 30. Both clusters had very high representations of growth-rate-correlated genes and constitute sets of statistically significant outliers to the overall distribution of slopes.

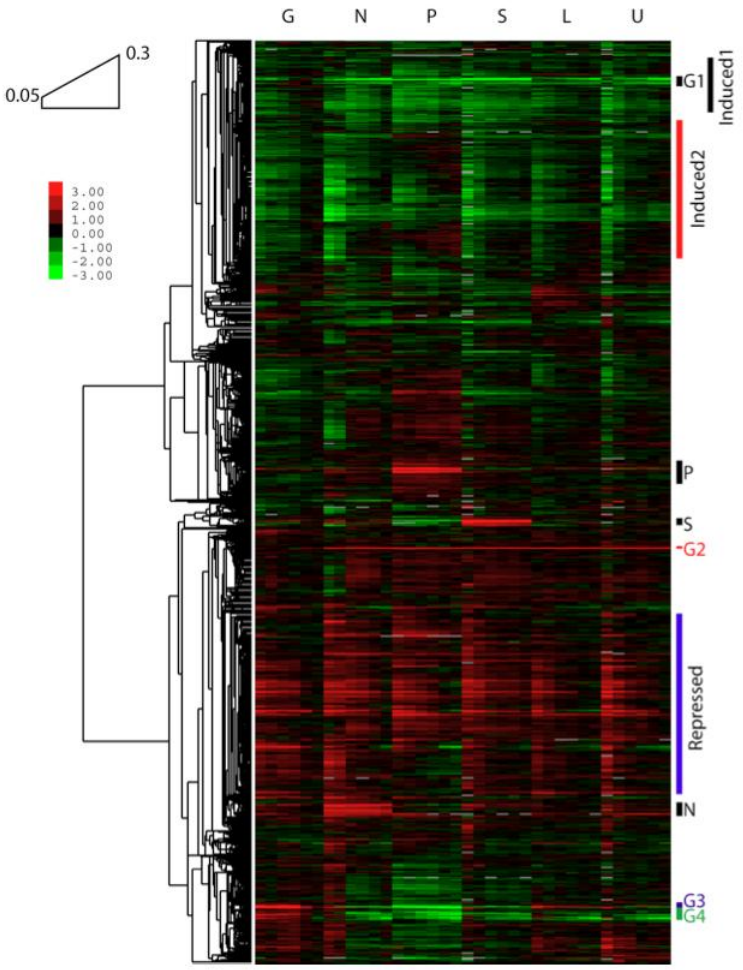


Figure 28: Hierarchical clustering of expression values across dilution rates and limiting nutrients. Clustering by Pearson correlation reveals many up- and down-regulated clusters spanning all nutrient limitations (e.g. Induced1, Induced2, Repressed) and a few smaller gene groups regulated in a nutrient-specific manner (e.g. G1-G4, P, S, and N). Reference for all samples is from a glucose-limited chemostat at 0.25 hr^{-1}.

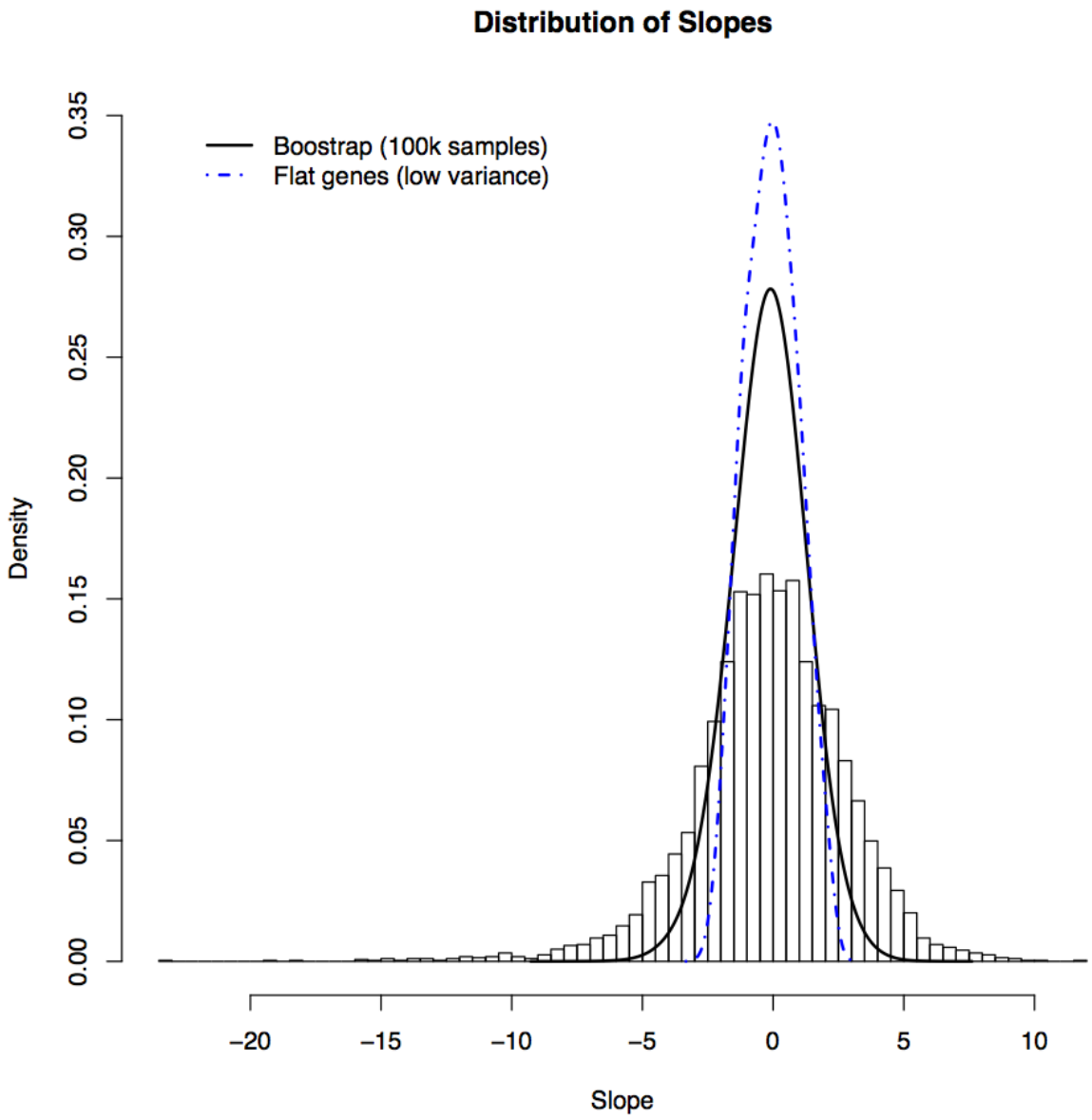


Figure 29: Distribution of experimental growth rate responses versus bootstrapped background distribution. A histogram of the estimated regression slopes for 5,537 genes is compared with a 100,000-point bootstrapped null distribution of slopes (density estimate; black, solid line) and to the distribution of slopes corresponding to genes that do not respond to growth rate (density estimate; dashdotted, blue line). The expression responses of genes in our microarray data are significantly broader than expected by chance, whereas genes we determine to be largely unresponsive to changes in growth rate have slopes near zero.

The 283-gene "ESR induced cluster" has p-values corresponding to Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney two-sample tests practically equal to zero, indicating that this cluster has a very significant representation of genes whose expression is negatively correlated with growth rate. In fact, nearly one quarter of the top 500 genes with negative growth rate regulation are in the ESR-induced cluster found by (Gasch, Spellman et al. 2000). The remaining 376 genes are enriched for GO process annotation terms relating to carbohydrate metabolism, particularly respiratory metabolism and oxidative phosphorylation. The component annotation terms indicate a significant enrichment in genes for proteins localized to the lytic vacuole and the peroxisome.

The "ESR repressed" cluster (585 genes) also has Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney two-sample test p-values of approximately zero, indicating a highly significant positive correlation with growth rate. Here again the 500 genes with the most significant positive correlations between expression and growth rate include 227 of the ESR repressed genes. The remaining growth rate correlated genes are significantly enriched for GO process terms that include membrane lipid biosynthesis and protein import into the mitochondrion.

Not all the ESR genes are expressed in such a way as to be significantly correlated with growth rate in our data. We detected no enrichment, for example, in genes for chaperone proteins and for some other classes of classical stress response genes. The significant majority of stress-induced genes that are expressed at increasingly low growth rates are related to oxidative metabolism; of 131 genes, 16% are oxidoreductases.

These results raise the possibility that many of the ESR genes as defined previously may in fact not be responding directly to stress, but instead are responding to a reduction in growth rate

secondary to the stress. A similar suggestion has recently been made in (Castrillo, Zeef et al. 2007) based on a similar set of observations.

Relative growth rate prediction in novel experimental settings

Our model of the growth rate transcriptional response can be used to predict relative instantaneous growth rates from any *S. cerevisiae* gene expression data. For example, Figure 31A shows our predicted growth rates for a gene expression time course sampled from a steady state culture exposed to a brief (<30s) heat pulse. The predictions clearly show a departure from steady state within five minutes of the heat pulse, followed by recovery within 15 minutes. Similar predictions over a range of chemostat flow rates (Supplemental Figure 2) reveal that this cellular behavior is consistent, although there is some variation in the degree of growth cessation during stress, in agreement with tolerance and sensitization models of the yeast stress response (Attfield 1997). Notably, standard experimental assays for growth rate (e.g. optical density) would be incapable of monitoring such a response, while our model is able to observe these growth changes on an instantaneous time scale.

A similar application of our model to predict relative growth rates for the stress response conditions of (Gasch, Spellman et al. 2000) is presented in Figure 31B (see Supplemental Figure 3 for complete results). These data represent batch yeast cultures assayed using a variety of different reference mRNA samples on a custom spotted microarray platform, none of which differences from our training data impair the growth rate estimation process. While there are no direct measurements of growth rate in these non-steady-state conditions, our predictions are consistent with known yeast biology and agree with expected growth behavior. Most shock time courses, including all heat shocks, peroxide, diamide, and hyper-osmotic stress, provoke an initial sharp decrease in growth rate followed by a return to initial or near-initial rate; shorter

shocks, such as DTT, menadione, and peroxide responses, capture only the rate decrease. Batch growth proceeds at a fairly constant rate until nutrients become depleted, at which point the rate decreases sharply; this pattern is also seen in intentional nitrogen depletion. Growth rates across varying temperatures peak as expected at 25C (Amberg, Burke et al. 2005), falling off at lower and higher temperatures. Finally, response to varying carbon sources is also as expected (Granot and Snyder 1993), with ethanol inducing the slowest growth and fructose, sucrose, and glucose allowing the most rapid. Our model's inference of growth rate from gene expression data alone thus allows both post hoc growth analysis (e.g. years after the original experiment) and an estimation of growth rates for cultures where it would be difficult or time consuming to measure directly.

When applied to microarray data from yeast mutants, in which one or more genes have been deleted, predicted growth rates can be used to approximate single mutant fitnesses. We used our model to analyze the knockout collection assayed in (Hughes, Marton et al. 2000); predictions on the complete data set are available in Supplemental Table 2. (Hughes, Marton et al. 2000) provides direct fitness measurements for 199 of the ~300 mutants assayed by the microarrays. Our predictions for these 199 growth rates correlate very strongly with their measured values (ρ=0.473, $p<10^{-11}$) and are derived solely from expression data. In contrast, methods for experimentally estimating single mutant fitness from high throughput growth curves showed substantially less agreement (ρ=0.321, $p<10^{-6}$ (Warringer and Blomberg 2003); ρ=0.108, $p>0.2$ (Jasnos and Korona 2007)) with the original publication's direct measurements. This represents a compelling argument as to the relevance of our growth rate model for fitness estimation.

Slope >1.5 SDs below average				337 genes
Process	**p-value**	**Gene hits**	**Term size**	**Term fraction**
Fatty acid β-oxidation	1.80E-04	6	8	75.0
Glutamine family amino acid catabolic process	4.20E-04	7	13	53.8
Energy reserve metabolic process	2.61E-05	12	36	33.3
Glucose metabolic process	8.10E-04	14	65	21.5
Monosaccharide metabolic process	1.10E-04	18	92	19.6
Hexose metabolic process	9.50E-04	16	85	18.8
Cellular carbohydrate metabolic process	7.75E-12	40	213	18.8
Monocarboxylic acid metabolic process	5.86E-06	23	124	18.5
Carbohydrate metabolic process	1.37E-12	43	233	18.5
Energy derivation by oxidation of organic compounds	4.59E-06	25	143	17.5
Generation of precursor metabolites and energy	5.02E-07	30	181	16.6
Coenzyme metabolic process	1.40E-04	22	135	16.3
Component				
Peroxisomal matrix	1.45E-06	8	12	66.7
Microbody	3.53E-05	14	57	24.6
Peroxisome	3.53E-05	14	57	24.6
Slope <1.5 SDs below average				291 genes
Process	**p-value**	**Gene hits**	**Term size**	**Term fraction**
Protein import into mitochondrial matrix	1.00E-07	11	22	50.0
Maturation of SSU-rRNA	1.68E-10	17	44	38.6
Protein import into mitochondrion	4.67E-07	13	37	35.1
Protein targeting to mitochondrion	2.23E-05	13	49	26.5
Mitochondrial transport	6.11E-05	14	62	22.6
Function				
snoRNA binding	5.52E-06	11	32	34.4
Structural constituent of ribosome	3.72E-43	71	230	30.9
Protein transporter activity	2.10E-04	12	53	22.6
Structural molecule activity	3.34E-30	72	357	20.2
Component				
Mitochondrial outer membrane translocase complex	2.04E-05	6	8	75.0
Cytosolic large ribosomal subunit (sensu Eukaryota)	3.72E-25	37	97	38.1
Cytosolic small ribosomal subunit (sensu Eukaryota)	1.27E-15	24	64	37.5
Cytosolic ribosome (sensu Eukaryota)	1.02E-43	64	176	36.4
Cytosolic part	2.36E-41	65	197	33.0
Large ribosomal subunit	8.13E-26	44	142	31.0
Ribosomal subunit	4.18E-44	73	240	30.4
Small ribosomal subunit	7.37E-16	29	98	29.6
Ribosome	9.44E-38	80	357	22.4
Small nucleolar ribonucleoprotein complex	3.39E-07	23	132	17.4
Nucleolar part	4.92E-10	31	179	17.3
Ribonucleoprotein complex	1.10E-35	101	623	16.2
Unresponsive genes				980 genes
Process	**p-value**	**Gene hits**	**Term size**	**Term fraction**
80 biological processes	<1.00E-03			>10.0
Function				
Exoribonuclease activity, producing 5'-phosphomonoesters	6.60E-04	13	23	56.5
Exoribonuclease activity	6.60E-04	13	23	56.5
Guanyl-nucleotide exchange factor activity	7.40E-04	17	37	45.9
General RNA polymerase II transcription factor activity	6.19E-05	25	62	40.3
GTPase regulator activity	5.50E-04	27	77	35.1
RNA polymerase II transcription factor activity	4.90E-04	37	123	30.1
Transcription regulator activity	2.01E-09	90	327	27.5
Protein binding	1.40E-04	121	585	20.7
Component				
~40 cellular components	<1.00E-03			>10.0

Table 7: GO annotation of genes according to growth rate response.

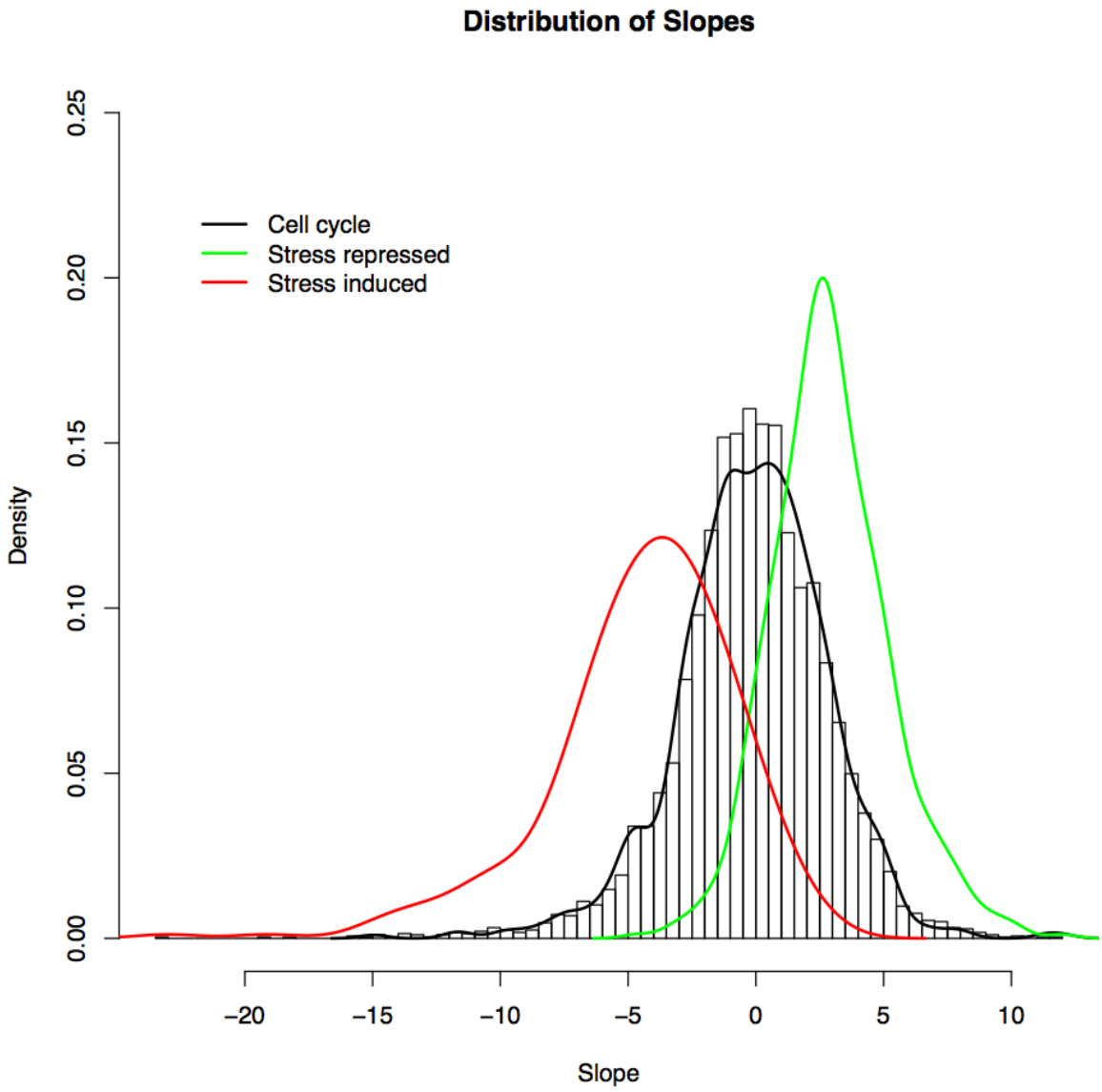


Figure 30: Transcriptional response of stress-related and cell cycle-related genes to changes in growth rate. Genes expressed periodically during the cell cycle (black line; (Spellman, Sherlock et al. 1998)) are distributed essentially as background, whereas genes induced (red line) or repressed (green line) by stress (Gasch, Spellman et al. 2000) tend to be conversely repressed or induced as growth rate increases.

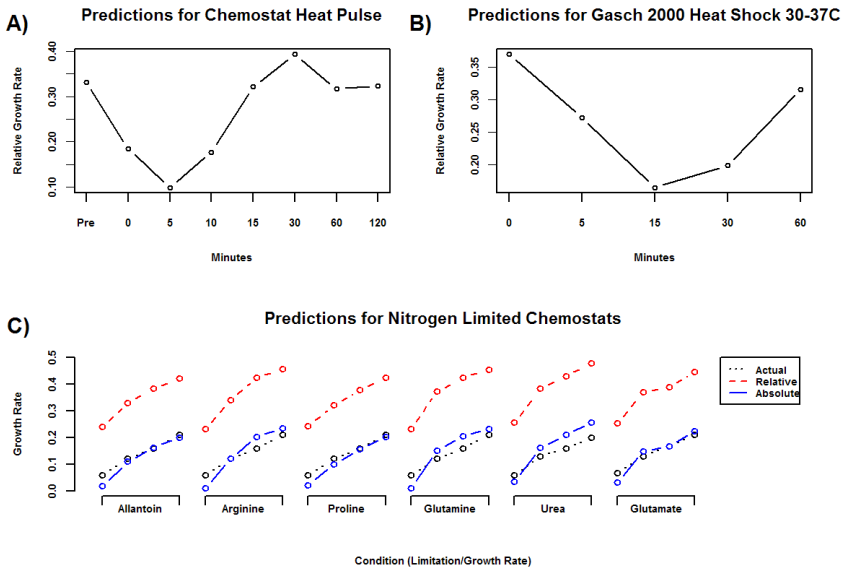


Figure 31: Predicted growth rates for *S. cerevisiae* gene expression data sets. Our model of the growth rate transcriptional response can be used to predict the growth rate of a cellular culture from gene expression data, robust to the originating biological conditions, growth regime, and experimental platform. Here, we apply the model to three selected data sets to infer relative and absolute growth rates. A) A brief (<30s) heat pulse was administered to a steady state chemostat culture immediately before time zero, and gene expression was assayed with a microarray time course (see Supplemental Figure 2). The relative growth rates inferred from this data show an abrupt departure from steady state growth, followed by a return to steady state (including a brief regulatory overshoot). Our predictions monitor these changes in growth rate at an instantaneous time scale (<5m) inaccessible by standard experimental assays for growth rate. B) Predicted growth rates for a portion of the (Gasch, Spellman et al. 2000) environmental stress response data, assaying the response to a 30–37C heat shock. Our model captures the cessation and resumption of growth induced by the stress, even for a batch culture in which the growth rate is not fixed a priori. C) A collection of 24 chemostats were run at four growth rates (0.05hr^{-1} through 0.2hr^{-1}) and limited on six different nitrogen sources. Using only expression data from each condition, our model predicts accurate relative growth rates. However, when provided with the known growth rate for a single condition, the model is additionally able to infer absolute growth rates for all other data sets sharing that condition's mRNA reference channel.

Absolute growth rate prediction with one shared reference

With a small amount of additional information, the relative growth rates inferred by our model can be made absolute in units of chemostat flow rate (hr^{-1}). Our model's predicted rates are typically relative values; this is due to the unknown quantitative effects of the reference mRNA in our dual-channel training data. It is impossible to know a priori the relationship between this reference channel and the relative (for dual-channel) or absolute (for single-channel) expression levels in new microarray data. However, if an absolute growth rate is known for some microarray condition in the data set, our model can make absolute rate predictions for other two-color data sharing the same reference channel.

Inferred absolute growth rates such as this can be seen in Figure 31C, which includes growth rate estimates for a collection of chemostats at various flow rates limited on one of several different nitrogen sources. On normalized dual-channel microarrays, the doubling of any gene's mRNA level in these conditions results in a constant increase in its expression readout. This allows one unit of our predicted relative rates to correspond to one unit of absolute chemostat flow rate. However, since the reference channel differs from that of our training data, all rate predictions are vertically shifted by a corresponding constant factor. By normalizing to any one of the 24 conditions' known growth rates, this shift can be automatically corrected for the 23 other microarrays employing the same reference channel.

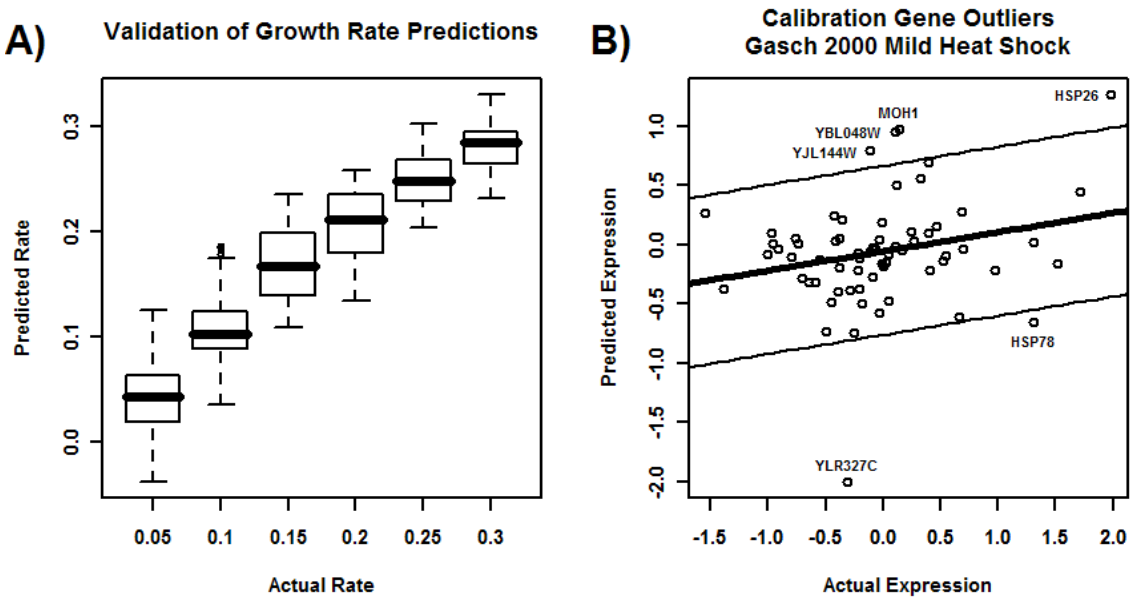


Figure 32: Assessment of accuracy and outlier detection during growth rate inference. A) We performed an out-of-sample cross-validation of our model by randomly subsampling 24 of the 36 training microarray conditions 1,000 times. We refit our linear model in each random sample, calculated bootstrapped null distributions for all gene parameters, and found sets of the most significant growth-specific genes. These were then used to infer growth rates for the 12 held-out conditions, providing an estimate of the accuracy of the model's growth rate predictions. B) When predicting the growth rate of a new collection of expression data, our model excludes any calibration gene with an expression level outside the inner fence (1.5 times the interquartile range below or above the first or third quartiles). This improves predicted growth rate accuracy while also calling out genes potentially responding to specific non-growth stimuli under some biological condition. For example, in the (Gasch, Spellman et al. 2000) mild heat shock time course, two of the six outliers are known heat shock genes (HSP26 and HSP78). The other four (YLR327C, MOH1, YBL048W, and TMA10) are uncharacterized genes, suggesting potential roles in the response to heat shock.

Robustness of the model and outlier detection

We assessed the quality of our growth rate predictions using 1,000 out-of-sample cross-validation experiments using the data from (Brauer, Huttenhower et al. 2008), as shown in Figure 32A. In

each experiment, we randomly withheld 12 of the 36 conditions for testing, fit our linear model on the remaining 24, derived bootstrapped null distributions using only these data, and determined calibration gene sets to use for growth rate inference on the held-out conditions. This out-of-sample validation allowed us to assess the accuracy and variability of our predictions on conditions with known growth rates not included in the model building procedure. In addition to the performance indicated by Figure 32A, this demonstrated robustness to the stringency of p-value cutoffs and number of growth-specific calibration genes; these ranged in number from ~50 to ~110 across the randomized validations (of a total ~5,500 possible genes), and changes of this magnitude in the final calibration gene set had little impact on predicted growth rates.

In the process of estimating growth rates and determining this confidence score, growth-specific calibration genes with outlying expression values are also detected. While most conditions induce few outlying growth-specific genes, when they occur, they are not indicative of the quality of growth rate predictions. We have found that neither the number of outliers nor their variability correlates with prediction error (data not shown), but they call out genes that may be responding to non-growth stimuli under specific biological conditions. Excluding outliers from the growth rate estimation process improves the accuracy of the predictions, and these outliers can in turn be biologically informative: an outlying calibration gene is likely responding specifically to a stimulus other than change in growth rate. For example, some of the only outliers in the mild heat shock time course from (Gasch, Spellman et al. 2000) occur towards the end of a shift from 29C to 33C (Figure 32B). These include HSP26 and HSP78, both known heat shock chaperones (Leonhardt, Fearson et al. 1993; Ferreira, de Andrade et al. 2006). Three genes of unknown function (YLR327C, MOH1 and the neighboring dubious ORF YBL048W, and TMA10) are also outliers in this condition, which is evidence that these genes may have specific

expression responses (and thus biological functions) during heat shock. HSP26 and YLR327C are frequent outliers in stress-related conditions, perhaps suggesting a more general stress response function.

Predicting growth rates in S. bayanus *and* S. pombe

While our growth rate model is based on a transcriptional growth signature in *S. cerevisiae*, the model can be applied to any organism with sufficiently orthologous transcriptional activity. This is likely to be the case within the *sensu stricto* yeasts, separated by ~25 million years of evolution (Gao and Innan 2004). By finding the ~50 *S. bayanus* genes orthologous to our ~70 *S. cerevisiae* growth-specific calibration genes (Kellis, Patterson et al. 2003), we can apply our model directly to *S. bayanus* expression data. Figure 33 demonstrates such a result for two *S. bayanus* time courses assaying the diauxic shift and a response to heat shock. These results have comparable profile to those from *S. cerevisiae* and are similarly biologically compelling. For example, the diauxic shift in *S. bayanus* results in a very similar growth pattern to the known response in *S. cerevisiae*, with a near-cessation of growth during the shift and subsequent rebound. Conversely, *S. bayanus* is less resistant to high temperatures than *S. cerevisiae* (Kishimoto and Goto 1995), and our growth rate inferences show a corresponding failure in its ability to grow following severe heat shock.

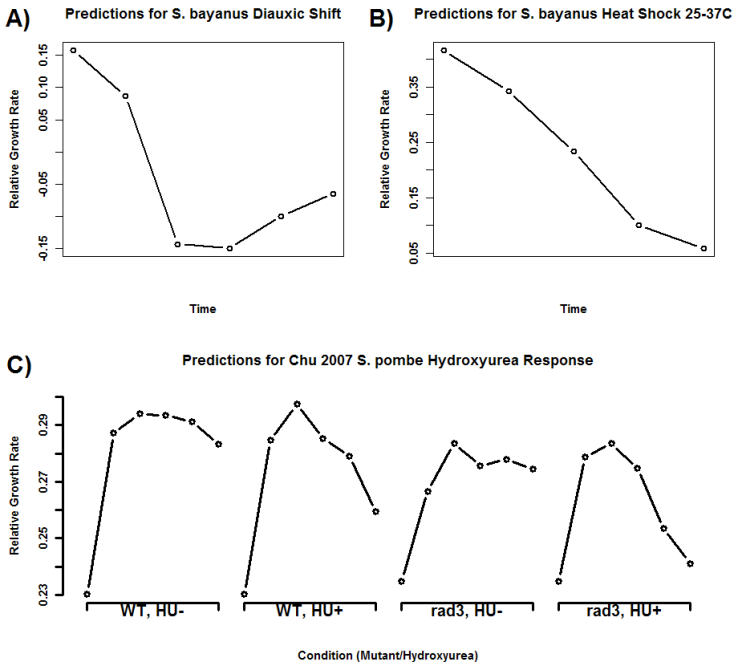


Figure 33: Predicted growth rates for *S. bayanus* and *S. pombe* expression data sets. By examining genes orthologous to our ~70 *S. cerevisiae* growth-specific calibration genes, we successfully applied our model to predict growth rates in *S. bayanus* (~50 orthologous growth-specific genes, ~20M years diverged) and *S. pombe* (~75 growth-specific genes due to one-to-many mappings, ~1B years diverged). A) Predicted growth rates for *S. bayanus* undergoing the diauxic shift from fermentative to respiratory growth. As observed for the *S. cerevisiae* diauxic shift in (Brauer, Huttenhower et al. 2008), growth pauses as glucose is exhausted and resumes as the yeast begins consuming ethanol. B) Predicted growth rates for *S. bayanus* exposed to a 25-37C heat shock. In contrast to Figure 31B, in which *S. cerevisiae* is observed to recover from a 37C heat shock, the less-thermotolerant *S. bayanus* (Kishimoto and Goto 1995) is predicted to halt growth at high temperatures. C) Predicted growth rates for *S. pombe* wild-type and *rad3*Δ time courses, grown normally and exposed to hydroxyurea (HU, an inhibitor of DNA synthesis and thus growth) (Chu, Li et al. 2007). Despite the wide evolutionary divergence between *S. pombe* and our *S. cerevisiae* training data, predicted growth rates are in substantial agreement with expected biology. Each time course begins with low growth in a synchronized culture. When the synchronization block is released, cells begin growing, wild-type more efficiently than the *rad3*Δ mutant. Exposure to HU decreases growth over time, and this effect is exacerbated by RAD3 deletion. While the *S. cerevisiae* RAD3 ortholog MEC1 is essential, knockouts of the MEC1 pathway members SOD1 and LYS7 have been previously observed to induce HU sensitivity (Carter, Kitchen et al. 2005).

We have also extended our model to a significantly further diverged yeast, specifically the yeast *Schizosaccharomyces pombe*, separated from *S. cerevisiae* by an estimated one billion years of evolution (Hedges 2002). A mapping of our growth-specific calibration genes to *S. pombe* using information from (Penkett, Morris et al. 2006) results in ~75 genes due to one-to-many correspondences, but these provide sufficient calibration information to make high quality predictions (Figure 33C). Calibration gene outliers and expression cohesiveness are not substantially changed relative to *S. cerevisiae* and *S. bayanus*, and the inferred relative rates reflect various biological expectations. All cultures (data from (Chu, Li et al. 2007)) show an initial increase from low growth rates due to stalled growth during synchronization. An expected decrease in growth rate is predicted during increased exposure to hydroxyurea (HU), and a *rad3*Δ deletion (*S. cerevisiae* ortholog MEC1) incurs a mild overall growth impairment as well as exacerbating HU sensitivity. While MEC1 is essential in *S. cerevisiae*, this sensitivity has previously been noted for deletions *sod1*Δ and *lys7*Δ, both members of the MEC1 pathway (Carter, Kitchen et al. 2005), which is necessary for the cell cycle checkpoint function.

The extent to which transcriptional regulation is conserved between *S. cerevisiae* and *S. pombe*, which allows us to successfully apply the model despite the evolutionary distance that separates these species, is reflective of cellular growth's central role, particularly in unicellular organisms. While this model would become less meaningful in metazoans, where the growth of individual cells is subjugated to the growth and differentiation of the organism as a whole, certain transcriptional growth behavior is of necessity conserved in single celled organisms (Rudra and Warner 2004). This is particularly true of the ribosome, one of the main contributors to our model's predictive power; rRNA regulation is purely transcriptional, and ribosomal proteins must be expressed stoichiometrically. Since any cellular growth requires translation, observation

of ribosomal transcription is a strong indicator of unicellular growth (Warner 1999). This is one aspect of the transcriptional growth response made quantitative by our model.

Insights into growth homeostasis

To further investigate the biological basis of growth rate correlated gene expression, we used our model to predict relative growth rates for two interesting cases: the yeast metabolic cycle (Tu, Kudlicki et al. 2005) and the mitotic cell division cycle (Spellman, Sherlock et al. 1998; Pramila, Wu et al. 2006). The microarray data published by Tu et al. was obtained for cells grown at high density in a glucose-limited chemostat. Under this regime, cells within the culture become metabolically synchronized and undergo periodic consumption of oxygen (defined as the oxidative phase of the metabolic cycle) followed by periods of undetectable oxygen consumption (termed the reductive building and reductive charging phases). The cell cycle data sets by Spellman et al and Pramila et al were obtained from experiments in which cells were arrested in growth using a variety of methods and then released from arrest to undergo the cell division cycle as a synchronous population.

Growth rate prediction applied to the yeast metabolic cycle data revealed a striking periodicity (Figure 34A). The cyclical pattern of growth rate variation occurs completely in concert with the metabolic cycle as defined by Tu et al. Specifically, the culture's growth rate is predicted to be at minima during the reductive phases of the metabolic cycle and reach maxima during the peak of the oxidative phases. In contrast, growth rate prediction for the cell cycle (Figure 34B and C) show virtually no variation in predicted growth rate during the different stages of cell division.

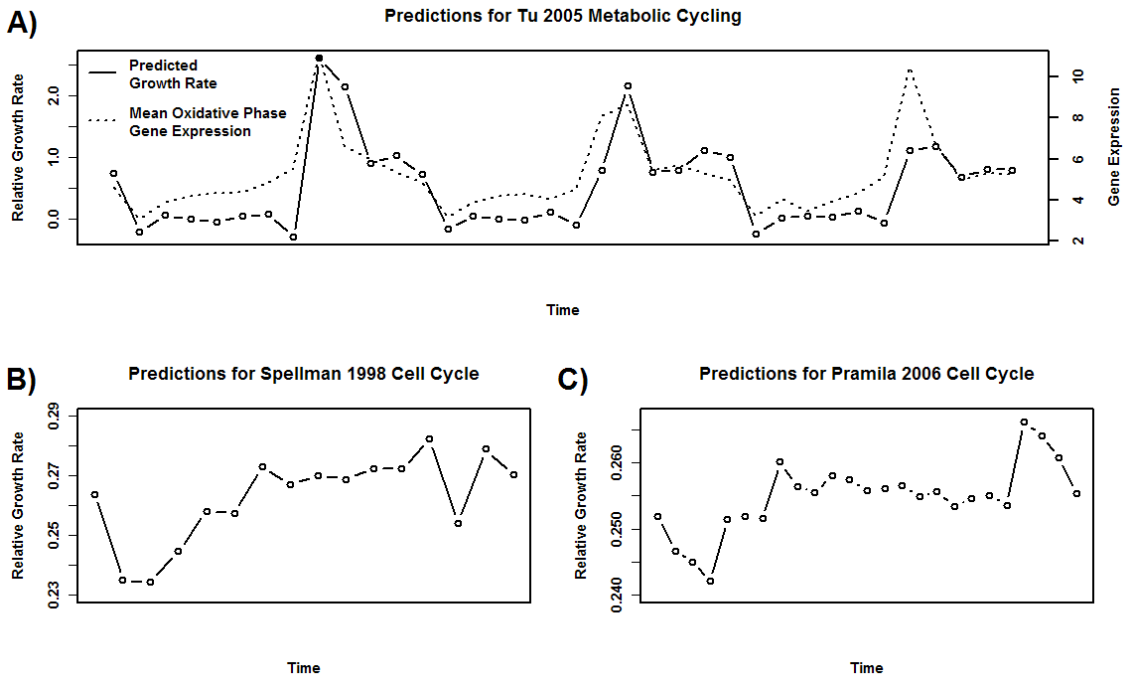


Figure 34: Differences in growth characteristics of a metabolically cycling culture compared to cells synchronously undergoing the cell division cycle. We predict periodic bursts of growth during the oxidative phase of the metabolic cycle as described by (Tu, Kudlicki et al. 2005). Conversely, we observe essentially no variation in growth in cultures synchronously undergoing the cell division cycle, which has been shown to primarily occupy the reductive phase of the metabolic cycle (Chen, Odstrcil et al. 2007). A) In cells undergoing metabolic cycling, growth rates are predicted to peak during the oxidative phase of the cycle, where (Tu, Kudlicki et al. 2005) also observes strong upregulation of translational and ribosomal genes. B) The predicted growth rate for the (Spellman, Sherlock et al. 1998) alpha-factor synchronized cell cycle is essentially constant, after an initial release from the synchronization block. C) Predicted rates for the (Pramila, Wu et al. 2006) alpha-factor synchronized cell cycle also show an initial resumption of growth after alpha-factor block followed by relatively constant growth rate. Taken together, these observations support the claim that growth rate regulation is not specific to any one cell cycle phase. This also agrees with the fact that rapidly growing (and thus fermenting) *S. cerevisiae* does not partition metabolism into discrete stages, a phenomenon only occurring when reductive metabolism is hindered by nutrient limitation or other stresses.

These data support and extend our previous assertions (Brauer, Huttenhower et al. 2008) that the there is a close connection between the metabolic cycle identified in (Klevecz, Bolen et al. 2004) and (Tu, Kudlicki et al. 2005) and the association we identify between growth rate and gene expression levels. This result is consistent with two possible explanations. The first is that there is variation in the growth rate of cells throughout the metabolic cycle. (Tu, Kudlicki et al. 2005) and (Chen, Odstrcil et al. 2007) have shown that under their specific experimental conditions, DNA replication and cell division is restricted to the reductive phases of the metabolic cycle. It is conceivable that growth per se (i.e. the accumulation of biomass) is paused during the reductive phases of the metabolic cycle so that the cell can replicate and segregate DNA and complete the complex processes of cell division; growth may then be restricted to the oxidative phase of the metabolic cycle. Alternatively, it is possible that as any heterogeneous culture grows faster, a greater fraction of cells are in the oxidative phase at any point in time. Thus, the growth rate gene expression signature we observe might reflect the increasing fraction of cells in the oxidative phase of the metabolic cycle.

The absence of growth rate differences during the cell division cycle (Figure 34B and C) supports our previous claim (Brauer, Huttenhower et al. 2008) that the growth rate expression signature is unrelated to the cell cycle. Moreover, since the relevant cell cycle experiments were performed in rich media using a fermentable carbon source, the results suggest that rapidly growing cells (which are almost exclusively fermenting) do not partition metabolic activity into discrete phases, as their energetic requirements are met in a continuously reductive metabolic state. It is only when slowed growth is imposed upon the cell, due to stress, nutrient limitation, or other suboptimal environments, that the metabolic cycle is required.

We sought to further distinguish whether nutrient availability determines growth rate (which in turn determines the gene expression pattern) or whether nutrient availability sets the transcriptional state (which in turn determines growth rate). To address this issue, we examined the regulatory circuit responsible for transcriptional changes in response to glucose availability in yeast. Glucose addition to cells growing on glycerol elicits a rapid and massive change in the pattern of gene expression, with more than half of all genes changing at least twofold in expression. Previous work has shown that the Ras/cAMP/PKA pathway is the major source for eliciting this transcriptional change in response to glucose addition (Wang, Pierce et al. 2004; Zaman, Lippman et al. 2008). Activation of the Ras/PKA pathway in the absence of environmental signals, through induction of an activated allele of RAS2 ($RAS2^{G19V}$), recapitulates in magnitude and direction more than 85% of the changes observed by glucose addition, and inhibition of PKA (concurrent with addition of glucose) blocks most of the glucose induced transcriptional changes (Zaman, Lippman et al. 2008). This mutation thereby represents a useful model connecting *S. cerevisiae*'s glucose sensory signaling to its transcriptional regulation of growth rate.

We constructed a *gal1*Δ strain carrying the activated allele $RAS2^{G19V}$ under control of the galactose inducible GAL10 promoter. Addition of galactose activates the Ras/PKA pathway, but since galactose cannot be metabolized by this strain, the metabolic state of the cell remains unaltered (Wang, Pierce et al. 2004). When grown on glycerol we predict a relative growth rate of ~0.2 for this strain (Figure 35A), which changes to ~0.6 within twenty minutes following glucose addition, consistent with the change in doubling time from 5.8hr to 2.6hr. When we performed the same experiment on glycerol media and induced the $RAS2^{G19V}$ by means of galactose addition, we detected a transcriptional response within sixty minutes. The predicted growth rate of $RAS2^{G19V}$

mutant strain was comparable to the addition of glucose despite the fact that galactose addition does not yield an increase in growth, as measured by optical density, since the cells are unable to metabolize galactose. In fact, while the model's summarization of gene expression state indicates that the culture is attempting to increase growth, induction of the $RAS2^{G19V}$ allele results in an immediate decrease in growth rate and complete cessation of growth within four hours (Fedor-Chaiken, Deschenes et al. 1990). These results are consistent with the cell setting its growth-specific transcription program on the basis of its perception of nutrients present in the environment, rather than on the direct availability of energy or metabolites produced from such nutrients. The mechanism by which the cell integrates this external state in order to set the appropriate growth rate expression state must be mediated, at least in part, through the Ras/cAMP/PKA pathway.

Potential transcriptional regulators of growth rate

To investigate the regulatory basis of growth-associated gene expression, we identified motifs enriched in the 3' and 5' regions of genes with strong growth rate responses (Figure 35B). We assigned genes to clusters based on their growth rate response parameter (β_g) using k-means clustering with k=10. Using the FIRE motif identification program (Elemento, Slonim et al. 2007), we identified enriched motifs in seven of the resulting ten clusters. Consistent with the functional enrichments of negatively growth rate correlated genes (Brauer, Huttenhower et al. 2008), we identified known binding sites associated with the stress responsive transcription factors Msn2p and Msn4p in genes negatively correlated with growth rate. Conversely, genes that are increased in expression with growth rate are enriched for the Rap1p consensus motif, which is commonly found upstream of genes encoding protein components of the ribosome.

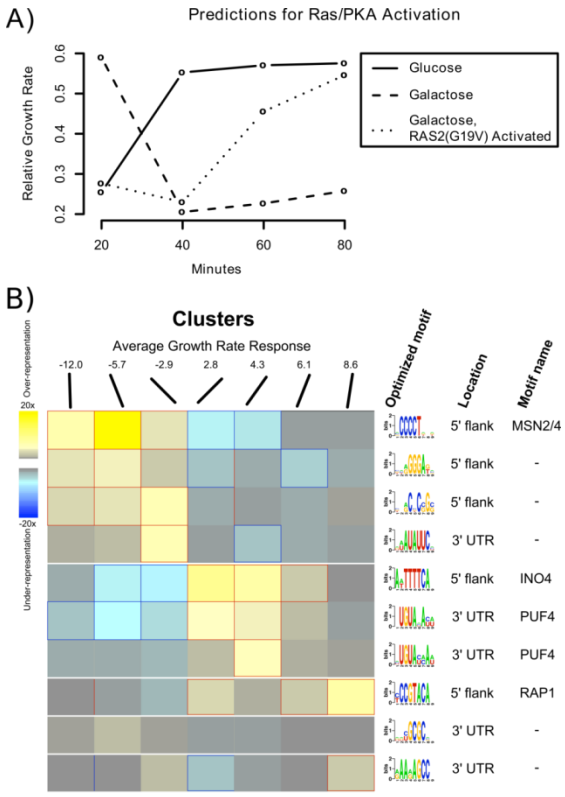


Figure 35: Perturbations and potential transcriptional regulators of the growth rate response. A) Predicted growth rates for *gal1*Δ cells shifted to glucose, to galactose, and to galactose with a constitutively active RAS2^{G19V} allele. On glucose, rapid growth is induced within ~40m; growth on galactose falls to low levels within ~40m, as it cannot be metabolized by this mutant. However, when glucose sensing is emulated by artificial activation of the Ras/PKA pathway, the transcriptional regulatory network attempts to induce rapid growth within ~60-80m despite the unavailability of appropriate nutrients. This disconnect between actual and perceived cellular state leads to cell death within 4-6 hours and suggests that nutrient sensing (as opposed to metabolic activity or internal cellular state) is responsible for a large portion of the transcriptional growth rate response. B) Regulatory binding sites enriched in growth up- and down-regulated genes. We clustered the yeast genome by degree of growth rate response, yielding ten clusters with average responses ranging from -12.0 (strongly downregulated with increasing growth rate) to 8.6 (strongly upregulated). The FIRE program (Elemento, Slonim et al. 2007) predicted 10 regulatory motifs in the upstream flanks and 3' UTRs of the most up- and down-regulated clusters. These included the known stress-responsive MSN2/4 binding sites in downregulated genes, the ribosomal regulators RAP1 and PUF4 in upregulated genes, and INO4 sites in upregulated genes (possibly corresponding to its role in the stress response and fatty acid biosynthesis (Santiago and Mamoun 2003)). We also identified five additional putative growth regulatory sites for which the binding factor is not yet known.

We also found enrichment of the Ino4p binding site in genes upregulated with increasing growth rate. Ino4p forms a heterodimer with Ino2p to activate genes involved in phospholipid, fatty acid, and sterol biogenesis, all of which are required in greater abundance with increased growth rates. Furthermore, Ino4p has been proposed to have an inhibitory effect on a number of genes, including those that encode the heat shock proteins (Hsp12p, Hsp26p) and catalase (Ctt1p) (Santiago and Mamoun 2003). We also identified two additional enriched motifs in the 5' UTR for which the binding factor is not known, suggesting additional activators of growth-related transcriptional programs.

In addition to 5' upstream motifs, we identified five enriched 3' UTR motifs, which are potential binding site for proteins that promote mRNA degradation. Only a small number of mRNA binding consensus sequences are known, all of which belong to the Puf family of mRNA binding proteins (Gerber, Herschlag et al. 2004). Our analysis identified five enriched motifs in 3' UTRs. Two of these motifs, found in genes positively correlated with growth rate, were identified by the FIRE program as being targets of Puf4p. As an independent test, we compared the distribution of growth rate responses in the known gene targets of the five Puf proteins with the overall distribution of growth rate slopes. Targets of both Puf3p (220 genes) and Puf4p (205 genes) are enriched for genes that are upregulated with increasing growth (Wilcoxon-Mann-Whitney two sample p-values $9x10^{-23}$ and $7.23x10^{-16}$, respectively). The consensus motifs of Puf3p and Puf4p are very similar; investigation of the PUF4 motif identified by FIRE suggests that the enrichment signal for at least one of the motifs denoted PUF4 is likely to result from a composite of Puf3p and Puf4p target genes (Figure 35B).

Overall, this analysis is consistent with tight transcriptional regulation underlying the cellular growth program; it is likely that mRNAs involved in this process are also subject to extensive

post-transcriptional control. Interestingly, since our method is sensitive to changes in gene expression levels occurring in just a few minutes, we expect that post-transcriptional regulation (both mediated decay of and stabilization of transcripts) is involved in this response. Experimental analyses of the effects of perturbations within this regulatory network promise to shed further light on its organization.

Discussion

We studied 36 yeast chemostat cultures growing at six different growth rates under six different nutrient limitations: glucose, sulfate, phosphate, ammonium, leucine (in a non-reverting *leu2* mutant) and uracil (in a non-reverting *ura3* mutant). By using a variety of different nutrients to limit growth rate, we could focus on quantitative relationships with growth rate per se, and not with the particular nutrient regime that limits the growth rate. Our data agree very well with the results of others who have done similar studies (Boer, de Winde et al. 2003; Saldanha, Brauer et al. 2004; Regenberg, Grotkjaer et al. 2006; Castrillo, Zeef et al. 2007), both with respect to genes that are responsive to particular limitations and with respect to genes that respond mainly to growth rate.

We present a statistical model of the gene expression response to changes in growth rate in *S. cerevisiae*. Based on microarray data from a variety of steady state growth rates and nutrient limitations, this model captures descriptive information regarding each gene's consistency and degree of response to growth rate. As detailed in (Brauer, Huttenhower et al. 2008), approximately half of the genome shows a significant transcriptional response to growth rate with strong functional cohesiveness; here, we extend this model to show its robustness, applicability to new data, and ability to provide insight into the biological systems driving cellular regulation of growth rate. New experiments with more complex models (quadratic and

hierarchical) demonstrated that additional model parameters did not provide substantial performance gains, particularly relative to their added complexity (data not shown). Similarly, changes in the stringency of definitions of responding genes or of growth-specific genes did not substantially alter results. This is reflected in the out-of-sample validation results, which quantify the model's accuracy in predicting relative growth rates from gene expression data.

Expression of about one-quarter of all yeast genes is correlated with growth rate, but the magnitudes of the slope of the relationship is characteristic for each gene

We identified a large number of genes (ca. 27% of all yeast genes) each of whose expression is linearly correlated (either negatively or positively) with the growth rate, independent of the limiting nutrient. Some of these genes were much more strongly affected by growth rate than others, again independent of the identity of the limiting nutrient. Hierarchical clustering of the entire chemostat dataset indicates that the correlation between the steady-state level of mRNA and the nominal growth rate applies to many genes. Notably, (Castrillo, Zeef et al. 2007) recently published a set of data entirely consistent with the one presented here for glucose, nitrogen, sulfur and phosphate limitations. Their analysis also led to the conclusion that many genes are expressed in a way correlated with growth rate, independent of the identity of limiting nutrient.

The combination of each gene's growth rate slope (i.e. strength of transcriptional response) and the bootstrap p-values of these slopes (i.e. their statistical significance) allows the rigorous identification of genes responding strongly to growth rate in a nutrient-independent manner. A histogram of the slopes for all yeast genes allows one to visualize the growth rate sensitivity of a single gene or a list of genes relative to the overall distribution (Figure 29). These methods (also available on the website http://growthrate.princeton.edu) facilitate the use of the information

from this dataset to make inferences for our own analysis, and, as we shall see, to analyze the results of others.

In order to focus on the biology of gene expression as a function of growth rate, we defined a subset of 1,608 genes that correlate with a characteristic slope: 337 had a negative slope, 291 a positive slope, and 980 a slope near zero (i.e. their expression was roughly the same at all growth rates). The point to be emphasized here is that this stringent a selection of 337 + 291 = 628 (i.e. about 10% of all yeast genes) necessarily underestimates the number of genes with non-zero slopes, since genes with smaller (positive or negative) slopes and/or noisy data are likely to fail the statistical tests. The clustering estimates (which suggest larger numbers of growth rate responsive genes) are surely closer to reality; we therefore suggest that expression of at least 27% of yeast genes is correlated with the nominal growth rate in chemostats, regardless of the nature of the limiting nutrient.

Functional roles of genes whose expression is most strongly related to growth rate

GO Term Finder analysis of the subsets of genes with well-defined slopes (Table 7) presents a very clear picture. The positive-slope subset of 291 genes focuses on the translation machinery, both cytosolic and mitochondrial. This result has very strong precedents in the literature of both bacterial and yeast physiology, where the correlation between the number of ribosomes and the growth rate was noted very early (Maaloe and Kjeldgaard 1966); see more recent reviews (Nomura 1999; Warner 1999; Zhao, Sohn et al. 2003). The biological logic for this relationship is virtually self-evident: to grow at a faster rate, more proteins must be made per unit time, which is facilitated by having more ribosomes per cell.

Many different regulatory mechanisms (most prominent among them the TOR1 signaling system) have been implicated in this connection (reviewed in (De Virgilio and Loewith 2006)). Of particular relevance are the results of Jorgensen et al. (Jorgensen, Rupes et al. 2004), who found a connection between ribosome biosynthesis and cell cycle entry at START (Hartwell, Culotti et al. 1974) via the regulation of both processes by the products of SFP1 and SCH9.

The negative-slope subset of 337 genes relates to functions associated with oxidative energy metabolism, especially those carried out in peroxisomes. Peroxisomes have been associated with fatty acid metabolism, with oxygen metabolism (particularly reactive oxygen) and, more recently, with autophagy (see reviews (Kim and Klionsky 2000; van Roermund, Waterham et al. 2003; Monastyrska and Klionsky 2006; Rottensteiner and Theodoulou 2006; Wanders and Waterham 2006)). The biological logic here is less obvious, although the benefits of engaging in autophagy and degradation of cellular materials during nutrient limitation are clear. Another possibility relates to the metabolism of reactive oxygen species, for which there might be more need when time between cell division cycles is longer. A purely metabolic logic (e.g. a need for more beta-oxidation of fatty acids at slow growth rates) is more difficult to rationalize. While reasonable when carbon (in our case glucose) is limiting, it is not obvious how this might work for the other limitations, especially those which leave high concentrations of residual glucose in the medium at steady-state.

The statistically derived growth-rate-independent subset (980 genes) is, in this context, equally informative. It includes a large number of GO terms that cover much of the remaining yeast cell biology. These 84 GO process terms notably include such areas as transcription, DNA metabolism, chromatin remodeling, proteolysis, protein secretion and even the cell cycle, among many others.

Instantaneous growth rate and the Environmental Stress Response

Among the positively correlated genes, we found many (but not all) of the genes whose expression declined during the "environmental stress response" as defined by (Gasch, Spellman et al. 2000); among the negatively correlated, we found many (but not all) of the genes whose expression increased in the Gasch experiments. Similar data were reported recently in (Castrillo, Zeef et al. 2007). These results raise the possibility that much (but probably not all) of what has been defined as environmental stress response might equally well be thought of as a general response to changes in the instantaneous growth rate. Since it consists mainly of the most growth-rate-sensitive genes, much of the response could be secondary to a much smaller number of specific responses to individual environmental stresses. It also is worth noting here that the steady state mRNA concentrations of the positively growth-rate-correlated genes fall remarkably rapidly after applications of stresses (Gasch, Spellman et al. 2000), consistent with the idea of regulation at the level of mRNA degradation (Grigull, Mnaimneh et al. 2004) as well as transcription.

Predicting instantaneous growth rates under novel experimental conditions

Our model can be applied to new gene expression data to infer the relative instantaneous growth rate of the originating cellular culture. This instantaneous rate represents a measurement of the transcriptional state of cellular growth rate control, and it provides insight into the cell's growth rate at arbitrarily short time scales inaccessible by experimental measurements (e.g. optical density). Moreover, genes with outlying expression values can be detected during growth rate inference, calling out probable biological responses to specific non-growth stimuli. The predictions based on this model are robust to changing biological conditions, experimental methods, and technological platforms; they also scale from our *S. cerevisiae* training data to the

related yeast *S. bayanus* and the highly diverged yeast *S. pombe*, suggesting that the transcriptional control of growth rate captured by the model are a fundamental aspect of unicellular biology.

Through further analysis of this regulatory network, we discovered several potential transcription factor binding sites enriched in growth-correlated genes, most notably the stress-responsive Msn2p and Msn4p, the Rap1p ribosomal factor, and Ino4p. Importantly, we have identified a likely role for post-transcriptional regulation in modulating transcriptional states related to growth rates. This finding is consistent with our ability to measure changes in growth rate over very short time scales using gene expression signatures. The abundance of any messenger RNA is a function of both its rate of production and of its rate of degradation; however, since transcription can be relatively slow, changes in abundance can be most rapidly effected by altering the stability of the extant mRNA population. The Puf proteins have known roles in mediating mRNA degradation (Olivas and Parker 2000) and in mediating the association of functionally related transcripts (Gerber, Herschlag et al. 2004). It has recently been proposed that modulation of mRNA stability is an important factor in metabolic regulation (Palumbo, Farina et al. 2008). The association of Puf protein binding domains in the 3' UTRs of genes with increased expression at higher growth rates suggests that modulating mRNA stability is also important in the regulation of the growth response at short time scales.

From a computational perspective, it is notable that a simple linear model accurately and robustly captures a specific biological phenomenon. The model represents a concise, functionally cohesive set of expression profiles regarding the genome's transcriptional response to growth. This description agrees with known aspects of the growth response, such as the transcription of ribosomal components, and provides initial data as to the mechanistic roles of internal feedback,

environmental sensing, and the stress response as growth rate varies. By monitoring an ensemble of genes - but with few parameters per individual gene - the model is easily applicable to new conditions and organisms and is robust to technical and biological variability. These features enable our model to serve both as a practical tool for growth rate estimation (available at http://function.princeton.edu/growthrate) and as a mechanistic building block in the pursuit of a systems-level understanding of cellular growth processes.

Effects of Aneuploidy on Gene Expression in *S. cerevisiae*

The gain or loss of entire chromosomes, referred to as aneuploidy, has been a known hallmark of cancer cells and other genetic disorders (particularly Down syndrome) for decades. Aneuploid cells can arise from a variety of mechanisms that have been intensely studied. In cancer cells, aneuploidies are thought to be caused mainly by the failure of mitotic checkpoints; meiotic aneuploidies, for which the root causes are less well understood, manifest primarily in spindle alignment failures. Once present, it is thought that aneuploidies can be selected for as beneficial traits e.g. in a tumorigenic environment, where they can activate oncogenes or deactivate tumor suppressors.

Beyond these general effects, however, it is not clear how a cell's internal regulatory network adapts to the massive imbalances imposed by aneuploidies. Without a moderating regulatory response, the presence of an extra chromosome would be expected to incur a correspondingly greater transcription, translation, and protein load on the cell; a missing chromosome would similarly limit a cell's ability to create specific proteins. Coupled with the fact that recent work has demonstrated aneuploidy to confer a general growth disadvantage (Torres, Sokolsky et al.

2007), it is thus of great interest to determine how a cell's regulatory network accommodates aneuploidy.

Saccharomyces cerevisiae presents a unique opportunity to study aneuploidies for several reasons, not least of which is its well-known amenability to genetic manipulation. Since yeast can grow naturally in both haploid and diploid states, this allows the examination of artificially monosomic (i.e. diploids missing a copy of one or more chromosomes), disomic (i.e. haploids with an extra copy of some chromosome), and trisomic (i.e. diploids with one or more extra chromosomes) cells. *S. cerevisiae* can also stably maintain yeast artificial chromosomes (YACs) carrying inactive, non-transcribed DNA. Finally, yeast can be easily cultured either under conditions of unlimited exponential growth (batch cultures) or at a controlled rate limited by some specific nutrient (continuous or chemostat cultures). The diversity of aneuploidy and growth conditions attainable by *S. cerevisiae* cultures, in combination with whole-genome techniques for monitoring gene expression, allow us to study both the transcriptional effects of specific types of aneuploidies and the global regulatory response to aneuploidy.

Here, we use a linear model to statistically determine the changes in *S. cerevisiae* gene expression attributable to specific aneuploidy conditions and to a global regulatory response to aneuploidy. By measuring the gene expression of cultures with four types of aneuploidy (monosomy, disomy, or trisomy for one or more chromosomes, or carrying a YAC) in three different growth environments (batch culture, chemostat cultures limited for phosphate, or chemostat cultures limited for uracil), we have the opportunity to decompose the resulting transcriptional programs into portions statistically attributable to each stimulus. We find that i) there is a broad, functionally diffuse regulatory response to all aneuploidies (even carrying an inactive YAC), ii) carrying an active aneuploidy upregulates ubiquitin-mediated protein degradation by the

proteasome, iii) the previously reported growth impairment by aneuploidy incurs a separate, largely unrelated regulatory response, and iv) trisomy specifically upregulates proteins involved in mitosis and downregulates mating signals. Additional regulatory perturbations were incurred by aneuploidies of individual chromosomes, e.g. disruption of chromosome XII strongly inhibits respiration and oxidative phosphorylation. Taken together, these results provide not only detailed insights into *S. cerevisiae*'s response to chromosomal abnormalities, but also form a starting point for investigating the broader impact of aneuploidy on gene regulation.

Methods

We applied a linear model to the study of gene expression in *S. cerevisiae* in response to a variety of induced aneuploidies (Table 8). Specifically, yeast mutants from a W303 background were constructed carrying one of four types of chromosomal aberrations: monosomes (diploids with only a single copy of some chromosome), disomes (haploids with two copies of some chromosome), trisomes (diploids with three copies of some chromosome), and YACs (yeast carrying an artificial chromosome with human or mouse DNA inserted). These mutants were then grown under one of three conditions: batch cultures, phosphate limited chemostats, or uracil limited chemostats. In each perturbation/environment combination, gene expression levels were assayed using microarrays with wild type yeast (grown in the same environment) as a reference; for additional biological details, see (Torres, Sokolsky et al. 2007). Here, we model each gene's expression as a linear combination of its baseline chromosomal copy number, the type of induced chromosomal aberration, the growth environment, and perturbations of individual chromosomes. This allows us to discover genes responsive to specific chromosomal abnormalities and a program of global gene regulation in response to aneuploidy.

Collection of gene expression data

Briefly, aneuploid yeast strains were generated by a chromosome transfer technique based on (Hugerat, Spencer et al. 1994). Disomic mutants were obtained by mating strains with integrated *HIS3* cassettes in each chromosome with a *kar1*Δ15 *cyh2*^{Q37E} strain and selected on cyclohexamide. The resulting progeny were in turn mated with a strain carrying the *kanMX6* cassette at the same locus as the *HIS3* integration, as well as the *can1*-100 allele. Progeny were selected on G418 media lacking histidine and containing canavinine. Trisomic strains were generated by replacing the *kanMX6* marker with *URA3*, mating the resulting diploids to a haploid strain containing *kanMX6* at the same location, and selecting on -His-Ura+G418 media. Yeast carrying YACs were generated in a similar manner by plating on media lacking uracil. For additional details, see (Torres, Sokolsky et al. 2007).

This procedure resulted in yeast with monosomic, disomic, trisomic, and YAC aneuploidies as specified in Table 8. These strains were then grown under one of three conditions: batch cultures grown to an OD600 of one in -His+G418 media, chemostat cultures limited for phosphate grown to steady state at 0.17hr^{-1}, or chemostats limited for uracil at 0.17hr^{-1}. Gene expression in the resulting cultures was assayed using Agilent yeast arrays using wild type yeast from identical strain backgrounds (but without aneuploidies) grown under the same conditions; biological and technical replicates were included when possible. Since standard microarray normalization assumes an expected change of zero between reference and test channels, which will not be true under whole-chromosome aneuploidies, the average log ratio of all non-aneuploid genes was subtracted from the log ratio expression value for all genes on each array. This resulted in global average $\log_2$ expression values of 0.76 (1.7 fold) for disomic genes, 0.47 (1.4 fold) for trisomes, -0.71 (0.61 fold) for monosomes, and -0.01 (0.99 fold) over all genes in YAC bearing strains.

	YACs	Monosomes	Disomes	Trisomes
Batch	YAC	I, I+IX	I, II, IV, V, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI, VIII+XIV, XI+XV, XI+XVI, I+YAC	-
Phosphate chemostats	YAC	V, VI, I+IX	I, II, IV, V, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI, XI+XVI, II+YAC	I, II, IV, V, VIII, IX, XI, XII, XIV, XV, XVI
Uracil chemostats	-	-	I, II, IV, V, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI, XI+XV, XI+XVI	-

Table 8: Aneuploid strains for which gene expression data was collected and analyzed. YAC carrying strains were either otherwise wild type or, in two cases (I+YAC and II+YAC), disomic for one yeast chromosome. Monosomic strains were diploid missing one copy of one chromosome (or two), disomic strains were haploid carrying an extra copy of one chromosome (or two), and trisomic strains were diploid carrying a third copy of one chromosome. When no data is present for a particular aneuploid/growth condition combination, those strains were not assayed under the specified condition; when a chromosome is not indicated in the list of specific aneuploidies, the necessary yeast strain was inviable or otherwise not obtained from the chromosomal transfer assay.

Linear model of aneuploidy response

The independent variables perturbed by this assay can be summarized by 22 binary values, each zero when the condition is not present and one when it is: five types of aneuploidy (YACs, monosomy, disomy, trisomy, and transcription, the last including all non-YAC conditions), three growth conditions (batch, phosphate limitation, and uracil limitation), and 14 possible chromosomal aberrations (I, II, IV, V, VI, VIII, IX, XI, XII, XIII, XIV, XV, and XVI, excluding the unobtainable III, VII, and X). Additionally, data was available to analyze six of the possible interaction conditions: monosomic, disomic, and transcriptional aneuploidies with batch and phosphate growth conditions. Finally, a gene's expression level was assumed to be directly dependent on its DNA copy number relative to wild type (0.5, 1.0, or 1.5) and to be transcribed at some baseline level (i.e. intercept or constant term). For each condition, these constants can be

summarized as a vector of constants $\boldsymbol{x}_c$; for each gene, the corresponding coefficients were represented as a vector $\boldsymbol{b}$:

$$\boldsymbol{b} = [\ \textit{base}, \textit{dna},$$
$$\textit{yac}, \textit{mono}, \textit{diso}, \textit{tris}, \textit{trans},$$
$$\textit{batch}, \textit{pho}, \textit{ura},$$
$$\textit{mono}{:}\textit{batch}, \textit{mono}{:}\textit{pho}, \textit{diso}{:}\textit{batch}, \textit{diso}{:}\textit{pho}, \textit{trans}{:}\textit{batch}, \textit{trans}{:}\textit{pho},$$
$$\textit{chrom}_{\mathrm{I}}, \textit{chrom}_{\mathrm{II}}, \ldots, \textit{chrom}_{\mathrm{XVI}}\ ]$$

Accounting for the variable availability of specific aneuploid/condition combinations and for missing microarray data, we obtained between 76 and 108 expression values for each gene. These dependent values $\boldsymbol{y}$ were then fit using least squares to a linear model:

$$\boldsymbol{y} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]\boldsymbol{b} + \varepsilon$$

That is, for each gene, we fit (at most; see below) 30 parameters, corresponding to the aneuploidies, environments, interactions, and chromosomal aberrations as described above, and to baseline (intercept) and copy number (DNA) parameters. A gene's expression values $\boldsymbol{y}$ are each fit using these coefficients $\boldsymbol{b}$ and a condition-specific vector of constants $\boldsymbol{x}_c$ (all binary save for the relative DNA content) that describe the combination of aneuploidies and environment making up the condition.

To avoid fitting noise with this sizeable collection of parameters, we employed a modification of the method described in (Knijnenburg, Wessels et al. 2008). When fitting each gene's linear model using R (R Core Development Team, Vienna, Austria), we began with an empty parameter set (i.e. $\boldsymbol{y} = \varepsilon$). Then, for each parameter, we used a bootstrap evaluation across the gene's expression values to determine which single parameter would, if added, provide the greatest reduction in

error. If this parameter was also significantly different from zero (at the 0.01 level Bonferroni corrected for the 30 possible parameters), it was added to the vector $\boldsymbol{b}$ and the process was repeated, halting at (and excluding) the first parameter not significantly nonzero. Interaction terms were only used if they reduced error more than the addition of both of their constituent individual terms. Unlike (Knijnenburg, Wessels et al. 2008), due to our much smaller biological effect sizes, no restriction was placed on the minimum value of each parameter. This resulted in zero to eight parameters being learned per gene, the former due to missing values in the microarray data and the latter extremely rare (only five genes using more than six parameters). This model fit significantly (Bonferroni-corrected $p<0.01$, minimum adjusted $R^2\approx0.2$) to the majority of the genome (4,997 genes), and except when specifically considering genes not fit by the model, further analysis was limited to these genes.

Bootstrap assessment of significance

To validate the significance of the linear model's fit to our yeast aneuploidy gene expression measurements, we used bootstrapping (Efron 1993) to randomly resample 10,000 synthetic "genes." In order to preserve the structure of the independent variables (e.g. conditions of monosomy, disomy, and trisomy are mutually exclusive, and any one of them entails a condition of transcriptional aneuploidy), we performed this resampling as follows. First, consider the set of 6,256 measured yeast genes $G=\{g_1, ..., g_{6,256}\}$ with a total of 740,677 expression measurements $\boldsymbol{e}=[e_1, ..., e_{740,677}]$, where $g(e_i)\in G$ indicates the gene providing measurement e_i. Each of these expression measurements e_i was taken under some conditions described by a vector $\boldsymbol{x}_i$ containing the (mainly binary) variables detailed above, inducing a matrix X with rows $\boldsymbol{x}_1$ through $\boldsymbol{x}_{740,677}$. For each gene g_i, let $n(g_i)$ be the number of expression measurements (i.e. elements of $\boldsymbol{e}$) assayed for g_i. Then for

our synthetic genes $G^s=\{g^s_1, ..., g^s_{10,000}\}$, we generated synthetic expression values $\boldsymbol{e}^s$ and conditions X^s:

1. For $g^s_i \in G^s$:
2. For a random $g \in G$, let $n=n(g)$.
3. Repeat n times:
4. Draw e randomly from $\boldsymbol{e}$.
5. Draw $\boldsymbol{x}$ randomly from the rows of X.
6. Set $g(e)=g^s_i$, add e to $\boldsymbol{e}^s$, and add $\boldsymbol{x}$ to X^s.

This yielded 10,000 synthetic genes, each with between 76 and 108 randomized expression measurements and conditions, for a total of 1,070,637 bootstrapped expression values. These data were fit using an identical model and procedure to the experimental data as described above.

The *S. cerevisiae* Phosphoproteome

We present an analysis of the yeast phosphoproteome based on data that uses endo-Lys C as the proteolytic enzyme, immobilized metal affinity chromatography (IMAC) for phosphopeptide enrichment, a 90 min, nanoflow-HPLC/electrospray-ionization tandem mass spectrometry experiment for phosphopeptide fractionation and detection, gas phase ion/ion chemistry (ETD) for peptide fragmentation, and the open mass spectrometry search algorithm (OMSSA) for phosphoprotein identification and assignment of phosphorylation sites. From a 30μg (~600pmol) sample of total yeast protein, we identify 1,252 phosphorylation sites on 629 proteins. Identified phosphoproteins have expression levels that range from <50 to 1,200,000 copies/cell and are encoded by genes involved in a wide variety of cellular processes. We note that most protein-

kinase recognition-motifs predicted by SCANSITE are significantly enriched in our phosphorylation data and present evidence for a novel motif recognized by one or more unidentified yeast kinases. We analyze the identified phosphoproteins in the context of interaction networks and find that they have a significantly higher number of interactions than expected and that yeast kinases themselves contain a disproportionately large number of phosphorylation sites. We note that the observed phosphoproteins, but not individual phosphosites, are likely to be conserved across very large evolutionary distances.

Results and Discussion

Our 629 phosphoproteins were identified from *Saccharomyces cerevisiae* grown on rich medium containing glucose. These proteins represent a random sample of yeast proteins covering most cellular processes (Figure 36A). The only GO (Ashburner, Ball et al. 2000) biological process term significantly underrepresented is *translation elongation*. We presume this is because we prepared the cells by centrifugation, and the high density of cells in the pellet leads to rapid starvation, which would be reflected by a rapid change in the phosphorylation status of translation elongation factors (de Haro, Mendez et al. 1996). The abundances of the phosphoproteins identified are also similar to the global distribution of protein abundance (Figure 36B) estimated from genome-wide protein affinity purification experiments derived from cells grown under similar conditions to those used in our studies (Ghaemmaghami, Huh et al. 2003). Together, these data suggest that the identified phosphopeptides are encoded by a representative sample of genes corresponding to a wide variety of cellular processes and are observed in proportion to their expression within the yeast proteome.

We find the identified phosphoproteins to be enriched in a small number of specific GO processes, particularly fermentation, protein synthesis, and phosphorylation-related processes.

This is not surprising, given that the cells were grown in rich medium under conditions favoring rapid growth and fermentation. Interestingly, the genes are also enriched in a subset cell division-specific processes; namely, budding, polarity and cytokinesis. This was unexpected and suggests that there is a high degree of phosphorylation-dependent regulation of these events. Phosphorylated proteins were significantly more likely to themselves possess known phosphotransferase activity. This suggests that phosphorylation is likely to be a common regulation mechanism for kinases in yeast.

We further analyzed the kinase-substrate relationships in our data set by using SCANSITE to predict motifs within the phosphorylated proteins from which the peptides were derived (Obenauer, Cantley et al. 2003). Nearly every protein kinase recognition motif predicted by the SCANSITE was significantly enriched in our phosphorylation data (Supplemental Figure 4). A summary of the SCANSITE kinase target-groups found in our data, using medium stringency criteria, appears in Supplemental Table 3. Basophilic sites make up the largest group of motifs, while acidophilic and proline-directed motifs are also well represented. These results are in agreement with data from other large datasets (Beausoleil, Jedrychowski et al. 2004; Nuhse, Stensballe et al. 2004). In addition to these known sites, our assay identifies 381 phosphorylation sites not found by SCANSITE. The relative occurrence of amino acids flanking the sites of phosphorylation appears as a heat map in Figure 36C, D, and E. The acidophilic (Figure 36C) and basophilic (Figure 36D) maps are as predicted from SCANSITE. The heat map for the novel sites (Figure 36E) shows an enrichment of proline and histidine at positions +1 and –1, respectively, to the site of phosphorylation. We suggest that this likely represents a novel motif for one or more protein kinases in yeast.

The majority of the phosphorylation sites in our data are on serine (82.3%), and the remainder on threonine (17.5%) and tyrosine (0.027%). This confirms the extremely low extent of tyrosine phosphorylation in yeast (Modesti, Bini et al. 2001). There are no true protein tyrosine kinases in yeast, but there are seven dual-specificity kinases predicted on the basis of sequence similarity: the MAPKK proteins (Ste7p, Mkk1p, Mkk2p, and Pbs2p) and three kinases that regulate the cell cycle (Rad53p, Mps1p, and Swe1p) (Hunter and Plowman 1997). None of the tyrosine-phosphorylated proteins are obviously related by functional annotation. However, one of the proteins, Cnm67p, is a structural component of the spindle pole body and is an excellent candidate for a substrate of Mps1p, a protein kinase that regulates spindle pole body duplication (Lauze, Stoelcker et al. 1995; Schaerer, Morgan et al. 2001).

We took advantage of diverse functional genomic and proteomic data to analyze protein kinase targets in the context of interaction networks. We began with affinity precipitation and yeast two-hybrid data consisting of 13,325 interaction pairs and 4,697 proteins (Bader, Donaldson et al. 2001; Breitkreutz, Stark et al. 2003). We found that phosphoproteins have a significantly higher number of total physical interactions than expected and, in particular, interact with other phosphoproteins more than expected (Figure 37B, Supplemental Table 4). One explanation for this is that signaling cascades are often organized upon molecular scaffolds that promote the physical association of proteins participating in the signal-transduction pathway (Bhattacharyya, Remenyi et al. 2006). Among genetic interaction data (4,775 interaction pairs, 1,469 proteins), we found that genes encoding phosphoproteins also exhibit a strong tendency towards a high degree of interaction and exhibit enriched interactions with other phosphoproteins in genetic interaction networks as well. This effect has been previously observed for essential proteins and has been implicated for conserved proteins (Jeong, Mason et al. 2001). Interestingly, phosphoproteins are

not statistically likely to be essential, but they are highly enriched for strongly conserved genes (Figure 37C). We propose that non-essential phosphoproteins play a central role in biological processes making their conservation evolutionarily advantageous.

One example of the propensity of phosphorylated proteins to be hubs and to interact with other phosphorylated proteins is shown in Figure 37A. The network involves proteins required for three successive stages of the cell cycle, and there are two interconnecting hubs that are phosphorylated protein-kinases having multiple interactions. Dbf4p is the regulatory subunit of the Cdc7p kinase that regulates the initiation of DNA synthesis, and Cdc5p is the polo-like kinase that is an important mitotic regulator. The network is especially interesting given that the cell must coordinate DNA metabolism with mitosis. Dbf4p interacts with two general classes of proteins, one (blue) required for the initiation of DNA synthesis (Cdc7p, Cdc45p, Mcm2p, Orc2p, Orc3p, Orc5p, Orc6p, and Swi5p) and the other (green) required for the DNA damage checkpoint (Chk1p, Ddc1p, Mec3p, Rad9p, Rad17p, Rad24p, and Rad53p). Similarly, Cdc5p also interacts with two general classes of proteins; one (dark grey) is required for chromosome structure and executing anaphase (Mcd1p, Smc1p, Smc3p, and Swe1p) and the other (light grey) for the exit from mitosis (Cdc15p and Mob1p). One interesting possibility is that the phosphorylation status of the hub dictates which of the two classes of interactions occurs. Alternatively, the phosphorylation status may regulate interactions between the hubs.

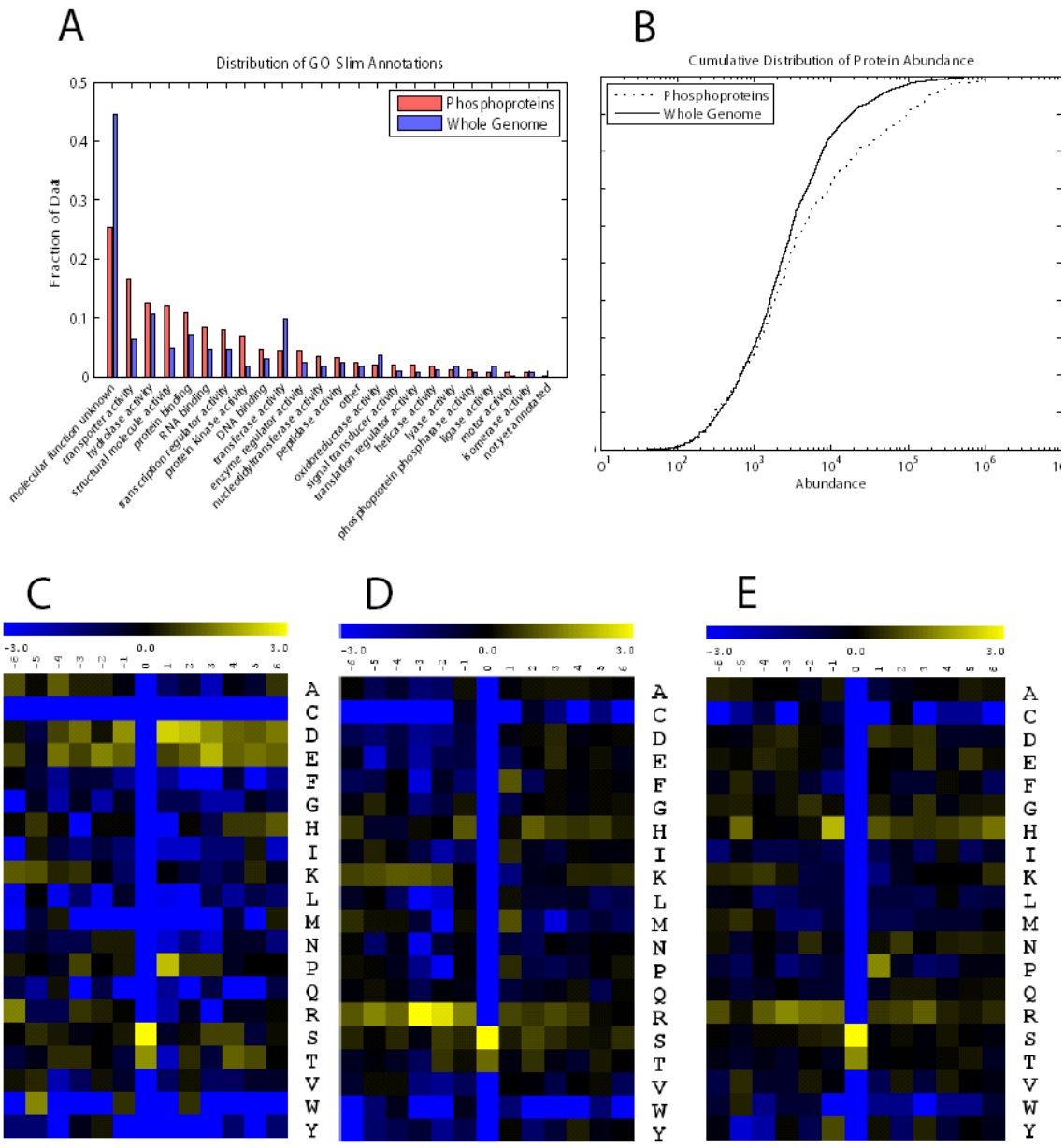


Figure 36: Phosphoprotein and amino acid frequencies. A) Comparison of GO Slim (Dwight, Balakrishnan et al. 2004) term frequencies between the whole genome and the phosphoproteins. B) Comparison of protein abundances between the entire genome and the phosphoproteins. The distribution of phosphoprotein abundances is comparable to that of the genomic background. C) $\log_2$ ratios of per-site amino acid frequencies relative to the genomic background, identified as acidophilic by SCANSITE, D) basophilic, or E) not recognized by SCANSITE.

Phosphorylation is expected to be evolutionarily conserved given the overall importance of phosphorylation in cell signaling and regulation. Indeed, we find that phosphorylated proteins are significantly more conserved as compared to other proteins in the proteome, even across large evolutionary distances (Figure 37C). Conserved phosphoproteins are much more likely to be conserved across large evolutionary distances covering all of the genomes we examined (*A. gossypi*, *C. elegans*, *D. melanogaster*, *H. sapiens*, *A. thaliana*) (Figure 37C, rightmost bars). Given the presumed importance of phosphorylation sites in the functionality of these proteins, it is expected that the phosphorylation sites themselves would be more strongly conserved than the surrounding protein.

We compared the conservation of phosphorylation sites within the sequenced fungal genomes (including separate tests against *sensu stricto* and *sensu lato* that span 10 and 300 million years of evolution, respectively) as well as with more distant model organisms (Supplemental Figure 5) (Balakrishnan, Christie et al. 2005). Surprisingly, phosphorylated serines and threonines were not found to be significantly more conserved than similar residues in the surrounding protein, regardless of evolutionary distance.

We have presented a strategy for the analysis of phosphoproteomes that uses endo-Lys C as the proteolytic enzyme, IMAC for phosphopeptide enrichment, ETD for peptide fragmentation, and the open mass spectrometry search algorithm (OMSSA) for phosphopeptide identification and assignment of phosphorylation sites. With this approach, we identified 1,252 phosphorylation sites on 629 proteins in a single experiment on 30 μg (~600pmol) of protein from a yeast whole cell lysate. Expression levels of identified phosphoproteins varied from <50 to more than 1,200,000 copies/cell. By implementing the ETD technology on LTQ-orbitrap and LTQ-FTMS instruments, it should be possible to sequence still larger phosphopeptides and, possibly, intact

phosphoproteins on a chromatographic time scale and thus to locate multiple phosphorylation sites that occur on the same protein molecule. Analysis of the resulting *S. cerevisiae* phosphoproteome reveals that phosphoproteins have a significantly higher number of interactions than expected and that kinases are in turn highly regulated by phosphorylation. Surprisingly, while phosphoproteins are highly enriched for essentiality and evolutionary conservation, individual phosphosites are not, suggesting an interesting evolutionary plasticity in the phosphorylation-based regulatory network.

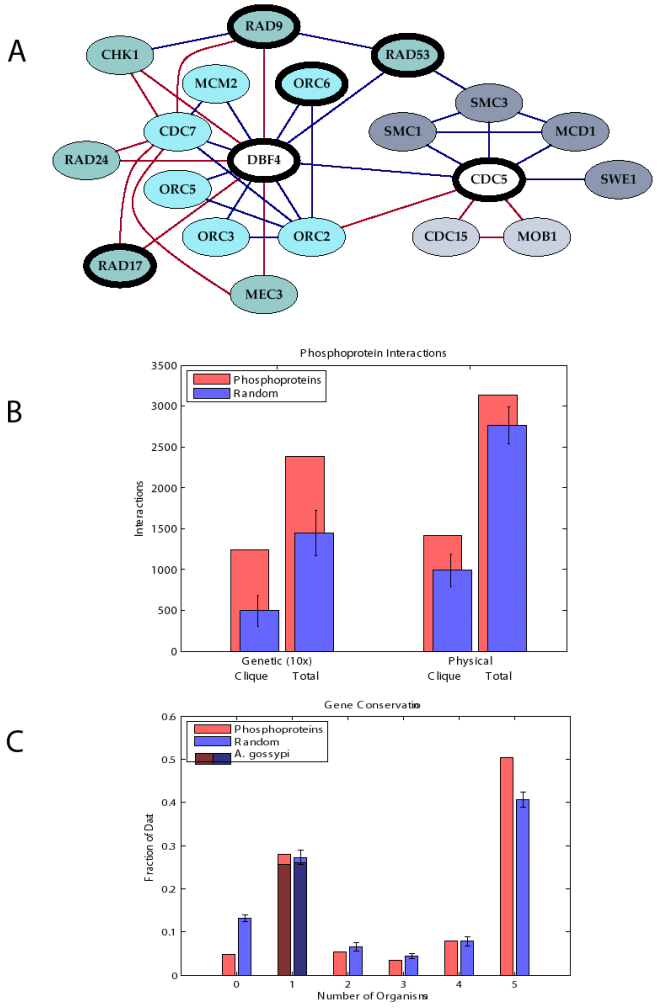


Figure 37: Phosphoprotein interactions and conservation. A) A subset of the KEGG sce04110 Cell Cycle pathway (Kanehisa, Araki et al. 2008). Proteins phosphorylated in this study appear as black nodes. Known physical interactions are represented by green edges, and known genetic interactions are shown as red edges. B) A comparison of phosphoprotein interactions to those of random genomic samples. Clique interactions represent genetic or physical interactions between phosphoproteins (or within random sub-samples), and total interactions contain all known genetic or physical interactions between phosphoproteins/sampled proteins and the yeast genome. C) A representation of genes with significant BLASTP hits across five model organisms (*A. gossypi*, *C. elegans*, *D. melanogaster*, *H. sapiens*, and *A. thaliana*). Phosphoproteins are much more likely than a random yeast protein to be conserved (leftmost bars), and conserved phosphoproteins are much more likely to be conserved in all five genomes examined (rightmost bars). Conservation in just one genome is largely explained by the data from the closest organism to *S. cerevisiae*, *A. Gossypi* (overlay).

Scaling Up: Large Data Collections and Complex Organisms

Throughout most of the history of biology, experimentation has occurred within basically two paradigms: macro-scale observations of phenotypes at the level of whole organisms or ecological communities, and micro-scale assays investigating the behavior and responsibilities of one or a few genes. The former has provided us with our modern body of knowledge regarding medicine, physiology, and ecology; the latter has developed into the immense field of molecular biology. The last decades of the 20th century provided two new developments that are still in the process of reshaping both areas, however, in the form of whole-genome sequencing (Goffeau, Barrell et al. 1996; Blattner, Plunkett et al. 1997) and high-throughput experimental assays (Fields and Song 1989; Schena, Shalon et al. 1995; Tong, Evangelista et al. 2001). For the first time, this data provided a view of biology that bridged the gap between individual biomolecules and organismal phenotypes, and the incredible subsequent developments in biotechnology and in basic science attest its effectiveness (Joyce and Palsson 2006).

Genome-scale data provides a wide range of immediate benefits. In particular, for unicellular organisms, it can monitor the entire organism's DNA (Solinas-Toldo, Lampel et al. 1997; Pinkel, Segraves et al. 1998), mRNA (Eisen, Spellman et al. 1998), and protein (Aebersold and Mann 2003; Hall, Ptacek et al. 2007) levels in response to any environmental stimulus. For metazoans, it can enumerate the gene expression (Kim, Lund et al. 2001) or signaling mechanisms (Garcia, Shabanowitz et al. 2005) characteristic of individual cell or tissue types, and it is increasingly capable of doing so on a single cell (as opposed to population) level (Bullen 2008). As has been observed in previous chapters, these advances have made possible the solution of specific

scientific problems by the collection and analysis of individual, focused high-throughput datasets.

As the availability and diversity of genomic data increased, so did interest in performing integrated analyses on these heterogeneous datasets (Marcotte, Pellegrini et al. 1999). This immediately raised a host of challenges, for inasmuch as high-throughput data can quickly answer many questions, it can quickly raise just as many new ones: genomic data is almost always noisier than classical experimental results (Altman and Raychaudhuri 2001), it requires various types of technological and biological normalization (Quackenbush 2002), it can be difficult to resolve measurements from different experimental platforms or environments (Bader and Hogue 2002; Troyanskaya, Dolinski et al. 2003), and in the worst case, near-replicate experiments can produce markedly different results (Bullinger, Dohner et al. 2004; Valk, Verhaak et al. 2004) (often for perfectly valid biological reasons (Nevins and Potti 2007)). There has since been a great deal of effort in integrating data of both related (Bork, Jensen et al. 2004; Rhodes, Yu et al. 2004; Huttenhower, Hibbs et al. 2006; Hibbs, Hess et al. 2007) and differing (Karaoz, Murali et al. 2004; Lee, Date et al. 2004; Myers and Troyanskaya 2007) experimental types, and the success of these methods has been one of the main supporting elements of the genomic revolution.

Now that high-throughput experimental techniques have been in widespread use for nearly a decade, tens of thousands of genome-scale datasets are publicly available (Barrett, Suzek et al. 2005), spanning hundreds of organisms, tens of thousands of genes, and hundreds of thousands of experimental conditions. The capabilities of whole-genome assays have immeasurably enriched our ability to ask - and answer - biological questions by simultaneously monitoring thousands of genes; what new questions can we ask when we can monitor not thousands of

genes, but thousands of genomes? Conversely, just as high-throughput data presented new computational challenges in order to be meaningfully analyzed, what new problems do we face in understanding and integrating this vast amount of information?

This chapter discusses three aspects of these issues and the biological discoveries that can be made when they are overcome. First, an overview of the MEFIT (Microarray Experiment Functional Integration Technology) system presents one scalable solution for integrated normalization and analysis of gene expression data. From a computational perspective, this process introduces methods for manipulating hundreds of microarray datasets and making them mutually comparable; from a biological perspective, this provides biologists with a way of monitoring genes and detecting common biological signals across thousands of experimental conditions, signals that might be too weak to detect with just a single assay. Second, we scale up from observations of single genes and proteins to an analysis of the systems-level functional structure apparent from large genomic data collections; where one high-throughput dataset might provide information on the interactions among individual genes, thousands of such datasets can detail the interactions and coregulation among entire pathways or cellular processes. Finally, we scale up to the genomic analysis a complex metazoan system with many times more genes, cells, and cell types: human beings. The HEFalMp (Human Experimental/Functional Mapper) system provides a means of integrating and exploring human genomic data through functional maps, which summarize the interactions among individual genes, groups of proteins in pathways or complexes, and genetic disorders, all in the context of differing biological processes or tissue types. Each of these systems addresses specific computational challenges in the normalization and manipulation of large, structured genome-scale datasets, and each

provides ways to discover new biology by leveraging the ever-increasing resources of genomic data.

We would like to thank Erin M. Haley and Hilary A. Coller for their experimental collaboration, particularly in the validation of the HEFalMp system.

MEFIT: Graphical Models for Large Scale Microarray Integration

Within the past decade, biological datasets have become available spanning not just whole genomes but multiple genomes, both within and across species. In particular, microarray coexpression studies routinely profile whole genomes simultaneously; and with shrinking costs, thousands of whole-genome experiments have become publicly accessible for many model organisms. Many methods have been proposed for extracting biological meaning from microarray data, including normalization and meta-analysis (Choi, Yu et al. 2003; Moreau, Aerts et al. 2003; Griffith, Pleasance et al. 2005; Hu, Greenwood et al. 2005), clustering (Eisen, Spellman et al. 1998; Heyer, Kruglyak et al. 1999; Butte, Tamayo et al. 2000; Cheng and Church 2000; Allison, Cui et al. 2006), signature algorithms (Ihmels, Friedlander et al. 2002; Bergmann, Ihmels et al. 2003; Ihmels, Bergmann et al. 2005; Kloster, Tang et al. 2005), detection of differential expression (Ideker, Thorsson et al. 2000; Baggerly, Coombes et al. 2001; Cui and Churchill 2003), and many others. Although complete analysis of individual microarray datasets is by no means a solved problem, it is of interest to begin examining the additional conclusions derivable from analysis of many microarray datasets. Integration such as this can enable broader understanding of gene regulation in the context of specific pathways and can allow the discovery of coexpression relationships too weak to be detected in individual experiments.

Such integrated analysis of microarray datasets is challenging because of differences in technology, protocols, and experimental conditions across datasets. Thus, any microarray integration system must be robust to such differences, and should easily adjust to new datasets, perhaps from technologies yet to be developed. Furthermore, in examining any diverse biological datasets (such as microarray results drawn from differing experimental conditions), it is critical to consider functional specificity, i.e. which biological processes are active in which experiments (Huttenhower and Troyanskaya 2006). For example, in a set of a thousand microarray experiments over *S. cerevisiae*, only ten experiments might have been performed under conditions inducing sporulation. These few microarrays might show strong coexpression of meiotic genes not expressed or not coregulated under other circumstances. This is a benefit in that it provides more specific information regarding meiosis-related genes, but such a relatively small signal can easily be lost during data processing. The problem of integrating many high-throughput data sources thus includes a problem of determining functional relevance; not only can such data reveal genes that are functionally related, it can also reveal the biological circumstances under which they relate.

To this end, we propose a Microarray Experiment Functional Integration Technology, MEFIT; this is a Bayesian framework facilitating the integration of multiple microarray datasets for predicting coexpression-based functional networks of proteins. Furthermore, each of MEFIT's predicted functional relationships is provided within the context of a specific biological process. These biological functions of interest can be provided directly by a biologist, or they can be derived automatically from functional catalogs such as the Gene Ontology (Ashburner, Ball et al. 2000) or MIPS (Ruepp, Zollner et al. 2004). In addition to its predicted functional relationships,

MEFIT's analysis process also provides a functional association score indicating how predictive each input microarray dataset is of each biological function.

Most prior work in large scale microarray integration has been performed in one of two contexts: statistical meta-analysis or the introduction of multiple microarray experiments into heterogeneous data integration systems. (Choi, Yu et al. 2003), (Rhodes, Yu et al. 2004), (Hu, Greenwood et al. 2005), and (Mulligan, Ponomarev et al. 2006) are representative examples of the former, all of which use meta-analysis to integrate microarray experiments for the detection of differential gene expression. In MEFIT, we take advantage of similar meta-analytic techniques in order to make disparate microarray experiments comparable, but we build upon the results to make predictions of global coexpression relationships and biological function and to determine the functional specificity of input microarray datasets.

(Pavlidis, Weston et al. 2002), (Clare and King 2003), (Troyanskaya, Dolinski et al. 2003), (Lee, Date et al. 2004), and (Butte and Kohane 2006) describe methods for the use of heterogeneous data integration to predict gene function or functional relationships, but none of these (or similar) systems focus specifically on the way in which microarray experiments are integrated. Most often, correlation over individual datasets or all datasets simultaneously is used with minimal inter-study normalization. This can result in a surprising amount of lost information, particularly since microarrays often represent by far the most extensive body of data available for integration (Pavlidis, Weston et al. 2002; Troyanskaya, Dolinski et al. 2003; Karaoz, Murali et al. 2004; Lee, Date et al. 2004). MEFIT improves on these prior systems by providing a scalable integration system specifically for microarrays that takes advantage of the functional diversity of coexpression data to improve prediction accuracy, to provide additional biological context for predicted functional relationships, and to identify biological functions in which individual

datasets are particularly informative. To our knowledge, none of these prior systems has provided a means of predicting both gene pair functional relationships and the specific biological processes in which those interactions are expected to occur.

The MEFIT system predicts functional relationships between genes within individual biological processes, consuming microarray datasets and known functional annotations as input. These predictions are generated as probabilities using a Bayesian framework trained in a function-specific manner. This training process allows one to derive relevance scores from the learned network parameters indicating how reliable the system finds each dataset to be within each biological process. Thus, MEFIT determines which microarray conditions are informative for a particular biological function in addition to predicting process-specific functional relationships.

Methods

The primary outputs of the MEFIT system are predicted probabilities of gene pair functional relationships within individual biological functions. These coexpression networks are derived from naive Bayesian networks trained on a per-function basis using microarray data and known functional annotations. The learned parameters of these networks also contain information regarding how predictive each microarray dataset is of each biological function. Biological functions of interest are provided to the system as simple gene sets (i.e. lists of genes annotated to processes such as mitotic cell cycle or pathways such as fatty acid biosynthesis), which are used to generate known positive pairwise relationships. Known unrelated gene pairs (negatives) are provided as a separate input to the system. For these experiments, we use functional annotations from the Gene Ontology (Ashburner, Ball et al. 2000) to generate both positive and negative gene pairs.

As Figure 38 summarizes, microarray data are preprocessed in order to serve as observations during training and evaluation of MEFIT's collection of naive Bayesian networks. The input functional annotations are used to derive known functional relationships (for training and evaluation) as well as to provide functional specificity by dictating the biological contexts for which separate Bayesian networks should be constructed. Finally, naive Bayesian inference on these per-function networks serves to produce predicted functional relationships for gene pairs within each biological function.

Microarray data preparation

We assembled *S. cerevisiae* microarray data available from the Stanford Microarray Database (Ball, Awad et al. 2005), the NCBI Gene Expression Omnibus (Barrett, Suzek et al. 2005), and several independent sources; see Supplemental Table 5 for a complete list. These data comprised 40 unique datasets (time courses or other cohesive collections of experiments) drawn from 34 publications for a total of 712 individual experiments, some single and some dual channel. For each individual dataset, genes missing in more than 50% of the experimental conditions were removed, and the remaining missing values were imputed using KNNImpute (Troyanskaya, Cantor et al. 2001) with k=10. Finally, replicated genes were averaged to ensure that each dataset contained at most one expression vector per open reading frame.

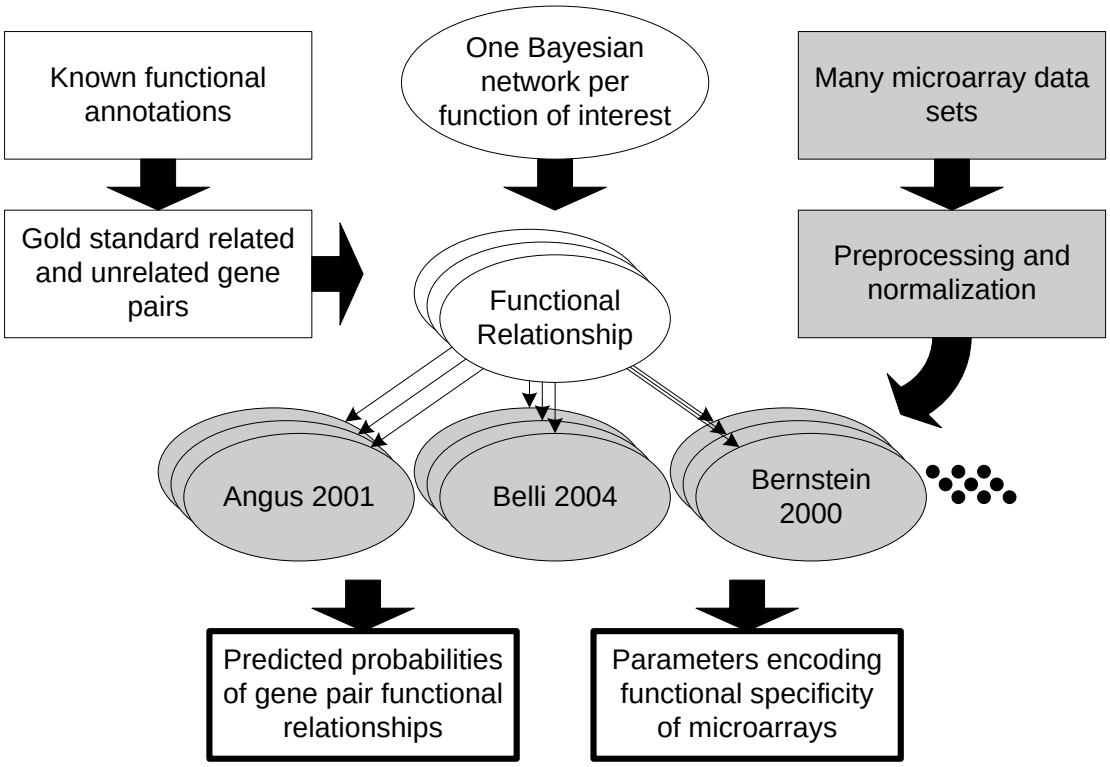


Figure 38: A schematic of MEFIT's data processing and control methodology. Microarray data sets are provided as input; these are preprocessed and quantized to serve as inputs for naive Bayesian networks. A single network structure is used for all biological functions, but the parameters of these networks are trained individually for each function of interest. Biological functions of interest and known functional relationships for training are derived from input sets of functional annotations. Finally, Bayesian inference produces a probability of functional relationship for each gene pair within each biological function.

For single channel data, expression values less than two were considered to be missing, and all single channel values were logarithmically transformed as a final preprocessing step. Since mismatch hybridization values were not available in many datasets, they were not used in this analysis.

Within each dataset, we calculated Pearson correlations between every pair of genes. These correlations were then normalized using Fisher's Z-transform (David 1949):

$$z = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

This maps a correlation ρ into a z-score, where the collection of pairwise z-scores within a dataset is guaranteed to be normally distributed. After dividing by the dataset standard deviation and subtracting the mean, this distribution will be $N(0, 1)$, making cross-dataset analyses more robust. Thus, from each microarray dataset, we produce a collection of gene pair z-scores representing the number of standard deviations their correlation lies away from the mean.

Bayesian data integration

MEFIT integrates multiple microarray datasets in specific biological contexts to allow for greater accuracy when predicting functional relationships. For each biological function of interest, MEFIT uses a naive Bayesian model to combine many microarray datasets and produce a single, integrated functional relationship score for every pair of genes, creating a function-specific, probabilistic coexpression network. Thus, a separate network is trained for each process or function of interest. Each of these networks generates predicted functional relationships within its particular biological function, and the parameters of the network encode how informative a dataset is within that function; an individual dataset is likely to provide varying degrees of predictive accuracy across disparate biological functions.

In each network, the probability of each dataset's observed correlation (represented as z-scores) is conditioned on the probability of functional relationship; each dataset's z-scores are discretized into five bins (below -1, -1 to 0, 0 to 1, 1 to 2, and above 2) and assigned to a single node in the model. Finer binning was found to lead to overfitting and data sparsity issues (data not shown). This results in a Bayesian network with one node predicting functional relationships (FR) and n

nodes conditioned on *FR*, each representing the value of some dataset D_i. For some gene pair (g_i, g_j) and supporting data $\{d_1(g_i, g_j), d_2(g_i, g_j), ..., d_n(g_i, g_j)\}$ with $d_k(g_i, g_j) \in \{0, 1, 2, 3, 4\}$, the probability of functional relationship is thus:

$$P_{i,j}(FR = yes) \propto \prod_{k=1}^{n} P(D_k = d_k(g_i, g_j))$$

The University of Pittsburgh Decision System Laboratory's SMILE library and GENIE modeling environment (Druzdzel 1999) employing the Lauritzen inference algorithm (Lauritzen and Spiegelhalter 1988) were used for Bayesian network manipulation, in addition to our own C++ implementations of basic naive Bayesian learning and inference (Neapolitan 2004). After training (discussed below), this method produces one probability of functional relationship per gene pair, referred to in this paper as the BAYESIAN integration process.

Testing and validation

MEFIT requires one or more sets of functionally related genes as input (positives), as well as a collection of unrelated gene pairs (negatives). For these experiments, we used 200 functions drawn from the Gene Ontology as positive sets (Supplemental Table 6); genes coannotated below these terms were considered to be functionally related. These terms were hand selected by a panel of six yeast genetics experts who were asked to evaluate whether each GO term would be informative enough to direct laboratory experiments. We constructed a set of all GO terms receiving four or more votes and added their descendants; we then trimmed this set by discarding any term for which all paths to the ontology root were blocked by another term in the set. Any pair of genes sharing an annotation beneath some term in this set was considered to be related (positive). To generate negative examples, any gene pairs not coannotated below some

GO term including at least 10% of the *S. cerevisiae* genome (roughly 645 genes) were considered to be unrelated (Supplemental Table 7). This resulted in a set of 619,278 related and 8,853,875 unrelated pairs.

Of the genes included in this set of pairs, 20% (951 genes) were randomly selected as test genes and held out of all training. Our test set consisted of any pair including at least one of these genes (241,408 positive and 3,320,786 negative pairs), and the remaining pairs were used for training. Performance was evaluated by areas under the ROC curves (AUC). All AUCs were calculated analytically using the Wilcoxon Rank Sum formula (Lehmann 1975). We generally observed only small differences between training and test performance. These training and test sets were used for the construction and validation of the global integration methods discussed below, and they were further subdivided for per-function analyses.

Global microarray integration

As discussed above, we implemented a version of our system that trains only one global network, referred to as BAYESIAN integration. For comparison purposes, we also implemented three non-Bayesian integration methods. Most naively, after preprocessing up to the gene averaging stage (excluding Fisher's z-transform), each microarray dataset was individually normalized per gene to have mean zero and standard deviation one. After this, all experiments were concatenated to create one large expression vector per gene, and pairwise Pearson correlations were calculated using these vectors. For each gene pair, conditions in which at least one missing value remained (due to genes not present in particular datasets) were removed from the correlation calculation. This resulted in the CONCATENATION integration technique.

One can also integrate microarray datasets using statistical meta-analysis similar to that discussed in (Choi, Yu et al. 2003). To accomplish this, we proceeded through preprocessing as described above to the point where each dataset was represented by a collection of pairwise z-scores drawn from $N(0, 1)$. For each gene pair, these z-scores were averaged over the datasets including that pair, producing the Z-SCORE integration data.

Finally, we implemented a version of the microarray integration method discussed in (Lee, Date et al. 2004) in the context of general data integration. In brief, pairwise Pearson correlations were calculated per dataset. The pairs in the training set were used to produce a modified precision/recall plot (a log-likelihood (LLS)/correlation plot) to which a sigmoid curve could be fit. This curve allowed transformation of correlations from the test set into a log-likelihood space from which datasets are integrated by taking the average LLS for a gene pair across all available data. This will be referred to as the LLS integration.

Functional analysis

For microarray integration on a per-function basis, it was necessary to further decompose the training and test sets into collections of gene pairs relevant to each biological process of interest. In all cases, a gene pair was considered relevant to some function if i) it represented a positive relationship and both genes were included in the function or ii) it represented a negative relationship and one gene was included in the function. This provides a definition of training and test sets for each function provided as input (e.g. the GO terms discussed above).

When evaluating the performance of the four global integration techniques on individual functions, training (for the BAYESIAN and LLS methods) was performed using the entire training

set. Evaluation was performed using each function's test set. Functions containing fewer than ten genes (45 gene pairs) were not considered during testing.

Using the same naive Bayesian framework, we also learned one network per function using the individual training sets - the MEFIT integration technique. In addition to the predictive benefits discussed below, this provided additional information relating each microarray dataset to each function of interest. Specifically, we calculated the average difference between the prior and posterior probabilities of a functional relationship for each dataset and function. For each biological function, dataset, and discretized value, we provided only that datum as input to the function's Bayesian network. We then averaged (over the five discretized inputs for a particular dataset) the absolute values of the differences between the network's prior and the posterior probabilities of functional relationship generated in this manner. This provides a measure of how "trustworthy" or influential each dataset is when predicting gene pairs in each function (Figure 41).

Results

Characteristics of functional relationships vary by biological process

The performance of the five integration techniques on selected GO functions can be seen in Figure 39. The MEFIT integration method yields an AUC increase of 5% or more (over the maximum of the other four methods) in 54 of the 110 functions for which evaluation of all five methods was possible. Performance increased by a smaller amount in 31 of the remaining functions and decreased by more than 5% in only two functions.

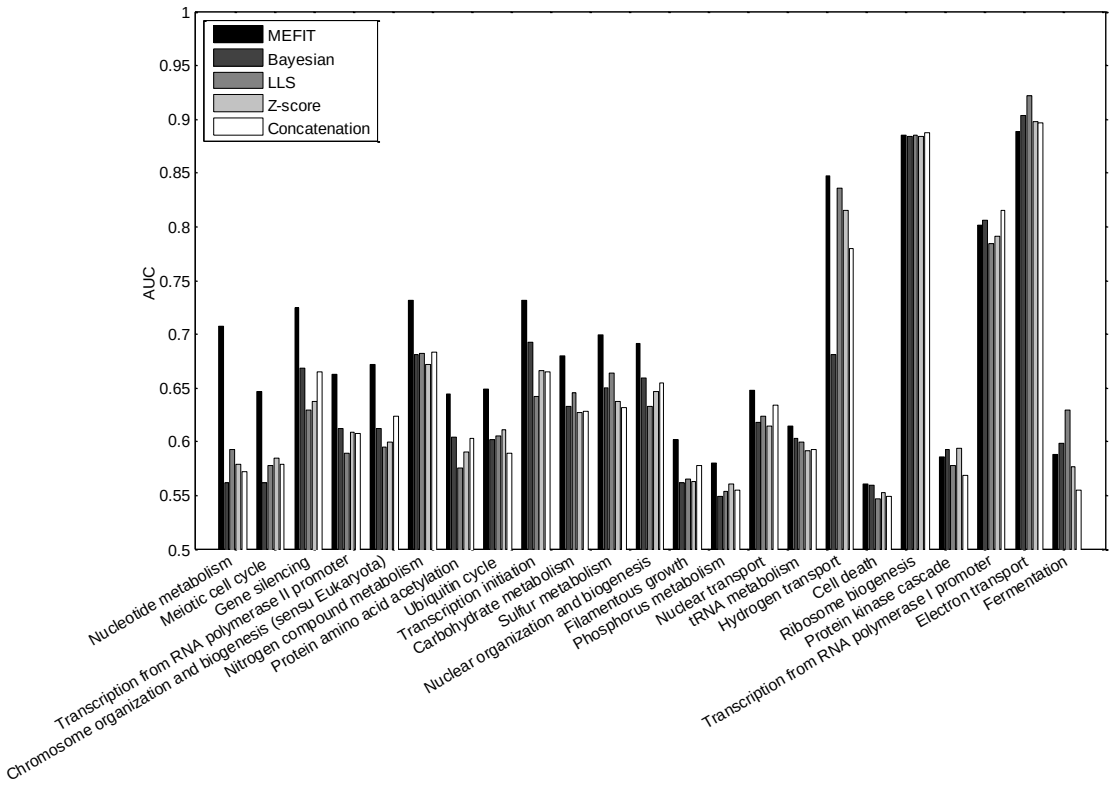


Figure 39: Areas under sensitivity/specificity curves (AUCs) for a selection of biological functions extracted from GO, ordered from most to least improvement and evenly spanning MEFIT's performance range. MEFIT showed an AUC increase of 5% or more over all other integration techniques in 54 of the 110 functions evaluated; AUC decreased by 5% or more in only two functions. AUC values range from random at 0.5 to optimal at 1.0. We measured performance for the CONCATENATION, Z-SCORE, LLS, BAYESIAN, and MEFIT integration techniques.

Interestingly, the functions in which the simpler CONCATENATION and Z-SCORE techniques perform well relative to the other three integration methods are also among those with the highest overall AUCs: *ribosome biogenesis*, *rRNA metabolism*, *RNA methylation*, *electron transport*, and *cellular respiration*. This may indicate that for such high-performing functions, little room for improvement exists given the currently available data. Indeed, these functions fall into two

categories, ribosomal processing and basic cellular metabolism, both of which are known to have clear "global" signals in microarray data (Eisen, Spellman et al. 1998; Jansen, Greenbaum et al. 2002). That is, given a collection of microarray experiments performed under almost any conditions, it is likely that genes related to ribosomes and cellular respiration will be coexpressed at detectable levels. This ubiquitousness makes these functions easy to detect by techniques such as CONCATENATION; even a modest signal present in most microarrays will be detectable in a correlation calculated across all experiments simultaneously. This accounts for the ease with which ribosomal function can be predicted from coexpression data (Gasch, Spellman et al. 2000; Karaoz, Murali et al. 2004; Lanckriet, Deng et al. 2004). Other functions in which MEFIT shows little improvement (such as *protein kinase cascade* or *fermentation*) are small or poorly studied functions in which data sparsity makes it impossible for any prediction technique to perform well.

Conversely, the functions in which MEFIT provides the most improvement tend to be specific functions that are reasonably well represented in the data but are poorly predicted by more global methods. For example, genes involved in the *meiotic cell cycle/response to pheromone/sporulation* group of functions should be coexpressed only under very specific circumstances; such a signal would be undetectable by correlation across all dataset simultaneously. Relative to CONCATENATION and Z-SCORE integration, the LLS method also provides some improvement for several such specific functions. Since MEFIT is designed to upweight datasets within functions where they demonstrate predictive power, this method is able to extract more localized signals originating in a few microarrays performed under appropriate conditions.

Bayesian integration of microarray datasets

It is of interest to compare the performance of the four global microarray integration methods - BAYESIAN, CONCATENATION, Z-SCORE, and LLS - on the entire answer set, without decomposing the results into specific biological functions. Additionally, one can reintegrate the individual components of the MEFIT output in a variety of ways to produce a global prediction set; relative to the BAYESIAN integration method, this has the benefit of preserving dataset/function associations by weighing each dataset by its relevance to each function. Ideally, each gene pair should be globally related if it is related in at least one function. In practice, noise in the predictions can make this assumption error-prone (a single overconfident prediction can dominate the overall probability), so we reintegrate each gene pair by taking the average probability of functional relationship over all functions in which that pair is predicted. This reintegration of the MEFIT output will be referred to as the AVERAGED MEFIT method.

Perhaps the most striking feature of a comparison of these global integration techniques (Figure 40) is the sharp decline in precision of the CONCATENATION method at low recall (i.e. high correlation). In other words, gene pairs strongly correlated across an extremely large vector of disparate microarray conditions tend to be functionally unrelated. This is caused by factors such as transposable elements and similar sources of homology (telomeric sequences, etc.) that can lead to non-functionally related coexpression under essentially any experimental conditions. These sequences are known to be problematic due to cross-hybridization in coexpression experiments (Bozdech, Zhu et al. 2003), and they are excluded from most microarrays; the CONCATENATION method interprets this absence as missing data and sees these genes only as very strongly correlated across the few datasets in which they are present, resulting in its poor performance at low recall.

Both the BAYESIAN and AVERAGED MEFIT results retain high precision at low recall cutoffs, with the AVERAGED MEFIT method also showing a substantial improvement in high recall areas. These integration techniques both explicitly encode the necessity to ignore or downweight inputs that tend to be overconfident (e.g. datasets in which a high correlation is not necessarily indicative of functional relationship), leading to their improved low recall behavior. AVERAGED MEFIT integration is able to perform the same downweighting on a per-function basis, which likely contributes to its greater precision at high recall.

Functional analysis reveals both expected and novel data content

In addition to per-function and global predictions of gene pair functional relationships, MEFIT also provides information on the relationships between microarray datasets and biological processes. Specifically, each per-function network learns during training how reliable it expects each dataset to be within its function. These reliabilities can be extracted as posterior probabilities after Bayesian inference, leading to a single confidence score for each dataset in each biological function.

Several aspects of these confidence scores (Figure 41) demonstrate clear agreement with the per-function results shown earlier and with existing biological knowledge. Nearly every dataset, for example, is highly informative regarding *ribosome biogenesis* and *rRNA metabolism* (Figure 41, light grey cluster), for the reasons discussed above; this is accompanied by a similar, weaker signal from the general *translation* function. Of the datasets in which ribosomal functions are not well predicted, (Angus-Hill, Schlichter et al. 2001) and (Rudra, Zhao et al. 2005) are knockouts in which ribosomal genes are specifically disrupted.

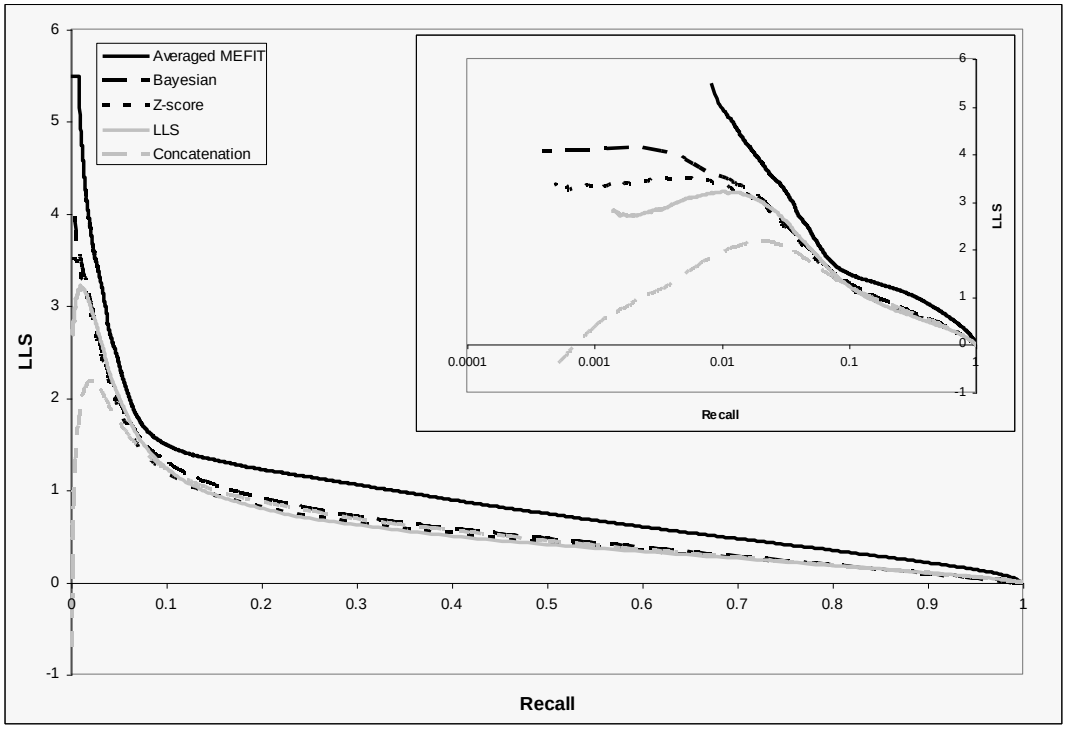


Figure 40: A comparison of the five global integration methods. A log scale inset is shown to emphasize the high precision area of biological interest; the minimum recall is limited to a minimum of 100 positive predictions to avoid noise. Performance is shown using the log-likelihood score $LLS=\log_2(TP \bullet N/FP/P)$ for P total positive pairs, N total negative pairs, and TP and FP the number of true and false positives at a particular sensitivity threshold.

Many datasets have moderately strong responses in *amine*, *amino acid*, and *organic acid metabolism* (Figure 41, black cluster), but the (Brem and Kruglyak 2005) and (Yvert, Brem et al. 2003) results particularly stand out. These are both recombination studies between lab (BY4716) and wine (RM11-1a) strains with a focus on regulatory relationships. Other strong signals arise from the (Hardwick, Kuruvilla et al. 1999) study investigating rapamycin treatment and nutrient response and from the two (Saldanha, Brauer et al. 2004) datasets for leucine and uracil limitation. All of these experiments have clear ties to amine and amino acid metabolism.

Three datasets are found to be particularly informative for *DNA recombination*, *M phase*, *meiotic cell cycle*, and *sporulation* (Figure 41, dark grey cluster); these are (Primig, Williams et al. 2000) (a sporulation time course), (Williams, Primig et al. 2002) (UME6 deletion, a known meiotic regulator), and (Jin, Laplaza et al. 2004). While the first two findings seem logical, (Jin, Laplaza et al. 2004) studies xylose metabolism and fermentation, which has no obvious connection to meiosis. Our functional relevance results in this case alert a biologist to the possibility of a sporulation response to an inhospitable medium, a pre-sporulation response (Pringle, Broach et al. 1997), or a disruption of the nutrient response pathways due to introduction of the XYL1, XYL2, and XYL3 genes (Jin, Laplaza et al. 2004), information that might not have been evident without such a per-function analysis.

A biologist could use such information in at least two ways. Given a set of existing microarrays and a pathway or process of interest, this functional decomposition reveals datasets with an increased likelihood of containing information regarding that pathway or process. Conversely, given a new microarray (possibly generated under experimental conditions spanning many functions), functional decomposition produces a summary of pathways potentially disrupted or activated under its conditions. These analysis methods would be lost in an integration technique not taking advantage of the functional specificity of microarray datasets and of functional relationships.

Novel functional predictions

Based on the integrated per-function coexpression networks predicted by MEFIT, we can make functional predictions for genes previously unannotated in the Gene Ontology. Specifically, we examined several functions in which MEFIT showed marked improvement over previous integration techniques and extracted the most confident predictions. Searching these predictions

for highly connected subgraphs involving both known and unknown genes produced several candidates, of which we chose to examine two: YML037C and YHR159W clustered around MMS4 in the *meiotic cell cycle* function, and YKR016W, YNL100W, and YNL274C clustered around INH1 and TIM11 in *hydrogen transport*.

All six of these genes are uncharacterized open reading frames annotated to *biological process unknown*, save for YNL274C's overly general *metabolism* annotation (which was unused in our analysis). In *hydrogen transport*, the GO term representing mitochondrial proton processing, INH1 and TIM11 are both proteins associated with the F1F0-ATP synthase (Brunner, Everard-Gigot et al. 2002). Of our predictions, YNL100W and YKR016W are known to localize to the mitochondrion (Huh, Falvo et al. 2003), and all three appear in the mitochondrial proteome (Sickmann, Reinders et al. 2003). Deletion of YKR016W also shows growth defects on non-fermentable carbon sources (Steinmetz, Scharfe et al. 2002), which we have confirmed in our lab (data not shown). YNL274C shows no strong localization, but its sequence contains a hydroxyacid dehydrogenase domain targeting NAD (Mulder, Apweiler et al. 2005), supporting our predicted role in cellular respiration.

Our predicted meiotic cell cycle group is centered on MMS4, which is a meiotic and mitotic gene involved in recombination and DNA repair (Xiao, Chow et al. 1998). YML037C shows a strong colocalization with clathrin coated vesicles (Huh, Falvo et al. 2003), appears to behave as a transcriptional activator (Titz, Thomas et al. 2006), and may be a substrate of the DBF2-MOB1 mitotic exit regulation complex (Mah, Elia et al. 2005). These characteristics point towards a potential mitotic or meiotic regulatory role for YML037C, in agreement with our prediction. YHR159W is thought to be a phosphorylation target of CDK1/CDC28, showing cell cycle regulation peaking in G1 (Ubersax, Woodbury et al. 2003); tests in our lab have shown that a

heterozygous deletion mutant appears to be defective in tetrad formation during sporulation, a phenotype that we are currently investigating in more detail.

Discussion

Here, we present MEFIT, a methodology for the simultaneous analysis of multiple microarray datasets using Bayesian integration augmented by per-function analysis. MEFIT's integration improves upon the general predictive power of existing methods for discovery of pairwise functional relationships from diverse microarray data. Additionally, it produces a per-function analysis for biologists, providing predictions in the context of individual pathways or biological processes (which may also be specified initially by the biologist).

Two strengths of MEFIT lie in its scalability and interpretability. Naive Bayesian learning and inference are both computationally inexpensive, and analysis can be performed simultaneously for hundreds of datasets spanning thousands of conditions. Additionally, given a single new dataset to integrate, no retraining need be performed - the conditional probabilities relevant to the new data can be learned independently of existing data. The statistics required for dataset normalization are fairly standard, and learned network parameters are readily interpretable and visualizable as probability distributions over each dataset and function.

As defined above, "functions" in this framework are simply gene lists defined by some prior method to be functionally related. These might consist of pathways or transcription factor modules specified by a biologist or of larger collections of genes; we have used groups of genes sharing annotations in the GO ontology (as well as performing initial validations with the MIPS hierarchy). This could easily be extended into other organisms, e.g. by using tissue types or

cancer pathways in mammalian systems. MEFIT learns to predict novel functional relationships similar to those specified in its input sets.

The output of MEFIT is one naive Bayesian network per function; dataset to function confidence values and per-function probabilities of gene pair functional relationships can be immediately derived from these learned networks. In other words, MEFIT produces one genetic interaction network per function in addition to a global interaction network; if desired, by interpreting pairwise probabilities as similarity scores, these predictions can be further visualized (e.g. as per-function clusters). Since functional relationships are frequently specific to individual biological processes (such as STE7/FUS3 interaction during pheromone response versus STE7/KSS1 interaction during nutrient limitation (Madhani and Fink 1997; Ptashne and Gann 2003)), this provides a biological perspective that is both more realistic and, by compartmentalizing interactions, more manageable.

We have made our test predictions available at the MEFIT web site (http://function.princeton.edu/mefit/) along with a collection of predictions for the entire *S. cerevisiae* genome constructed by training on all known data and evaluating all gene pairs (including unknowns). This site includes an interface for browsing these predictions and the large collection of microarray datasets used to generate them. We expect that this microarray integration methodology will be useful in the context of heterogeneous data integration tools, where it can provide more informative preprocessing of coexpression data. We have already established substantial biological evidence for several of MEFIT's predictions, and we hope that it will provide a useful tool for guiding future laboratory and high throughput experiments.

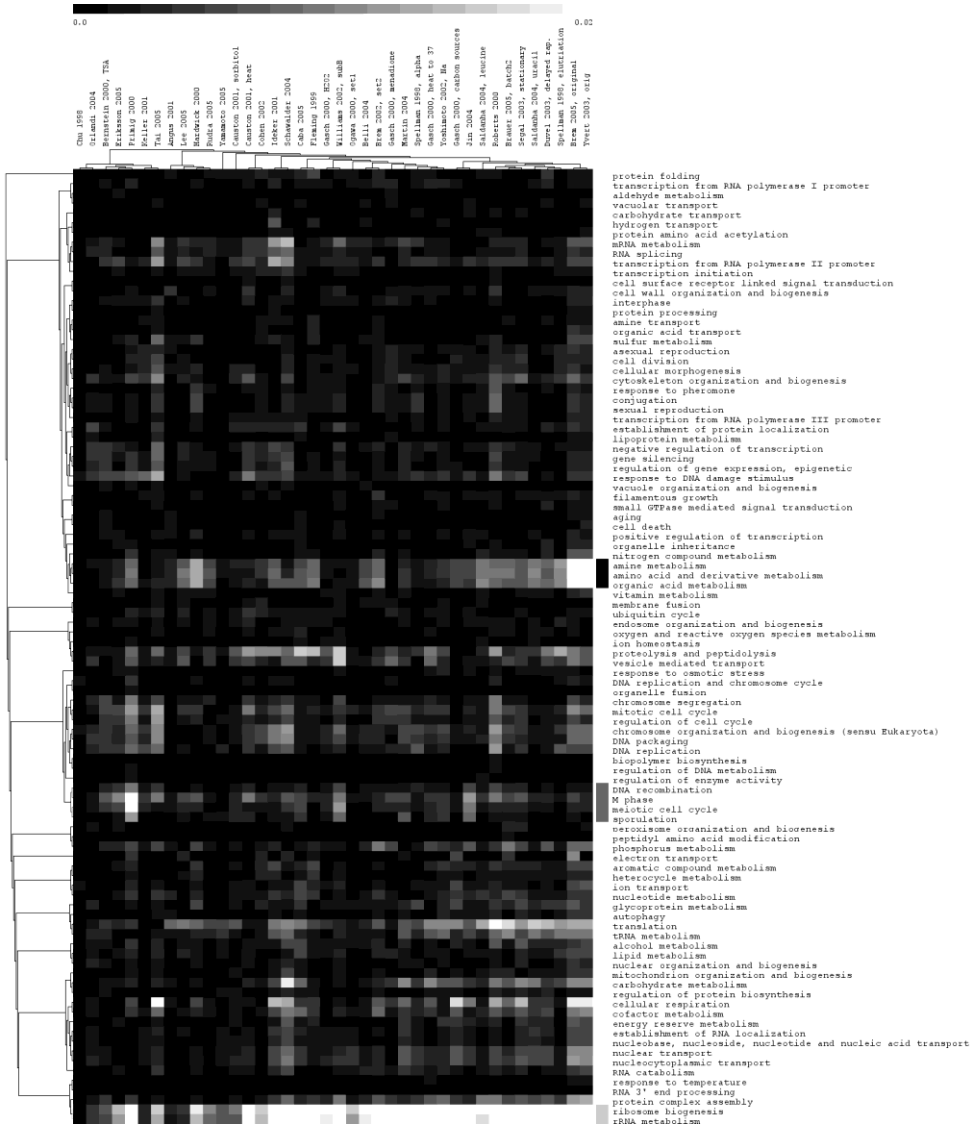


Figure 41: A portion of the per-function data set confidence scores learned by MEFIT. Brighter cells indicate a higher average posterior probability of functional relationship given input from a particular microarray data set in a particular biological function. These are calculated from networks averaged over a five-fold cross validation and are small due to the volume of microarray data employed (the maximum average difference for permuted data is ~0.005). Data sets and ontology terms have been clustered to visually show similarities in predictive power. The three colored clusters (amine metabolism in black, meiosis in dark grey, and ribosomal in light grey) represent interesting predictions discussed in the text. The heat map was generating using TIGR MeV (Saeed, Sharov et al. 2003).

Assessing the Functional Structure of Genomic Data

The technological developments of the past several decades have driven a continuing expansion of our understanding of molecular biology and a similar expansion in the analysis techniques applied to this data. In particular, genome-scale assays for coexpression (Eisen, Spellman et al. 1998; Spellman, Sherlock et al. 1998), genetic interactions (Giaever, Chu et al. 2002; Tong, Lesage et al. 2004), physical interactions (Gavin, Bosche et al. 2002; Ho, Gruhler et al. 2002), protein localization (Huh, Falvo et al. 2003), and regulatory networks (Zhu and Zhang 1999; Harbison, Gordon et al. 2004) have all opened up new opportunities for computational data mining that have been richly explored. Data such as these have been used in a variety of machine learning and other computational contexts (Jansen, Yu et al. 2003; Troyanskaya, Dolinski et al. 2003; Karaoz, Murali et al. 2004; Lee, Date et al. 2004; Franke, van Bakel et al. 2006).

As the amount of available genome-scale data has continued to increase, it has become possible to ask higher-level questions about the systems-level functional associations between entire pathways and processes. These associations represent the complex interplay between linked biological processes: DNA replication and mitosis are distinct cellular processes, for example, but they are functionally associated in their biological goals (cell division), regulation, and genetic participants. Understanding this network of associations between processes is a critical link between functional relationships at the single-gene level and phenotypes at the organismal level.

By deriving an understanding of large-scale functional structure based directly on genome-scale datasets, we also gain an understanding of the data itself. An examination of the pathways and processes perturbed by whole-genome experiments allows those experimental results to be described in terms of their functional activity. For example, microarrays performed under

conditions of heat shock and oxidative stress might both show functional activity related to an environmental stress response; this similarity of functional activity reveals biological commonalities between otherwise disparate experiments. By combining these two lines of inquiry - functional associations between processes and functional similarities between datasets - we gain insight into unexpected relationships in existing data, and we can direct experimenters to biological areas that are currently unexplored. All of these analyses deal with the high-level functional structure of genome-scale data and biological processes, which allows us to answer increasingly complex questions using the ongoing flood of high-throughput data.

We present such an analysis of functional associations among 141 biological processes and over 180 datasets (spanning >950 publications, >2,300 microarray conditions, and several thousand interaction, localization, and sequence-based data) in *S. cerevisiae*, where a functional association entails cooperation, coregulation, or other interaction between pathways and processes to perform a cellular task. These associations are derived by examining functional relationships between many individual genes, which are in turn predicted in a process-specific, probabilistic manner from heterogeneous data integration. This provides a global view of the functional structure of biological processes in yeast, including the degree of data-driven associations between processes, the experimental cohesiveness of gene behavior within each process, and the coverage of individual biological processes by currently available data. Likewise, we obtain measures of functional activity within each dataset - that is, which biological processes are covered by a dataset, independently of experimental platform. This high-level functional analysis technique is not specific to yeast and is extensible to any organism with a sufficiently large body of experimental data.

This analysis of functional structure produces a number of findings useful for guiding future experimental efforts and further computational studies. Specifically, we provide maps of data-driven associations between biological processes and of similar functional activities among datasets. By examining associations between processes, we observe several biological processes that could benefit from additional high-throughput data coverage, including *ion homeostasis and transport* and *mitochondrion organization*. We also highlight biological processes likely to be performed by currently uncharacterized genes (e.g. *autophagy*). Similar functional activities among datasets demonstrate commonalities in several large microarray studies and consistency between protein localization, synthetic lethality, and protein-protein interaction screens. These similarities also expose specific biological relationships, such as a subtle effect due to strain background we discovered in three otherwise diverse microarray datasets. All of these relationships are fundamentally driven by similarities in gene and protein response across hundreds of datasets, and this high-level analysis of such large-scale functional structure is valuable for guiding future experimentation and in understanding systems-level associations among biological processes.

Methods

In summary, we analyzed the large-scale structure of functional relationship networks predicted based on Bayesian integration of genomic data. Functional associations between biological processes from the Gene Ontology (Ashburner, Ball et al. 2000) were derived by further integration and analysis of these networks in a context-sensitive manner. Functional activity information for each dataset was calculated during the integration process, and this was used to further characterize functional similarities between datasets. The resulting process/process, process/dataset, and dataset/dataset association networks were mined for subgraphs and

interactions of high weight. All network visualization was performed using Graphviz from AT&T (Gansner and North 2000).

Data collection and gold standard generation

Data collection. The data employed in this study is a union of that from (Hibbs, Hess et al. 2007) and (Myers and Troyanskaya 2007). Non-expression data includes pairwise physical and genetic interaction data from a variety of databases (Alfarano, Andrade et al. 2005), (Stark, Breitkreutz et al. 2006), protein localization (Huh, Falvo et al. 2003), and sequence and TFBS similarities (Harbison, Gordon et al. 2004), (SGD 2006). Pairwise interaction data were represented as binary presence/absence values; where applicable, interaction profile similarities were calculated between genes from binary data using an inner product. For details, see (Myers and Troyanskaya 2007).

Expression data were collected from ~80 publications comprising ~120 datasets and ~2,300 conditions as described in (Hibbs, Hess et al. 2007) and initially processed as described in (Huttenhower, Hibbs et al. 2006). Datasets containing fewer than four experiments were initially merged, creating a merged microarray set that was subsequently processed identically to the remainder of the datasets. Each of these was converted from expression values to gene pair similarity scores using Pearson correlation normalized using Fisher's z-transform (David 1949) and subsequently z-scored:

$$fisher(g_i, g_j) = \frac{1}{2}\log\left(\frac{1+\rho(g_i, g_j)}{1-\rho(g_i, g_j)}\right)$$

$$z(g_i, g_j) = \frac{fisher(g_i, g_j) - \mu_f}{\sigma_f}$$

That is, the Fisher's transformed score between any two genes g_i and g_j is a transformation of their Pearson correlation ρ, and the final similarity between two genes $z(g_i, g_j)$ is the pair's Fisher score minus the mean Fisher score μ_f divided by the Fisher score standard deviation σ_f (both over all gene pairs).

After z-scoring, each expression dataset was quantized using the binnings (-∞, -1.5), [-1.5, -0.5), [-0.5, 0.5), [0.5, 1.5), [1.5, 2.5), [2.5, 3.5), [3.5, ∞); these represent steps of one standard deviation in z-score space. Mutual information was calculated between the resulting sets of discrete values, and any pairs of datasets sharing more than 15% of the possible information were merged by averaging z-scores. PISA (Kloster, Tang et al. 2005) modules (a biclustering algorithm) were also calculated for the expression data collection and transformed into pairwise scores for our analysis by counting the number of times each pair of genes coclustered after 500 iterations. These biclusters offered an orthogonal analysis of the microarray data capable of providing different information than the normalize correlation scores.

Gold standard generation. To perform supervised learning, we generated a gold standard of known functionally related and unrelated gene pairs. Biological processes of interest were selected from the Gene Ontology (Ashburner, Ball et al. 2000) using a method based on (Myers, Barrett et al. 2006). The standard developed in (Myers, Barrett et al. 2006) is specific to *S. cerevisiae*; using a similar voting method and polling six biologists, a set of 433 GO terms were selected for this study to be experimentally informative independent of organism. 141 of these have at least ten gene annotations in *S. cerevisiae*, and these were selected as processes (gene sets) of interest (Supplemental Table 8).

An answer set was derived from these processes of interest as described in (Huttenhower, Hibbs et al. 2006). Gene pairs coannotated to any of the 141 terms were considered to be related. A gene pair was unrelated in the gold standard if i) the two genes were both annotated to some term in the set of 141, ii) the genes were not coannotated to any of these terms, and iii) the terms to which the genes were annotated did not overlap with hypergeometric p-value less than 0.05. All other gene pairs were omitted from the standard (i.e. they were neither related nor unrelated for training and evaluation purposes).

For context-specific learning, this answer set was decomposed into subsets relevant to each process of interest. A gene pair was considered to be relevant to a biological process if either i) both genes were annotated to the process or ii) one of the two genes was annotated to the process and the pair was unrelated in the standard (i.e. not coannotated to another process).

Bayesian analysis

Learning Bayesian classifiers. One naive Bayesian classifier (Neapolitan 2004) was learned per biological process of interest; experiments with other network structures were shown to provide negligible performance improvements (Huttenhower and Troyanskaya 2006). Briefly, a global classifier was learned in which the class to be predicted was gene pair functional relationships (as defined in the gold standard) and each dataset formed one node in the network. 141 function-specific networks were learned with identical structures, each using a subset of the global gold standard as described above. When fewer than 25 gene pairs were available for a particular dataset/relationship combination, the global probability distribution was used for that condition. This defines the predicted probability of functional relationship between genes as a weight:

$$w_F(g_i, g_j) \propto \prod_D P_F[D = d(g_i, g_j)]$$

That is, the weight between genes g_i and g_j in function-specific network F is proportional to (using Bayes rule) the product over all datasets D of D's probability of experimental value $d(g_i, g_j)$ for the two genes.

All Bayes network manipulation was performed with a combination of custom C++ software and the SMILE library from the University of Pittsburgh Decision Systems Laboratory (Druzdzel 1999).

Predicting functional relationships. Each naive Bayesian classifier directly implies a functional relationship network in which nodes represent genes and edge weights consist of the posterior probabilities of functional relationships between gene pairs. The 141 function-specific networks were combined to form a predicted global interaction network by transforming each network's edge weights to z-scores (subtracting the mean predicted probability and dividing by their standard deviation) and averaging each gene pair's weight across all available networks.

Functional relationship and dataset enrichment predictions

Process/process relationships. As described above, for the purposes of this analysis, a biological process was defined as a set of related genes. The strength of a predicted functional relationship between two processes F and G was calculated as the average edge weight in the global interaction network within the edge set:

$$E_{F,G} = \{(g_i, g_j) \mid g_i \in F, g_j \in G, g_i, g_j \notin F \cap G\}$$

That is, the predicted functional relationship strength between functions F and G is the average weight of all edges in the global interaction network between genes g_i and g_j spanning the two gene sets and not coincident to any gene in their intersection. Note that this specifically excludes

process similarity due to overlapping curated annotations and retains only data-driven functional relationships.

Similarly, the functional cohesiveness of a process was measured as the ratio of the average edge weight in the process to the average edge weight incident to the process:

$$cohes(F) = \frac{2 \mid G \mid \sum_{g_i, g_j \in F} w_F(g_i, g_j)}{\mid F - 1 \mid \sum_{g_i \in F} \sum_{g_j \in G} w_F(g_i, g_j)}$$

where F is the function of interest, G is the genome, and $w_F(g_i, g_j)$ is the edge weight between genes g_i and g_j in F's predicted functional relationship network. This normalizes for processes that are inherently more interactive and have uniformly higher probabilities of functional relationship. tRNA genes are omitted for the purposes of these calculations, since they represent a large class of very related genes for which essentially no data is available (thus generating a large number of misleadingly low weights).

Process/dataset relationships. The predicted enrichment of each dataset within each biological process was derived from the conditional probability tables learned for that dataset's node within the appropriate function-specific Bayesian classifier. Specifically, the predicted enrichment for process F in dataset D was calculated as the weighted sum of the difference in posterior probability of functional relationship induced in F's classifier by evidence from each possible value of D:

$$rel(F, D) = \sum_{d \in D} P_F[D = d] \left| P_F[FR] - P_F[FR \mid D = d] \right|$$

For example, suppose the prior probability of functional relationship in the *ribosome biogenesis* process is 2% ($P_{rb}[FR] = 0.02$). The GRID- and BIND-based yeast two-hybrid dataset has two possible values, 0 representing no observed binding and 1 representing binding, thus $D = \{0, 1\}$. After learning, the Bayesian classifier for ribosome biogenesis indicates that a lack of binding makes little difference ($P_{rb}[FR | yth = 0] = 0.025$), but gene pairs that bind are very likely to be functionally related ($P_{rb}[FR | yth = 1] = 0.4$). However, there are relatively few such pairs ($P_{rb}[yth = 1] = 10^{-4}$), since most gene pairs in the genome have not been observed to interact by available two-hybrid data ($P_{rb}[yth = 0] = 0.9999$). Thus the strength of relationship between the process of ribosome biogenesis and the yeast two-hybrid dataset is $r = 0.9999|0.02 - 0.025| + 10^{-4}|0.02 - 0.4| \approx 0.005$. The exact value may differ due to rounding in this example.

The estimated coverage of a process in currently available data was calculated as the average of $rel(F, D)$ over all datasets in our study.

Dataset/dataset relationships. This calculation of predicted process/dataset enrichments results in a vector of 141 values in the range [0, 1] for each dataset. To determine the functional similarity between two datasets, each value is first transformed to a log ratio against the average across all datasets:

$$rel'(F, D) = \log(rel(F, D) \cdot |D| / \sum_{d \in D} rel(F, d))$$

This normalizes against the fact that certain biological processes are inherently more apparent in most high-throughput data (e.g. most microarray datasets have strong signals for processes such as *translation*). The functional similarity between datasets is then the Pearson correlation of the resulting r' vectors across all datasets.

Gene/function relationships. For the purpose of predicting gene function based on "guilt by association" with known genes in some process, the connectivity of a gene to a process was assessed as follows. Each gene/process pair was assigned a functional association score equal to the ratio of its average probability of functional relationship to the process over the process's cohesiveness:

$$assoc(g_i, F) = \frac{\sum_{g_j \in F} w_F(g_i, g_j)}{|F| cohes(F)}$$

This calculation was also used to predict each biological process's predicted association enrichment with unknown genes. A list of 1,451 genes with no annotation below biological process was extracted from the Gene Ontology. A function's strength of association with unknowns was then the sum of its association scores for these 1,451 genes.

Robustness. A robustness study was carried out by randomly shuffling data points within each dataset prior to Bayesian learning. The resulting networks had average dataset functional enrichment scores of $4.46 \cdot 10^{-5} \pm 1.57 \cdot 10^{-4}$, biological processes cohesiveness of 1.37±1.32, and association between processes of $7.14 \cdot 10^{-3} \pm 0.0293$, the last due to the greatly reduced differentiation between processes. In contrast, the averages for these values in our results are $2.43 \cdot 10^{-4} \pm 6.02 \cdot 10^{-4}$, 15.1±35.9, and $1.94 \cdot 10^{-3} \pm 0.141$, respectively.

Dense subgraphs. An implementation of a modified greedy heuristic for discovering heavily weighted subgraphs (Charikar 2000) was used to mine interaction networks for cohesive modules. Briefly, to discover each module within the network of interest, a node set was initialized with the most cohesive pair in the network. Nodes were added to this set greedily based on edge weight until no node could be added without reducing the average cohesiveness

of the node set below the network baseline. The average edge weight of the set was then subtracted from each edge between nodes in the set, and the process was iterated to discover the next module. In pseudocode:

1. $N = \text{argmax}_{\{gi,gj\}}\ cohes(\{g_i, g_j\})$
2. Loop:
3. $\quad g = \text{argmax}_g\ cohes(N \cup \{g\})$
4. $\quad$ If $cohes(N \cup \{g\}) < 1$, stop
5. $\quad N = N \cup \{g\}$
6. If $|N| > 2$, output N
7. Let $\bar{w}$ be the average edge weight among nodes in N
8. For each $g_i, g_j \in N$
9. $\quad w(g_i, g_j) = w(g_i, g_j) - \bar{w}$
10. Repeat from 1

Results

By analyzing functional associations among biological processes and functional similarities between high-throughput datasets in a purely data-driven manner, we summarize knowledge from thousands of whole-genome experiments in a biologically informative way. This includes descriptions of the cohesiveness, data coverage, and associations of biological processes (Figure 42), which can guide experimenters towards promising targets for future experimental work (Table 9). Datasets can also be compared based on functional activity, allowing the detection of large-scale functional similarity between the effects of experimental perturbations (Figure 43 and Figure 44). These analyses provide an important global summary of interplay between pathways,

and they identify processes, process associations, and dataset similarities likely to benefit from experimental investigation.

Discovering data-driven functional associations between biological processes

Two or more biological processes can interact and work together to perform cellular functions in a manner analogous to a relationship between individual genes. A pair of genes might be functionally related if they operate in the same complex, pathway, or transcriptional module. Our focus is at a higher level, where two processes might be functionally associated if they interact to achieve the same cellular goals; for example, nutrient sensing and the translation of new proteins at the ribosomes are distinct processes, but they interact to allow controlled cellular growth. These process-process associations are thus an extension of gene functional relationships: processes are functionally associated if they achieve related cellular goals, and we predict such an association if their constituent genes behave similarly in datasets determined to be good functional indicators. A small segment of our predicted process association network appears in Figure 42, made up of only the most confidently associated biological processes.

The edges in this process association network summarize information regarding the interactions between biological processes. A single biological process is internally cohesive in the currently available experimental results if its constituent genes also show strong individual functional relationships. If most gene pairs within a process are confidently functionally related, that process is reflected well by the available data: its annotations are in agreement with measured cellular behavior. If gene pairs within a process are related with low confidence, it often indicates an area of biology where further experimentation or annotation efforts may be most beneficial. The cohesiveness of biological processes in Figure 42 is represented by node color, where more cohesive processes appear in brighter yellow.

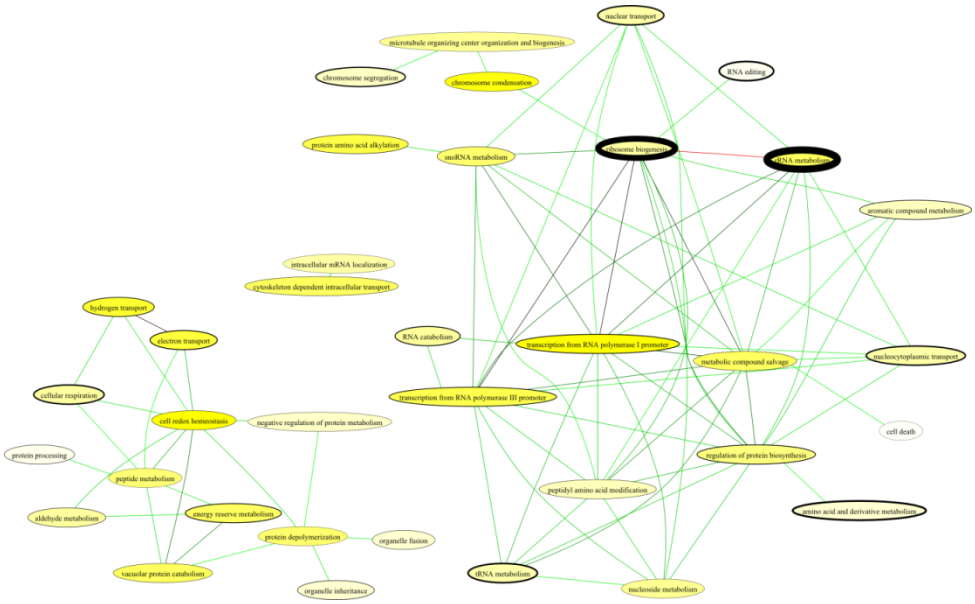


Figure 42: High-confidence associations between biological processes predicted from large-scale data integration. Each node represents a biological process extracted from the Gene Ontology; edges represent predicted functional associations between these terms based on their constituent genes' behavior in a compendium of >180 *S. cerevisiae* datasets. Node color intensity represents cohesiveness of the process, a measure of predicted relationship density within the process's gene set (white indicates background cohesiveness, yellow maximum cohesiveness). Border thickness summarizes estimated coverage of the biological process by available data. These edges represent only the strongest associations in the complete network, so coloration is relative, ranging from green (least strong) through black to red (strongest). Biological processes with high cohesiveness but low data coverage represent particularly promising targets for future experimental screens.

Finally, we also determined the degree to which each biological process is covered by available data. Our integration method provides a statistical measure of how active each biological function is within each dataset; we can thus sum over all datasets to estimate a biological

process's total representation within the data. This coverage measure is summarized by border width in Figure 42, with thicker borders indicating well-covered processes. Cohesive biological processes (yellow nodes) not covered well by available data thus represent promising candidates for future investigation: they show evidence of strong functional similarity, but they may not yet have been specifically targeted by high-throughput studies.

This interplay between functional associations, cohesiveness, and data coverage is evident in several of the example processes in Figure 42. *Ribosome biogenesis* and *rRNA metabolism*, for example, are processes strongly evident in most microarray data (Myers, Barrett et al. 2006), and this ubiquity is demonstrated by their extremely strong coverage and association. They are not as cohesive as many other processes, however, due to the large number of snRNAs and rRNAs annotated to these processes for which little or no high-throughput data is available. This analysis thus highlights an area for future exploration, even in an area as thoroughly studied as the ribosome. Other processes with relatively low coverage for their size (not shown in Figure 42) include *protein complex assembly*, *ion homeostasis and transport*, and *mitochondrion organization*, all representing opportunities for future directed screens. Processes with low cohesiveness can either be particularly diverse (e.g. *amino acid and derivative metabolism*, *protein processing*) or not yet fully characterized, representing further opportunities for future experimental investigations.

Processes predicted to be enriched for uncharacterized genes. Networks of functional associations between processes represent a richly structured summarization of high-throughput data; they implicitly encode predicted details regarding pathway structure, association between gene sets, and the functional diversity of currently available data. In addition to associating known processes and pathways, though, similar relationships can also be inferred to find areas of

biology enriched for uncharacterized genes. These represent specific processes for which targeted genomic screens might uncover substantial new information.

A selection of processes that we find to be highly associated with uncharacterized genes is shown in Table 9, in addition to statistics describing the processes (see Supplemental Table 8 for complete results). The *autophagy* term, despite being the smallest and most cohesive process in this subset, still maintains a very strong association with uncharacterized genes. It is moderately well covered by available data, falling roughly in the middle of our 141 coverage estimates; it is thus possible that further information regarding autophagy could be gleaned from existing data, even though few experiments have specifically investigated the process in yeast. However, this predicted association with uncharacterized genes also suggests that substantial new functional assignments could be made by targeted screens for involvement in autophagy.

Similar functional activity in high-throughput datasets

While most high-throughput experiments are designed with fairly specific goals in mind, almost every dataset contains information about a variety of biological processes, and our analysis provides several ways of exploring these data. Our Bayesian learning process results in a probabilistic score indicating the activity of each biological process within each dataset. Collecting all such scores for a single dataset results in a functional profile for the dataset, and these numerical vectors can be compared between datasets to evaluate functional similarity. The network in Figure 43 contains a selection of datasets with similar functional activities.

Process	Size (Genes)	Cohes.	Rel. Data Coverage	Assoc. with Unch. Genes
carbohydrate metabolism	233	2.09	3.75	972.1
phosphorus metabolism	201	1.95	2.35	895.3
reproductive physiological process	308	1.87	1.95	863.5
establishment of protein localization	279	1.82	1.77	862.0
sporulation	120	2.48	1.68	832.7
autophagy	40	3.69	1.22	797.6
one-carbon compound metabolism	94	1.94	2.57	794.9
cell wall organization and biogenesis	196	2.11	1.40	788.2
chromosome organization and biogenesis	557	1.96	4.53	773.1
cofactor metabolism	169	2.60	2.52	743.8

Table 9: Biological processes highly associated with yeast genes currently uncharacterized in the Gene Ontology. Association with uncharacterized genes is measured as the sum of predicted functional relationships between genes in a process and uncharacterized genes, normalized by the cohesiveness (and thus size) of the process. The cohesiveness of a process indicates the ratio of average in-process relationship weight to the average out-of-process relationship weight (with 1.0 thus the genomic background). Relative data coverage is a scaled sum of all datasets' predicted association weight with the given biological process. Because of their likely association with uncharacterized genes, these processes represent good candidates for future genomic screens.

Even in this small subset of analyzed datasets, several patterns are apparent. On the left, the first of the two main clusters contains primarily localization data from (Huh, Falvo et al. 2003). Within the localization subsets, dataset similarity is correlated with cellular localization: the periphery and bud are associated with the main body of data by way of actin, the Golgi stages are

associated with each other, the endosome and peroxisome are related, and so forth. Three synthetic genetic array screens are also similar to the localization data. (Davierwala, Haynes et al. 2005) is associated primarily with the Golgi and ER, and one of the primary findings of this study was the characterization of PGA1, a gene essential for ER activity. (Krogan, Kim et al. 2003) and (Zhao, Davey et al. 2005) show similar functional activity to a variety of localization subsets (including several not shown in Figure 43) and to (Krogan, Peng et al. 2004), all of which are enriched for nuclear functions (DNA packaging, chromosome organization, transcription, RNA elongation, etc.) These functional similarities were generated solely by automatic data mining and call out important biological associations between disparate experimental results.

On the right, the cluster of microarray data is centered around a core of large datasets exploring very diverse conditions and thus enriched for many different biological processes (Hughes, Marton et al. 2000; Brem, Yvert et al. 2002; Yvert, Brem et al. 2003; Brem and Kruglyak 2005). The other main components of the cluster are stationary phase growth and carbon metabolism (Ideker, Thorsson et al. 2001; Stuart, Segal et al. 2003; Martin, Demougin et al. 2004; Pitkanen, Torma et al. 2004; Brauer, Saldanha et al. 2005) and various stresses (Gasch, Spellman et al. 2000; Jelinsky, Estep et al. 2000; Bro, Regenberg et al. 2003; O'Rourke and Herskowitz 2004). Interestingly, (Chitikila, Huisinga et al. 2002), (Bulik, Olczak et al. 2003), and (Schawalder, Kabani et al. 2004) are all likely included due to their use of galactose-inducible promoters while investigating other diverse processes; these datasets all share a *carbohydrate metabolism* enrichment in addition to their more specific targets (e.g. *biopolymer biosynthesis*, a parent of *chitin biosynthesis*, in (Bulik, Olczak et al. 2003)). This demonstrates the power of associative functional analysis to uncover both primary and secondary enrichments, a consideration essential to getting the most out of any experimental result.

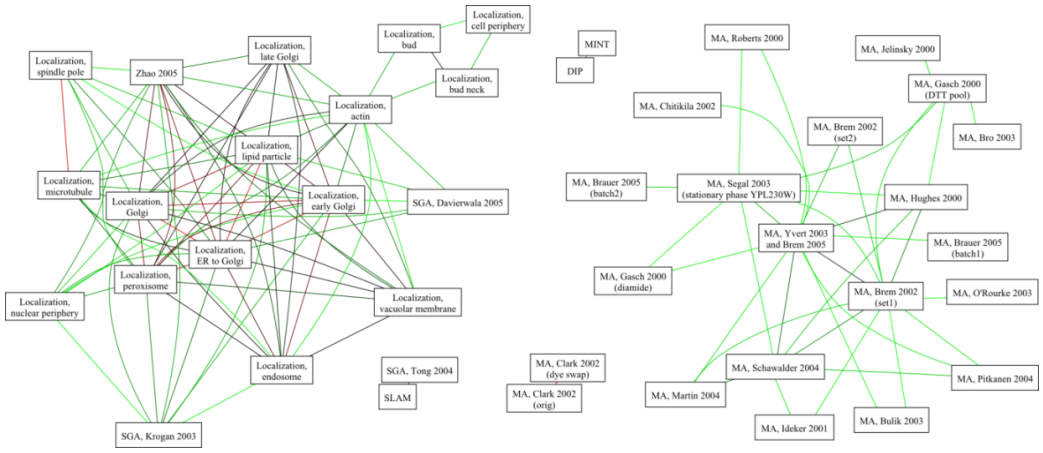


Figure 43: Similarities in functional activity between high-throughput datasets. Each node represents a dataset, each edge the correlation between two datasets' functional activity profiles. These edges represent only the strongest correlations (by Kendall's τ), so coloration is relative from green (least strong) to red (strongest). This associates collections of datasets that explore related areas of biology, either by specific experimental design (e.g. protein localization) or by provoking similar biological responses (e.g. the diauxic shift and stationary phase growth). This also confirms that multiple genetic (SLAM and Tong et. al. 2004) and physical (DIP and MINT) interaction collections offer similar functional coverage.

Simultaneous association of datasets and biological processes

Because our method assesses functional activity within datasets, functional similarities between datasets, and associations between biological functions, it provides a means of coclustering datasets and processes in a biologically meaningful way. This raises the possibility of exploring complex data, potentially summarizing millions of individual measurements, in an intuitive manner. Each predicted weight between two datasets, two processes, or a dataset and a process represents a measure of similar biological function, and thus an investigation of heavily weighted subgraphs in this space provides a way of exploring groups of related data and processes.

An example of such a cluster appears in Figure 44, which highlights one of the densest functional areas and the datasets in which these functions are most active. This consists of metabolic processes including *alcohol*, *aldehyde*, and *carbohydrate metabolism*, *cellular respiration*, *hydrogen* and *electron transport*, and *mitochondrion biogenesis*; while they have been removed for visual clarity, several other related processes are also members of this cluster, including *cofactor metabolism*, *autophagy*, and *aging*. The group of associated microarrays again represent a combination of broad genomic response (Yvert, Brem et al. 2003; Brem and Kruglyak 2005), carbon metabolism (Segal, Shapira et al. 2003; Schawalder, Kabani et al. 2004), and stresses (Gasch, Spellman et al. 2000), the latter likely included due to the relationship between stress response and growth rate (Brauer, Huttenhower et al. 2008). These are linked into the cluster of biological processes primarily through *carbohydrate metabolism*, but also through the biclustering modules (PISA). These biclustering results incorporate all of the available microarray conditions, in contrast to the normalized correlation scores used to analyze individual datasets. Biclustering thus represents a view of expression data orthogonal to pairwise correlations and tends to be more sensitive to metabolic functions in general (*phosphorus*, *amino acid*, and *nitrogen compound metabolism* in addition to those appearing in Figure 44).

The non-microarray datasets associated with this functional cluster are diverse, including mitochondrial localization (in association with several mitochondrial and respiratory functions), cytoplasmic localization (in association with more general metabolism), two sequence-based analyses (downstream sequence similarity and shared transcription factor binding sites from (Harbison, Gordon et al. 2004)), and synthetic lethality interaction profiles from GRID (Stark, Breitkreutz et al. 2006) and BIND (Alfarano, Andrade et al. 2005). Synthetic lethality profiles and shared binding sites both provide good coverage of many biological processes and are included

largely due to moderate association with many of the functions within the cluster (most edges are not shown in Figure 44); this is reflected in their relative isolation in the network. Broad downstream (and upstream) sequence similarity tends to capture structural features of the genome, in this case the close positional association of the GAL genes.

A case study: detecting a specific biological response in diverse data

At a more specific level, these interprocess associations and functional descriptions of datasets can be used to uncover detailed biological responses in high-throughput data. We were struck by the correlation in functional activities between three seemingly diverse datasets: (Chitikila, Huisinga et al. 2002), an investigation of TBP inhibitors, (Martin, Demougin et al. 2004), an analysis of *tor2* mutants described in (Helliwell, Howald et al. 1998), and (Pitkanen, Torma et al. 2004), a *pmi40* deletion assayed over varying mannose concentrations. These three microarray collections share functional enrichments with other datasets assaying similar conditions (e.g. the nutritional cluster discussed above including (Martin, Demougin et al. 2004) and (Pitkanen, Torma et al. 2004)), and no one pair of the three correlations is unusually high. They also represent two different experimental platforms: (Martin, Demougin et al. 2004) and (Pitkanen, Torma et al. 2004) both employ single channel microarrays, while (Chitikila, Huisinga et al. 2002) uses a two-color array. However, the average functional correlation between the three datasets is highly significant ($\overline{rel'} = 0.316$, $p<10^{-3}$) for arrays under such apparently diverse conditions.

All three datasets are enriched for activity in distinct biological processes, and all three present unique biological conclusions that are in no way undermined by this unexpected similarity. Upon inspection of the three datasets' experimental protocols, however, the common factor appears to be the use of a specific plasmid shuffle transformation employing a strain background of the form *ura3 trp1 leu2 his3* or *his4*. We have confirmed this similarity in a fourth dataset we are

currently developing investigating temperature-sensitive *dbf4* mutants (Myers, Robson et al. 2005). Although the overarching biological conditions of our dataset share little in common with (Chitikila, Huisinga et al. 2002), (Martin, Demougin et al. 2004), and (Pitkanen, Torma et al. 2004), our mutants were also constructed using a similar plasmid transformation, and the resulting microarrays produce highly correlated functional profiles. Even when strain background and reference channels (when applicable) are all properly controlled, the plasmid shuffle process and associated auxotrophies result in subtle changes in global transcription detectable by large-scale functional analysis.

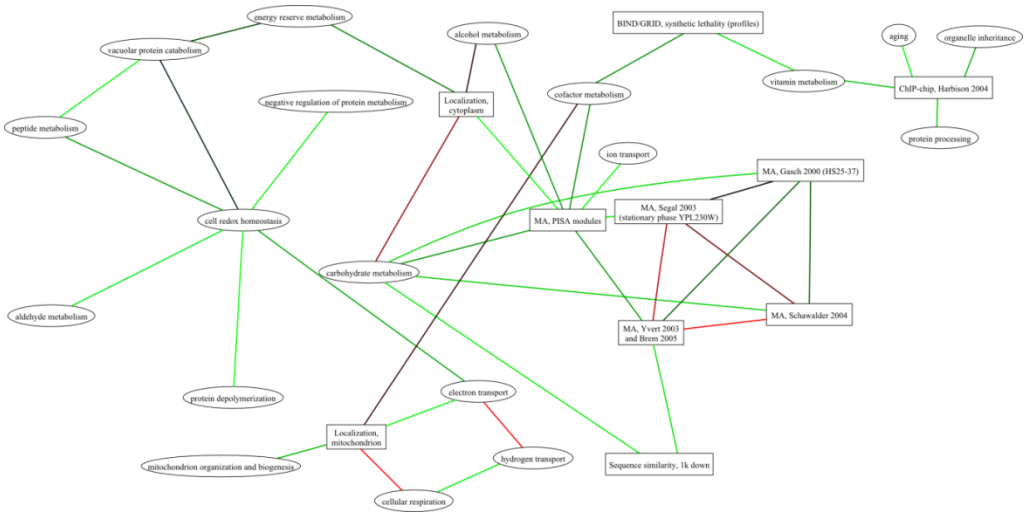


Figure 44: Coclustering datasets and biological processes in an area of dense functional associations. By mining associations between biological processes for dense subgraphs, we recover a collection of processes (rectangular nodes) predicted to be highly related based solely on experimental data. We then extract the datasets (oval nodes) most informative for those processes and display the most confident process/process, dataset/dataset, and dataset/process associations among these nodes. Each edge type is individually weighted, and only the strongest edges are shown, ranging in weight from green (least strong) to red (strongest). This network thus represents a snapshot of one area of yeast biology, the interconnections among its constituent processes, and datasets exploring these processes.

This effect is quite subtle, a fact which we stress for two reasons. First, it is a secondary effect within the more prominent biological features assayed by these three datasets, and it is only by large-scale analysis of their functional content in the context of many other datasets that the similarity was discovered. Second, we emphasize that it in no way diminishes these datasets' primary results, and instead provides additional functional insight into their coexpression measurements. Most previous computational data integration has focused on associating genes with functions or genes with genes. As more high-throughput data becomes available, it opens up opportunities for associating entire datasets with broad functional activity and with other datasets, allowing the detection of biological signals and similarities that would remain undetectable at smaller scales.

Discussion

We present a high-level functional analysis of very large compendia of genomic data and apply it to *S. cerevisiae*. By computationally summarizing thousands of whole-genome experimental conditions, we elucidate the current data coverage of *S. cerevisiae* biological processes, the cohesiveness of its functional annotations, and associations among these processes based on high-throughput experimental results. We also determine the functional activity in high-throughput datasets, allowing us to discover subtle relation-ships such as shared strain backgrounds in otherwise diverse microarray conditions. This analysis begins with specific functional relationships between individual genes predicted from large-scale data integration, and it extends into high-level information including functional associations between datasets, uncharacterized genes, and biological processes.

A primary application of this system lies in directing future experimental efforts. In particular, high-throughput screens of any sort can be costly to implement and assay fairly general

conditions; for example, if two proteins bind only during fermentation, their interaction will not be observed in a genomic screen during respiratory growth. A high-level functional analysis serves to call out underrepresented biological processes and those with increased likelihoods of novel discovery, which can in turn provide focus for experimental screens. This is analogous to candidate gene selection at a whole-genome level, a form of "candidate process" selection, just as our predicted associations between biological processes represent functional relationships at a larger scale.

High-level functional analysis also provides very specific information on individual experimental results, in addition to its larger scale applications. This is exemplified by the functional signature of the plasmid shuffle strain discussed above; given any new high-throughput dataset, microarray or otherwise, we provide a means for establishing its functional activity in the context of existing data. Both this post-hoc analysis and the a priori predictions of underrepresented functions are of particular use in less well-studied organisms. By designing experiments to explore processes shown to lack functional coverage and by leveraging all available data to interpret new results, laboratory work can be quickly guided to areas of biological interest and potential.

Finally, the functional information summarized by our system can also be employed in the continuous process of functional cataloging. While we have used examples from the Gene Ontology, any sets of functionally related genes could drive analyses such as this, and the results can guide annotators in cataloging existing data much as they can guide experimenters in generating new data. By providing a means of directing annotators to potentially under-annotated functions and the datasets associated with them, our analysis simplifies a curation and cataloging task that grows with each new publication. By analyzing and presenting the large-

scale functional structure of genome-scale data, we hope to guide annotators and experimenters alike in exploring the potential of the ongoing genomic revolution.

HEFalMp: A Functional Map of the Human Genome

The completion of the Human Genome Project and the subsequent flood of genomic data and analyses have provided a wealth of information regarding the entire catalog of human genes. Comprehensive assays of gene expression, protein binding, genetic interactions, and regulatory relationships all provide snapshots of molecular activity in specific cell types and environments, but turning these biomolecular parts lists into an understanding of pathways, processes, and systems biology has proven to be a challenging task. This abundance of data can sometimes obscure biological truths: the size of the human genome, the complexity of human tissue types and regulatory mechanisms, and the sheer amount of available data all contribute to the analytical complexity of understanding human functional genomics.

In order to take advantage of large collections of genomic data, they must be integrated, summarized, and presented in a biologically informative manner. We provide a means of mining tens of thousands of whole-genome experiments by way of functional maps. Each map represents a body of data, probabilistically weighted and integrated, focused on a particular biological question. These questions can include, for example, the function of a gene, the relationship between two pathways, or the processes disrupted in a genetic disorder. Functional integrations investigating individual genes' relationships have been successful with smaller data collections in less complex organisms (Lee, Date et al. 2004; Date and Stoeckert 2006; Myers and Troyanskaya 2007), although (as discussed below) it is particularly challenging to scale these

techniques up to the size and complexity of the human genome. Each functional map, based on an underlying predicted interaction network, summarizes an entire collection of genomic experimental results in a biologically meaningful way.

While functional maps can readily predict functions for uncharacterized genes (Murali, Wu et al. 2006), it is important to take advantage of the scale of available data to understand entire pathways and processes. Cross-talk and co-regulation among pathways, processes, and genetic disorders can be mapped by analyzing the structure of underlying functional relationship networks. This includes the association of disease genes with (potentially causative) pathways; for example, many known breast cancer genes are involved in aspects of the cell cycle and DNA repair, and novel associations of this type can be mined from high-throughput data. Similarly, associations between distinct but interacting biological processes (e.g. mitosis and DNA replication) can be quantified by examining functional relationships between groups of genes, allowing the identification of proteins key to interprocess regulation.

The functional maps we provide for the human genome include information on protein function, associations between diseases, genes, and pathways, and cross-talk between biological processes. These are all based on probabilistic data integration using regularized naive Bayesian classifiers. While naive Bayesian systems have been used successfully to analyze protein-protein interaction data (Rhodes, Tomlins et al. 2005; von Mering, Jensen et al. 2007) and to perform functional integration in simpler organisms with smaller data collections (Date and Stoeckert 2006; Myers and Troyanskaya 2007), they have not previously been scaled to provide a functional view of the human genome driven purely by experimental results. In addition to challenges of computational efficiency in the presence of hundreds of genome-scale datasets, naive classifiers assume that all input datasets are independent; this becomes increasingly untrue and problematic as more

datasets are analyzed, resulting in a paradox of decreasing performance with increasing training data. To address this, we use Bayesian regularization (Steck and Jaakkola 2002), a process by which an observed distribution of data can be combined with a prior belief in a principled manner. Intuitively, this results in groups of datasets containing similar information making a more modest contribution to the integration process, upweights unique datasets, and prevents overconfident predictions. Our regularization of the naive classifier parameters using a score based on mutual information up- and down-weighted appropriate subsets of data, maintaining both efficiency and accuracy.

While Bayesian regularization enables the prediction of functional relationship networks from very large genomic data collections, functional mapping further analyzes these networks to answer specific biological questions. Naive Bayesian classifiers alone can accurately weight individual experimental results, e.g. high microarray correlation or physical protein binding, with respect to their ability to predict functional relationships. Regularization adds the ability to weight entire datasets with respect to their predictive power, and both of these weightings can be applied in a process-specific manner (e.g. a microarray dataset may be highly predictive for transcriptionally regulated processes but cannot detect post-translational modifications). Thus, regularized Bayesian integration summarizes billions of data as millions of predicted relationships; this is clearly still too large a body of data to explore directly. Functional mapping solves this problem by further summarizing sets of functional relationships into specific predicted associations between genes, pathways, or diseases. For example, two related processes such as DNA synthesis and the cell cycle are likely to share substantial regulatory cross-talk, i.e. the distribution of predicted functional relationships spanning genes in the two processes will be significantly different from that expected by chance. Functional mapping can statistically identify

such associations based solely on genomic data, predict specific proteins serving as potential regulatory hubs, and highlight the experimental data underlying these predictions.

We applied our functional maps to a specific biological question in the area of autophagy, the process by which a cell can recycle its own biomass under conditions of starvation or stress (Klionsky 2007). Among many proteins predicted to participate in this biological process by an early version of our maps, we chose to investigate LAMP2 and RAB11A in the laboratory. We demonstrated through multiple lines of experimental evidence that these proteins are indeed involved in macroautophagy in amino acid-starved human fibroblasts, a specific type of autophagy in which bulk cytoplasm is lysosomally degraded. The results of our integration are available through a web-based interface, HEFalMp (Human Experimental/Functional Mapper), at http://function.princeton.edu/hefalmp. This tool allows a user to interactively explore functional maps integrating evidence from thousands of genomic experiments, focusing as desired on specific genes, processes, or diseases of interest.

Results

Using the system outlined in Figure 45A, we generate functional maps of predicted gene functions, pathway and process associations, and genetic disorders in 229 biological areas, incorporating information from ~30,000 genome-scale experiments. Within each biological area, maps are derived from a functional relationship network predicted using regularized Bayesian integration of the genomic data. The features and contents of the resulting interaction networks are analyzed to produce gene-, process-, and disease-centric functional maps specific to each biological area. We have experimentally confirmed two genes newly predicted to be active in the area of macroautophagy, LAMP2 and RAB11A.

Data integration for functional mapping

A functional map is a view of genomic data focused on a particular area of interest: genes, processes, diseases, and their associations and interrelationships. To derive these maps, we analyze functional relationship networks predicted based on Bayesian integration of ~30,000 genome-scale experiments. These are organized into 656 datasets (grouped by related microarray experiments, individual interaction databases, and so forth) and probabilistically weighted based on their functional activity in 229 biological areas of interest (e.g. *autophagy*, *mitotic cell cycle*, *protein processing*, etc.). As summarized in Table 10, one product of this integration process is an estimate of the biological processes active in each dataset. Further, as highlighted in Table 11, over 25% of our predicted functional relationships are supported by at least 100 datasets, and many genes' predictions include information from over 500 genome-scale datasets.

Using only the information in these predicted functional relationship networks before they have been further processed into functional maps, we can accurately recapitulate known biology from catalogs such as the Gene Ontology (Figure 45). As observed in (Myers and Troyanskaya 2007), functional integration benefits substantially from context-awareness, a fact we take advantage of in our use of process-specific functional maps. Performance differs only slightly between an evaluation of the entire genome and of a held-out test set, demonstrating naive classifiers' robustness to overfitting. Most significantly, Bayesian regularization provides a dramatic increase in performance by downweighting groups of similar datasets and upweighting unique, informative datasets in each biological process.

Features of the functional relationship networks

Functional maps are generated by analysis of functional relationship networks, and each network is based on probabilistic integration of genomic data within a particular biological area. In

addition to providing maps of higher-order associations among processes and diseases, these functional relationship networks can be examined directly to provide insights into protein function, functional modules, and characteristics of the integrated experimental data. Table 11 presents summary statistics for several of the networks we analyzed. A substantial fraction (26%) of the networks' edges are supported by evidence from more than 100 datasets, and ~10,000 edges are supported by over 500 datasets. There is strong variation in probabilities and dataset weighting between biological processes, with the most confident coverage offered by reintegration across all available processes. While different genes tend to be highly connected in each process-specific network, commonalities emerge in the global networks and interprocess averages. These proteins (HNF4A, RUNX2, GHRHR, and others from the rightmost table column) tend to be components of complexes or receptors; they are thus predicted to have a relatively small number of extremely confident relationships with their other complex members or ligands. This is confirmed by the fact that these genes are also among the most variable, although their predictions are not generally supported by the most datasets. Instead, to find these particular relationships, subsets of appropriately reliable data are upweighted by our integration system in a process-specific manner.

Individual functional relationship networks can also be used to predict protein function using "guilt by association," as diagrammed in Figure 46A. If a gene has many strong, specific predicted relationships with genes in a particular biological process, it is itself likely to participate in that process. ALOX5AP, for example, is a membrane protein required to activate ALOX5 for leukotriene synthesis; this pathway is a clinical target for the treatment of asthma, heart disease, and obesity (Peters-Golden and Brock 2003; Mehrabian, Allayee et al. 2005). Our integration

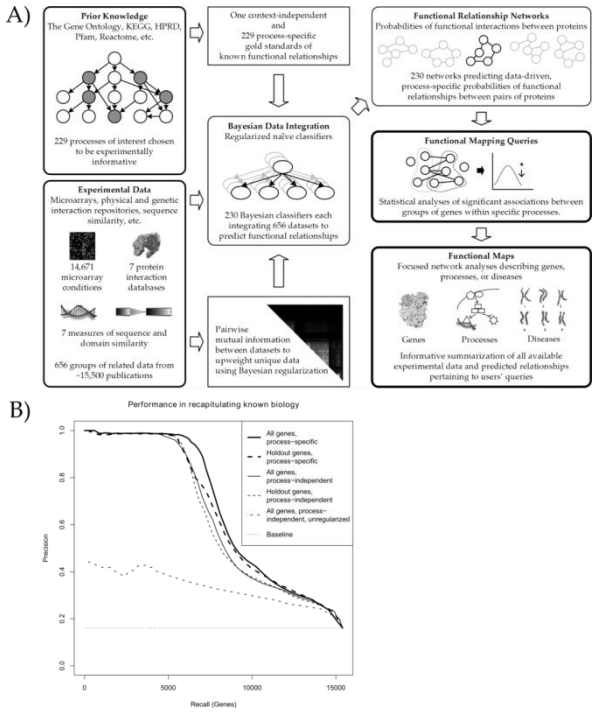


Figure 45: Overview and performance of genomic data integration for functional mapping. A) Data from ~30,000 genome-scale experiments (~15,000 microarray conditions and ~15,000 interaction and sequence-based assays) were organized into 656 related datasets. These datasets were used as inputs for 229 process-specific naive Bayesian classifiers each trained to predict functional relationships specific to a particular biological area and one process-independent global classifier. Mutual information was calculated between each pair of datasets and used to regularize these classifiers and prevent overconfident predictions. Each classifier was used to infer a predicted functional relationship network for a particular biological process. These networks were then analyzed to find statistically significant sets of functional relationships spanning gene groups of interest. This results in functional maps focusing on individual genes, groups of genes, biological processes, or genetic disorders. Each map provides an informative summarization of the genomic data collection focused on the current biological entity of interest. B) Performance of predicted functional relationship networks in recapitulating known biology. To confirm that the predicted functional relationships underlying our functional maps were accurate, we scored their ability to recover information from a held-out portion (25% of genes) of our gold standard. This evaluation includes the global process-independent network tested on all genes and the holdout set, a global mean of the process-specific networks tested on all genes and the holdout set, and an unregularized global process-independent network tested on all genes. Precision is well above baseline, and since naive classifiers are generally robust to overfitting, performance of the holdout set is only slightly below that of the entire genome. Bayesian regularization provides a large performance increase at low recall by preventing overconfident predictions.

	Data points	Sets	Pubs.	Conds.	Mean max. posterior	Mean norm. weight	Most informative functional areas
Interactions (physical and genetic)	11,244,053	14	>15,000	>15,000	0.375	0.000286	Response to DNA damage, membrane potential, regulation of cell cycle, cell death, DNA metabolism
Sequence comparisons (nucleotide and protein)	452,199,430	7	6	NA	0.162	0.00197	Cell adhesion, cell surface receptor signal transduction, phosphorus metabolism, chromosome organization
Microarrays	27,248,177,875	635	417	14,671	0.0270	0.000606	Cell surface receptor signal transduction, cell adhesion, RNA splicing and metabolism, ion transport
All data	27,711,621,358	656	>15,500	~30,000	0.0378	0.000619	

Table 10: Summary of integrated genomic data. 21 interaction and sequence-based datasets were assembled from various sources consolidating >15,000 publications; 635 microarray datasets spanning >14,000 conditions were downloaded from GEO (Barrett, Suzek et al. 2005). The mean maximum posterior and normalized weights are calculated across the 229 analyzed functional areas. Particularly active functional areas are determined for each dataset based on the weight given to the data by each process-specific classifier; microarrays, for example, are particularly good at detecting the strong transcriptional signals of RNA processing and co-complexed proteins such as ATP synthases. While genetic and physical interactions are generally the most reliable data sources, they are also the least common. This results in them being given a high weight (posterior) during Bayesian integration, but when this weight is normalized by the amount of available data (prior probability), sequence-based data (shared protein domains, transcription factor binding sites, etc.) are found to provide the best balance between coverage and informativity.

system predicts it to have many specific functional relationships with other membrane proteins involved in the inflammatory chemotaxis response in leukocytes (among other predicted relationships). While neither ALOX5AP nor ALOX5 are annotated to a chemotactic pathway in

the Gene Ontology, one of their immediate biosynthetic products, LTB4, is a known activator of chemotaxis (Peters-Golden and Brock 2003). This is an example of uncovering an uncataloged protein function by functional mapping, and we provide details below of our experimental confirmation of novel predicted functions for LAMP2 and RAB11A in autophagy.

	Ave. rel. conf.	Rel. above prior	High-conf. rel.	Genes with >10 high-conf. rel.	Most connected genes	Most variable genes
Global (process-independent)	0.0381 (0.117)	60,189,940	51,890	2,278	RUNX2, PRLR, GHRHR, ATP2B2, OPRM1	RUNX2, GHRHR, OPRM1, PRLR, ATP2B2
Representative processes						
Autophagy (20 genes)	0.000561 (0.0113)	30,054,992	5,981	234	CDK4, SUMO1, PPM1G, HPRT1, HINT1	CDK4, PPM1G, SUMO1, RAN, HINT1
Chemotaxis (137 genes)	0.0103 (0.0644)	42,265,832	137,957	3,784	GHRHR, HTR4, FSHR, SERPINA4, OPRM1	GHRHR, HTR4, FSHR, SERPINA4, MLN
Cell death (724 genes)	0.00968 (0.0313)	17,919,145	9,818	348	KPNB1, HNRPK, VEGFA, MSH2, HNRPA2B1	HNF4A, GRB2, KPNB1, TP53, YWHAZ
Average across individual processes	0.0111 (0.0186)	42,515,815 (34,336,803)	66,663 (126,498)	1,135 (1,179)	HNF4A, GHRHR, FSHR, HTR4, RUNX2	GHRHR, HTR4, OPRM1, HTR6, ADRA1A
Global (process-aware)	-0.001570 (0.444)	NA	1,871,380	11,614	HNF4A, COPS5, VBP1, DDX1, PSMD14	TP53, GRB2, PCNA, COPS5, HDAC1
Nodes (genes)		**Edges (rel.)**		**Rel. with >100 datasets**		**Rel. with >500 datasets**
24,433		298,473,528		78,519,235		10,317

Table 11: Features of functional relationship networks predicted from data integration. We inferred 231 networks predicting functional relationships among 24,433 human genes. 229 of these are process-specific and provide interaction probabilities within a particular functional area. The remaining two are global (non-process-specific) and indicate probabilities of functional relationship either without consideration for biological process (process-independent) or as a normalized average across all processes (process-aware), respectively. The 229 process-specific networks and the global non-process-specific network consist of probabilities in which the threshhold for high confidence was 0.95. The global integrated-process network is normalized to contain z-scores, which can be negative, and an equivalent high-confidence threshhold was set at 2.0. Data for three representative processes of varying sizes are shown, in addition to averages across process-specific networks (standard deviations in parentheses). As detailed in (Huttenhower, Hibbs et al. 2006), reintegration across processes produces a more confident and reliable global network than is obtained from ignoring process-specificity. Many predicted relationships are supported by several hundred datasets, and protein interactions vary strikingly between biological areas as they participate in different pathways and processes.

By extracting highly connected clusters from functional relationship networks, we can also discover putative functional modules showing high similarity in experimental data without being directly associated with pre-annotated gene sets or processes. These modules may represent novel pathways, complexes, or other groups of proteins interacting to carry out cellular tasks. The modules can be merged to create a hierarchical structure reminiscent of catalogs such as the Gene Ontology; a small subset of our predicted functional modules appears in Figure 46B. The most specific module in the hierarchy links the transcriptional regulators PIAS3, MITF, and PAX6 with very strong evidence drawn from multiple direct binding assays in the BioGRID (Stark, Breitkreutz et al. 2006). This module has two main branches of more general parents in the hierarchy. The first contains several cell growth, death, and differentiation transcriptional modulators, including JUN, NFKB1, and BCL3. The second contains multiple cell cycle related oncogenes, oncogene activators, and TGF-β family mediators, almost all of which are also transcriptional modulators (Kim, Wang et al. 2000). This is likely indicative of two interrelated regulatory programs, the former focused on cell development and differentiation and the latter responding more specifically to extracellular signaling. We have automatically mined and hierarchically organized ~17,000 functional modules from our integrated data, spanning all ~25,000 genes for which we have data and ranging in size from three to 5,600 genes.

Functional associations: genetic disorders and biological processes

By examining the behavior of entire pathways in integrated genomic data, we derive functional maps of cross-talk between related biological processes (Figure 47A). Just as functional relationships between genes are predicted by finding significant agreement among many integrated datasets, functional associations between processes are discovered by observing strong relationships among many of their constituent genes, based on similar behavior of the processes'

genes in many genomic data sources and not on prior knowledge of genes shared between processes.

For example, if we focus on the process of *cell fate commitment*, we predict associations with many specific processes of cell differentiation and development. Several of these associations are driven by proteins known to be involved in multiple processes, e.g. the association with *gastrulation* involves many shared genes including TGFB2, BMP4, TBX6, and TRIM15. On the other hand, an apparently similar association with *axis specification* is driven mainly by genes not yet cataloged as involved in a cell fate decision (e.g. TDGF1, T, MDFI, etc.) Maps associating interrelated biological processes (and detailing the proteins predicted to drive those associations) can be derived from high-throughput data for any biological area of interest. This provides a way of exploring pathway cross-talk in genomic data and quickly identifying potential regulatory hubs.

In a similar manner, groups of known disease-related genes can be associated with each other or with (potentially causative) pathways and processes. An example in Figure 47B focuses on ovarian cancer, currently recorded in OMIM (OMIM 2008) as being influenced by at least seven genes. While known shared genes drive some of these associations (e.g. MSH6 in *aging* or ERBB2 in *epithelial cell proliferation*), others are more surprising. For example, AKT1, a protein known to contribute to ovarian cancer, is predicted to be related to B3GNTL1 and PHKG2 in *biopolymer biosynthesis* (i.e. DNA synthesis) due mainly to high microarray correlation across a wide variety of conditions; these proteins are also involved in the estrogen and insulin pathways, respectively, which have been observed to interact (Hamelers and Steenbergh 2003). Similarly, while there is a growing understanding of the link between breast and ovarian cancer and hormone stimulus (Dumeaux, Fournier et al. 2005), we predict explicit molecular connections driven by LYN, EIF2B5, and MMS19L. We also observe links between ovarian cancer and other cancers, including

breast cancer, osteosarcoma, colorectal cancer, and hepatocellular carcinoma, mainly due to interactions or high microarray correlation with BRCA1, MSH6, and other known cancer-related proteins. Functional mapping can thus call out potentially overlooked associations between diseases as well as posit new molecular connections between biological processes and genetic disorders.

Finally, if an investigator has a specific biological hypothesis in mind, it can be explored using functional mapping of user-provided gene sets. Figure 47C demonstrates a query of known autophagy genes ATG7, BECN1, and MAP1LC3B with test genes LAMP2, RAB11A, and VAMP7 in the context of *autophagy*. This produces two clear clusters, a group of known autophagy genes related to a group of vesicular and transport genes (including the three test genes). These two clusters are associated primarily by RAB11A/BECN1, CLTC/BECN1, ARPC5/CLN3, and SH3GLB1/MAP1LC3B relationships, as well as less heavily weighted links through DPM1 and PSMC2. The four primary relationships are driven by a wide variety of microarray correlations, led by datasets investigating retinal pigment epithelium (Tian, Ishibashi et al. 2004), macrophage infection (Detweiler, Cunanan et al. 2001), bone marrow (Graf, Iwata et al. 2002), and DNA damage (Rieger and Chu 2004). The secondary relationships are also predicted based on diverse microarray data and information from the GSEA gene sets (Subramanian, Tamayo et al. 2005). All of these genes are known to be involved in ER/Golgi trafficking, the secretory and vesicular system, and protein degradation; these associations led us to investigate whether LAMP2, RAB11A, and VAMP7 play roles in the specific activation of macroautophagy. Our experimental confirmation of two of these predictions is detailed below.

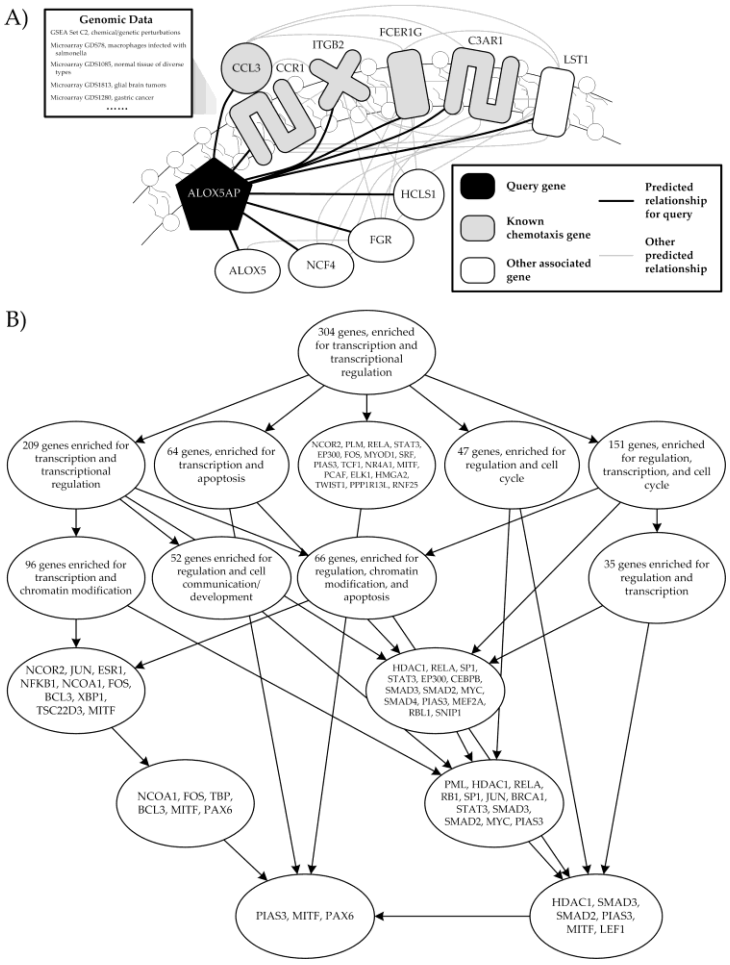


Figure 46: Analyses of functional relationship networks predicted from data integration. The process-specific functional relationship networks underlying functional maps can themselves provide information on individual genes' and modules' behavior in the underlying genomic data. A) Focusing on ALOX5AP, a membrane protein participating in leukotriene synthesis, highlights a predicted association with the process of chemotaxis in leukocytes, driven by multiple predicted relationships with known chemotaxis proteins. While ALOX5AP has not been formally cataloged as participating in chemotaxis, its immediate biosynthetic product LTB4 is a known activator of chemotaxis (Peters-Golden and Brock 2003). B) A subset of the functional modules predicted by mining highly-connected clusters from functional relationship networks. Each module consists of genes predicted to be related based on multiple informative genomic datasets. Here, a specific module consisting of PIAS3, MITF, and PAX6 generalizes through two main branches into modules enriched for various transcriptional regulation activities in the cell cycle, apoptosis, and intercellullar signaling. We have automatically mined and hierarchically organized ~17,000 functional modules of varying specificities from our integrated data.

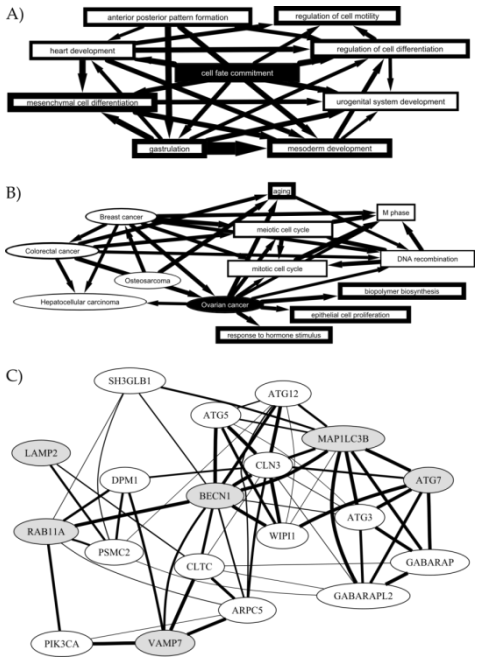


Figure 47: Results of functional mapping. Functional maps derived from experimental data integration provide information on groups of genes, including cross-talk between pathways, processes, and genes associated with genetic disorders. In all figure parts, thicker arrows indicate stronger associations, and directed arrows point to the gene group in which the background connectivity was calculated. A) Associations between biological processes derived by functional mapping. A focus on the process of cell fate commitment predicts it to be associated with a cluster of cell development and differentiation processes. Arrow width indicates the strength of predicted association, and border thickness indicates the internal cohesiveness of each process in the integrated genomic data. These predicted associations are based on a combination of proteins known to participate in multiple processes and novel data-driven predicted relationships. B) Associations between genetic disorders and biological processes. Focusing on ovarian cancer, known to be influenced by at least seven genes (OMIM 2008), we predict associations with the cell cycle, cell proliferation, and hormone stimulus, as well as with several other cancers. These associations are each based on relationships among individual genes predicted from integrated genomic data. C) Visualization of a functional map generated by querying a custom gene set. We chose to focus on the known autophagy proteins ATG7, BECN1, and MAP1LC3B, in addition to genes of interest LAMP2, RAB11A, and VAMP7, in the context of autophagy. This extracts two clear clusters of predicted autophagy-specific functional relationships, one consisting mainly of known autophagy proteins and one enriched for ER/Golgi and vesicular trafficking proteins (including the three test genes). This led us to experimentally test and confirm the hypothesis that LAMP2 and RAB11A are involved in macroautophagy in amino acid-starved human fibroblasts.

LAMP2 and RAB11A are required for macroautophagy in human fibroblasts

Autophagy is the process by which cells can consume their own biomass in order to survive when starved or otherwise stressed. Particularly in human biology, it is an area of active research, with recent work discovering links to tumorigenesis and bacterial infection (Klionsky 2007). Specifically, macroautophagy is the process of engulfing and degrading the contents of bulk cytoplasm, while chaperone-mediated autophagy and microautophagy employ different mechanisms to target specific proteins to the lysosome (Yorimitsu and Klionsky 2005). We will use the terms autophagy and macroautophagy interchangeably, as we focus here only on macroautophagy. Using an early version of our functional maps, we chose to experimentally investigate three proteins predicted to function in autophagy: LAMP2, RAB11A, and VAMP7. Previous work has shown these proteins to be involved in the lysosome and vesicular trafficking (Chen, Pan et al. 1985; Prekeris, Klumperman et al. 2000; Ward, Pevsner et al. 2000), with LAMP2 playing a known role in chaperone-mediated autophagy (Cuervo and Dice 1996), but they have not been specifically associated with macroautophagy. Punctate localization of the MAP1LC3 protein to autophagy-specific vesicles known as autophagosomes and its cleavage from the MAP1LC3-I to the MAP1LC3-II isoform are common markers for cells undergoing autophagy; both of these markers are obviated by the inhibition of proteins necessary for autophagy, e.g. ATG5 (Kabeya, Mizushima et al. 2000; Mizushima, Yamamoto et al. 2004). We found these markers to be decreased in primary human fibroblasts in which LAMP2 or RAB11A have been knocked down by siRNA, suggesting that these two proteins are required for successful autophagy (Figure 48).

LAMP2 and RAB11A depletions both significantly diminish autophagy as measured by quantification of fluorescent GFP-tagged MAP1LC3 (Figure 48A). Automated image analysis

using CellProfiler (Carpenter, Jones et al. 2006) demonstrates a significant drop in MAP1LC3 fluorescence (and thus autophagosome formation) under starvation conditions when ATG5 (positive control), LAMP2, or RAB11A levels are depleted by siRNA. In addition to overall decreased levels of fluorescence, both proteins' depletions also specifically abrogate the localization of MAP1LC3 to autophagosomes, as quantified by the number of fluorescent MAP1LC3 labeled puncta in a collection of 80 microscopic images (Figure 48B and C, Supplemental Figure 6). A VAMP7 knockdown showed no effect in any assay, which is possibly due to known variation in its behavior in different cell types; this is discussed in more detail below. The modest decrease in MAP1LC3-II incurred by the RAB11A knockdown (see Supplemental Figure 6), as opposed to its strong fluorescence and localization effect, raises the interesting possibility that it participates in the formation of autophagosomal membranes containing MAP1LC3 after it has been processed by ATG3 and ATG7 to the MAP1LC3-II isoform (Kabeya, Mizushima et al. 2004). Further investigation is necessary to determine the specific roles of LAMP2 and RAB11A in mammalian autophagy, but these assays provide strong evidence for their involvement in as predicted by functional mapping.

HEFalMp: a web-based interface for interactive functional mapping

Our functional maps can be explored interactively using the HEFalMp (Human Experimental/Functional Mapper) tool at http://function.princeton.edu/hefalmp. As shown in Figure 49, HEFalMp provides an interface through which a user can focus on a particular subject of interest - a gene, group of genes, biological process, or disease - and examine its predicted associations. For example, this can predict gene function (gene/process associations), cross-talk between pathways (process/process associations), or processes associated with genetic diseases,

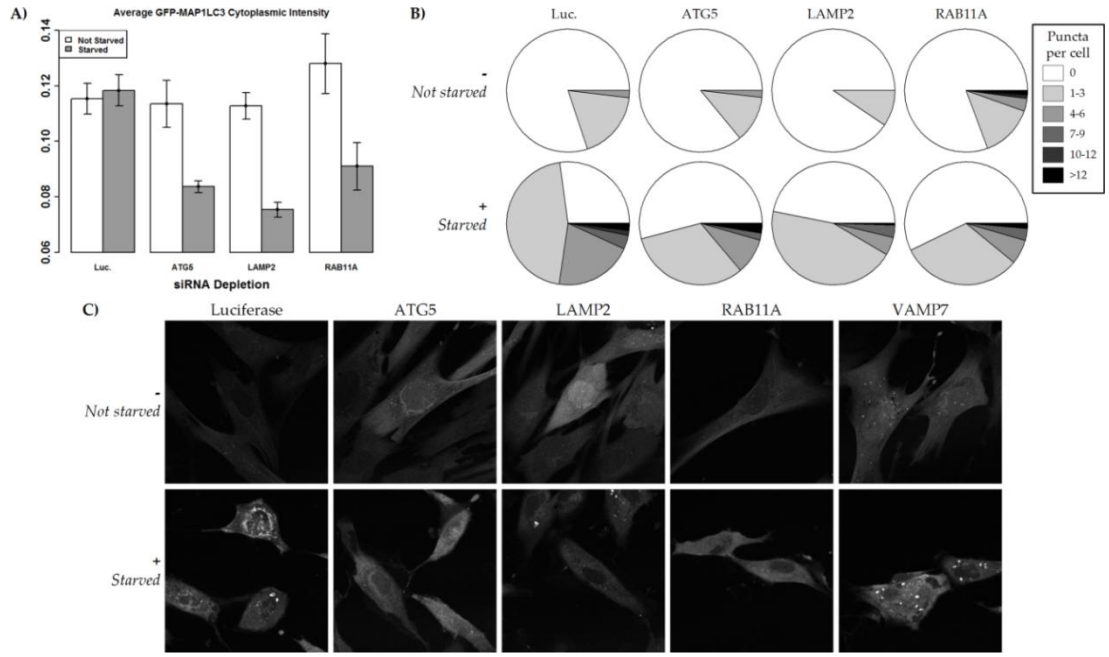


Figure 48: Impaired autophagosome formation confirms the predicted involvement of LAMP2 and RAB11A in human macroautophagy. An early version of our functional maps predicted LAMP2, RAB11A, and VAMP7 to be involved in autophagy, the process of recycling cellular biomass in order to survive under conditions of starvation or stress. While VAMP7 knockdowns showed no effect (see Discussion), siRNA knockdowns of LAMP2 and RAB11A inhibited normal autophagy. A) Automated image analysis detects a significant decrease in fluorescent GFP-tagged MAP1LC3 under starvation conditions for ATG5, LAMP2, or RAB11A knockdowns. Bars show standard error of average cytoplasmic intensity per cell as quantified by CellProfiler (Carpenter, Jones et al. 2006) over a collection of 10 images per condition (80 total). This decrease in fluorescence indicates that normal MAP1LC3-II processing (and thus autophagy) is impaired when ATG5, LAMP2, or RAB11A levels are reduced. B) Quantification of punctate autophagosome formation. The numbers of fluorescent puncta (MAP1LC3-II labeled autophagosomes) per cell were averaged over counts from three independent investigators in 10 images per normal (-) or starvation (+) condition, unlabeled and randomized (80 images total; see Supplemental Figure 2 for standard errors). The resulting distribution of puncta frequencies is low under all non-starved conditions and significantly increased under a negative control (luciferase) condition. It is only slightly increased for the ATG5 positive control and for the LAMP2 and RAB11A predictions. C) Punctate localization of fluorescent GFP-LC3 to the autophagosome during autophagy. Under normal conditions (-), MAP1LC3-I is localized diffusely through the cytoplasm; starvation (+) induces autophagy and localization to the autophagosome membrane. Knockdowns of ATG5 (positive control), LAMP2, and RAB11A abrogate this localization, indicating that these proteins are required for successful macroautophagy.

and all predictions can be made in any of the >200 biological areas for which we have constructed functional maps. A variety of visualizations are used for different query types, and all results can be downloaded for offline analysis. All predictions between groups of genes can be expanded into the specific functional relationships driving the analysis, and individual functional relationships can always be traced to the genomic datasets on which they are based. HEFalMp provides a convenient and informative way to explore functional maps summarizing data from ~30,000 genome-scale experiments.

Discussion

While the growing amount of publicly available genomic data can answer a wide variety of biological questions, usefully integrating, mining, and summarizing these data is an ongoing challenge. Using information from over 650 genome-scale datasets drawn from thousands of publications, we produce functional maps that provide specific information focused on an investigator's area of interest. This can include gene function, functional modules, cross-talk between pathways and processes, or interactions among genetic disorders. We have experimentally confirmed predicted involvements of RAB11A and LAMP2 in human macroautophagy, and we provide the HEFalMp web-based interface for biologists to explore our results and to generate new functional maps in their areas of interest.

Applications of functional mapping

Functional mapping can guide further laboratory and computational investigations by taking advantage of large collections of genomic data in a biologically meaningful way. As demonstrated by our confirmation of the participation of LAMP2 and RAB11A in autophagy, functional associations of individual genes with pathways and processes can be used to suggest directed laboratory experiments. In the area of human disease, this can be even more significant,

since functional mapping predicts associations of genetic disorders with potentially causative processes and with specific individual genes. It is key that computational methods take advantage of modern high-throughput biology to guide researchers to novel disease genes based on information from thousands of experimental results.

Functional mapping can further leverage high-throughput data to better inform functional cataloging and annotation efforts. As seen above with ALOX5AP, many human proteins have ample literature evidence to link them to established pathways and processes but have not yet been fully annotated in catalogs such as GO or KEGG. Functional mapping can rapidly direct annotators to such under-annotated genes, providing an opportunity to substantially improve functional catalogs based on existing literature evidence.

Bayesian regularization enables very large scale data integration

It is notable that previous data integration techniques do not scale adequately to the size of the human genome and the amount of currently available genomic data. Bayesian structure learning has been applied successfully to very small groups of genes with focused datasets (Sachs, Perez et al. 2005), but its computational complexity makes it inapplicable on a whole-genome scale. Even TAN classifiers, which are only minimally more complex than naive networks, can be inefficient to learn from very large, incomplete data collections (Tian, Wang et al. 2005). While naive Bayesian classifiers can perform rapid data integration and can be learned and evaluated very quickly, their inherent independence assumption can produce overly confident predictions in the presence of many datasets (Supplemental Figure 7). In order to maintain accuracy when dealing with very large data collections, we use Bayesian parameter regularization (Steck and Jaakkola 2002) to assign a uniform prior to each dataset with belief inversely proportional to the

amount of unique data in the dataset. This allows particularly diverse, informative datasets to efficiently provide a stronger contribution to the integration and mapping process.

Mutual information, which we use to evaluate similarities between datasets when performing regularization, also reveals surprising large-scale structure in our collection of genomic data (Figure 50). While most datasets share very little information by an absolute measure, small but consistent patterns emerge when considering hundreds of datasets spanning thousands of experimental conditions. Since most available genome-scale data is expression based, microarray platform is one of the broadest factors by which datasets cluster. Within these large platform-based groups, other similarities are detectable based on a variety of factors ranging from tissue type to array normalization algorithm. It is striking that a straightforward data mining measure such as mutual information, when applied to a sufficiently large collection of genome-scale data, can discover various underlying classes of datasets. Even though the amount of information shared based on factors such as array platform is small, its ubiquity violates the independence assumption of naive classifiers, and it thus provides the basis for the performance improvement we observe when using regularized parameters.

Next steps: tissue specificity and temporal resolution

A variety of biological features and prior knowledge could be added to further improve functional mapping's integration of genomic data. Most significantly, tissue and cell type is a key aspect of metazoan biology that is not currently taken advantage of by our functional maps. This is perhaps evident in our investigation of VAMP7, a vesicle-association membrane protein known to show widely varying behaviors in different tissue types (Advani, Yang et al. 1999; Siddiqi, Mahan et al. 2006). It has characterized roles in the late endosome/lysosome, and our functional maps predict extensive relationships with other synaptosomal proteins, in agreement

with VAMP7's function in neuronal morphogenesis (Rossi, Banfield et al. 2004). While we found that decreasing the expression of VAMP7 in human fibroblasts did not detectably influence their induction of autophagy, it is possible that VAMP7 participates in autophagy in other cell or tissue types.

Similarly, just as many functional associations are cell-type specific, others are dependent on subcellular localization or on temporal characteristics (e.g. cell cycle phase). Our results, as well as previous work (Myers and Troyanskaya 2007), show that explicitly modeling functional relationships within individual biological processes significantly improves accuracy. Differences in cell type, localization, and temporal character represent equally significant cases in which the same proteins can carry out different functions. Incorporating information such as cell and tissue types is thus an important way in which the mapping process can be further developed in the future.

The features, diversity, and amount of genomic data will certainly continue to increase, and functional maps provide a flexible means by which this data can be informatively summarized and explored. By integrating over 650 datasets spanning thousands of experimental conditions, we have predicted functional relationship networks specific to a variety of individual biological processes. Mapping these networks allows an investigator to mine this data from several different perspectives, focusing on associations between genes, pathways, processes, or genetic disorders of interest. We have experimentally confirmed predicted participation of LAMP2 and RAB11A in the process of macroautophagy, demonstrating that functional mapping can accurately direct experiments to specific genes and functional areas. These predicted associations can be extended to any group of genes, e.g. allowing an experimenter to investigate novel

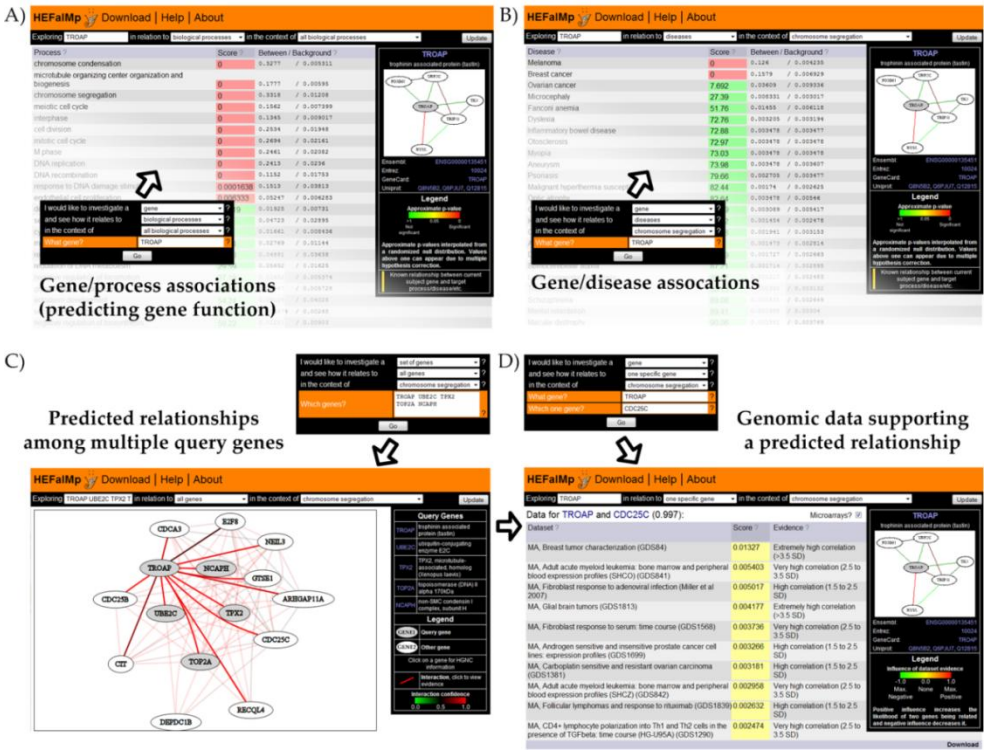


Figure 49: The HEFalMp tool for functional mapping. We have provided a web interface, the Human Experimental/Functional Map (HEFalMp), at http://function.princeton.edu/hefalmp for interactively exploring our predicted functional maps. A user can focus on a gene, gene set, biological process, or genetic disorder of interest and investigate its predicted associations with other genes, processes, or diseases. These predictions are presented using a variety of visualizations, and all data is downloadable for further analysis. A) Associating a gene with biological processes. An investigator wishes to study which biological processes the TROAP protein is predicted to participate in. B) Associating a gene with genetic disorders. In the context of one of TROAP's most likely biological processes, *chromosome segregation*, it is predicted to be particularly associated with genes causing melanomas and breast cancer. C) Visualizing a predicted functional relationship network for specific genes. Focusing on a gene set consisting of TROAP, two of its most likely relationship partners (UBE2C and TPX2), and two of its most likely partners in chromosome segregation (TOP2A and NCAPH) retrieves a predicted functional relationship network specific to the area of *chromosome segregation*. D) Viewing genomic data contributing to a prediction. Clicking on a predicted functional relationship or specifically focusing on TROAP's relationship with CDC25C displays the genomic data used to generate the prediction. Here, TROAP is predicted to relate to CDC25C, a highly conserved mitotic regulator, due to very high correlation between the genes' expression in a variety of microarray conditions. Taken together, this evidence suggests that TROAP is strongly cell cycle regulated and may play an as-yet-uncharacterized role in mitosis.

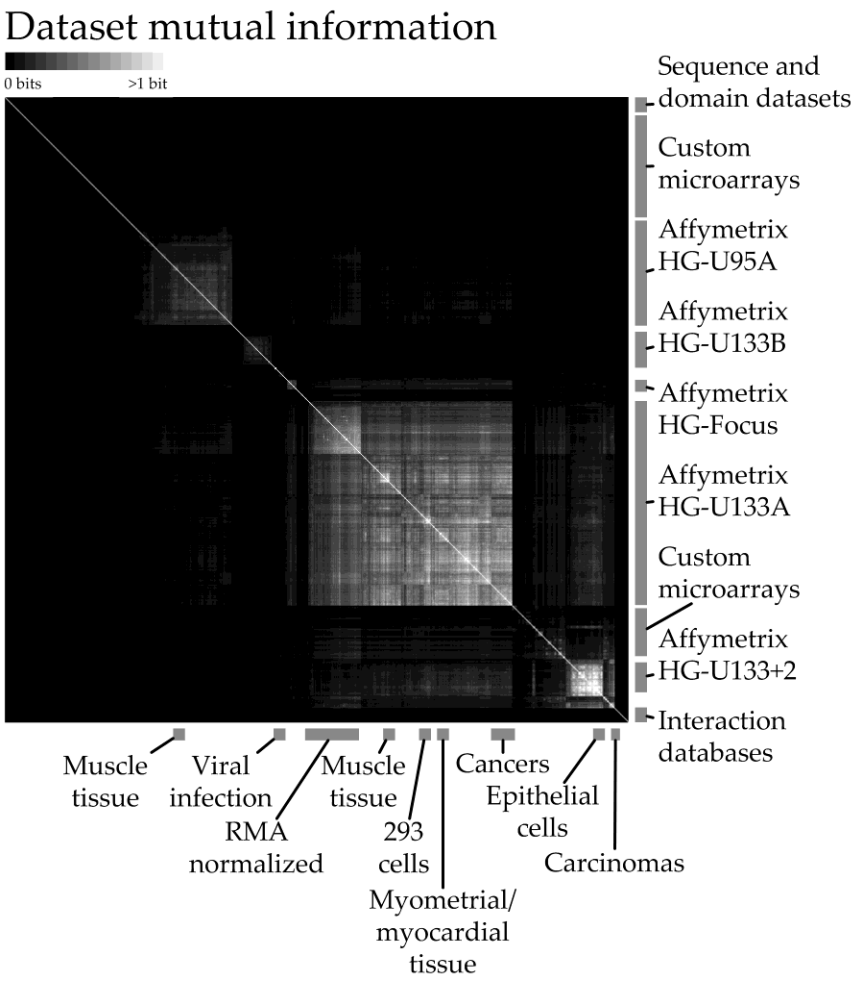


Figure 50: Overview of hierarchically clustered mutual information (MI) between genomic datasets. We used MI among 656 genomic datasets to perform regularization of the parameters of our 230 process-specific Bayesian classifiers. Datasets with a greater proportion of shared information were more heavily mixed with a uniform prior, resulting in the overall upweighting of particularly unique and informative data. Additionally, a global view of the mutual information scores reveals structure in the data. Primarily platform-based effects can be observed among the expression datasets we obtained from GEO (Barrett, Suzek et al. 2005), most of which use Affymetrix arrays; tissue type, cell type, and array normalization algorithms can all cause small amounts of information to be shared between many datasets. For example, Robust MultiArray (RMA) normalization causes a noticeable shift in the information shared among HG-U133A arrays. While the amount of MI between any two datasets is generally low (this figure saturates at one bit of shared information), an accumulation of many small overlaps can result in overconfidence during Bayesian data integration, accounting for the success of parameter regularization.

associations among genes linked to genetic disorders. Our results and functional maps have been made available to the community through the interactive HEFalMp tool at http://function.princeton.edu/hefalmp.

Methods

We integrated 656 genome-scale datasets, comprising ~15,000 microarray conditions and ~15,000 interaction and sequence-based results, to predict process-specific functional relationship networks in 229 biological areas. Data integration was performed using naive Bayesian classifiers, with parameters regularized using a mutual information score between datasets. The resulting functional relationship networks were analyzed to generate functional maps for genes, processes, and diseases within each biological area. Evidence from immunoblotting and fluorescent microscopy was used to confirm novel predictions of the involvement of the LAMP2 and RAB11A proteins in macroautophagy.

Briefly, functional mapping relies on the construction of process-specific functional relationship networks. These are interaction networks in which each node represents a gene, each edge a functional relationship, and an edge between two genes is probabilistically weighted based on experimental evidence relating those genes. We integrate evidence from many datasets, with each dataset weighted in a process-specific manner. To generate functional maps, these networks are mined for functional associations between groups of genes, which might represent individual genes, pathways, processes, or diseases. A functional association summarizes the overall strength of predicted association between the two groups, and it takes four features into account: relationships between genes spanning the two groups, relationships within the groups, each group's background strength of relationship to the entire genome, and the baseline probability of

relationship for all genes. These four features are converted into a p-value by comparing their ratio to a randomized null distribution.

Data preparation

We collected 635 human microarray datasets from the NCBI Gene Expression Omnibus (GEO) repository (Barrett, Suzek et al. 2005) comprising 14,671 conditions. These were processed largely as in (Huttenhower and Troyanskaya 2008), with additional manipulation to handle single channel data and the ambiguity of human probe mapping. Within each dataset, negative and very small (<2) single channel values were removed, genes with missing values in >30% of the conditions were removed, and the remaining missing values were imputed using KNNImpute (Troyanskaya, Cantor et al. 2001) with k=10.

Probe IDs were mapped to HGNC symbols using the appropriate GEO platform files. When multiple probes mapped to a single HGNC symbol, a consensus set of probes was generated by finding pairwise Euclidean distances more likely to have been generated from the dataset's distribution of intra-gene probe pairs than from the distribution of inter-gene probe pairs. If this consensus set contained at least half of the probes mapping to a gene symbol, the consensus set's average value became the expression vector for that gene.

Within each dataset, a similarity score for each pair of genes was generated by first calculating the Pearson correlation ρ between the vectors. These correlations were normalized using Fisher's Z-transform, shifted by the mean, and divided by the dataset standard deviation, yielding a collection of pairwise scores with distribution $N(0, 1)$. Finally, these were binned into one of seven discrete values in the ranges $(-\infty, -1.5]$, $(-1.5, -0.5]$, $(-0.5, 0.5]$, $(0.5, 1.5]$, $(1.5, 2.5]$, $(2.5, 3.5]$, $(3.5, \infty)$.

Non-microarray pairwise datasets were, for the most part, discretized into two bins: interaction and no interaction/no data. In some cases, negative interactions were explicitly recorded by a third bin. Pairwise data was generated from sequence information (transcription factor binding sites, protein domains, etc.) by calculating either the inner product or the Euclidean distance of the occurrence vectors for each gene pair.

Gold standard construction

Biological processes of interest were selected from the Gene Ontology (Ashburner, Ball et al. 2000) by polling a panel of six biologists as described in (Huttenhower and Troyanskaya 2008). Of the 433 GO terms selected to be experimentally informative, 229 had at least ten human gene annotations, becoming our processes of interest.

An answer set of known functionally related and unrelated proteins was derived by combining these gene sets with information from KEGG (Kanehisa, Araki et al. 2008), HPRD (Mishra, Suresh et al. 2006), Pfam (Finn, Mistry et al. 2006), Reactome (Vastrik, D'Eustachio et al. 2007), the Pathway Interaction Database (PID) (Schaefer 2006), and the curated GSEA pathways (Subramanian, Tamayo et al. 2005), all of which represent manually curated databases of functional interactions. A gene pair was considered functionally related if annotated as such in any of these databases and unrelated if annotated to two different terms in GO, KEGG, or PID (the other databases not providing explicit negatives). Genes pairs annotated to terms overlapping with a hypergeometric p-value below 0.05 were excluded from unrelated pair generation (i.e. they were neither related nor unrelated for training and evaluation purposes). This resulted in a gold standard containing 16,184 genes, 8,692,471 functionally related pairs, and 45,712,399 unrelated pairs.

To train and evaluate process-specific classifiers, this answer set was decomposed into subsets related to each biological area of interest. A gene pair was used for training/evaluation in a particular biological process if either A) both genes were annotated to the process in GO or B) one of the two genes was annotated to the process and the pair was unrelated in the standard (i.e. not coannotated to another process).

Evaluation was performed using a holdout set of 6,129 genes (~25% of the genome). Any gene pair including at least one of these genes was withheld from training and used for evaluation of precision/recall, with AUCs calculated analytically using the Wilcoxon rank-sum test.

Data integration

One naive Bayesian classifier was trained per biological area of interest, using the appropriate subset of the gold standard as described above, in addition to one global process-unaware classifier trained using the complete gold standard. Each classifier *f* consisted of a class node predicting the binary presence or absence of a functional relationship (FR) between two genes and n nodes conditioned on FR, each representing the value of a dataset D_k.

Parameter regularization was performed as described in (Steck and Jaakkola 2002) using mutual information between datasets to estimate a strength of prior belief for each dataset. While a large amount of shared information does not guarantee a redundant dataset, since the same subset of information could be shared many times, it provides a valuable quantitative estimate of dataset uniqueness. For each dataset D_k, we calculated a heuristic sum of shared information U_k relative to the dataset's entropy:

$$U_k = 1 + H(D_k)^{-1} \sum_{i \neq k} I(D_i; D_k)$$

We then used this value to weight the strength of prior belief in a uniform distribution for the dataset, based on the technique in (Steck and Jaakkola 2002). This exponentially decreased the weight of a dataset as its shared information increased. Let us notate $|D_k|$ as the number of possible observations in dataset D_k (discretization levels). For some gene pair (g_i, g_j), supporting data $\{d_1(g_i, g_j), d_2(g_i, g_j), ..., d_n(g_i, g_j)\}$, and an effective document count of two, the probability of a FR in function f is thus:

$$P_{i,j}^{f}(\mathrm{FR}) \propto \prod_{k=1}^{n} \frac{2P[D_k = d_k(g_i, g_j)] + 2^{M_k} - 1}{2 + |D_k|(2^{M_k} - 1)}$$

When fewer than 25 gene pairs were available for a particular dataset/relationship combination, the global probability distribution was used for that condition. Remaining zero counts were Laplace smoothed.

An additional global process-aware FR network was generated by transforming each set of process-specific probabilities into z-scores and averaging the results for each gene pair across all processes. Specifically:

$$Z_{i,j}(\mathrm{FR}) = \frac{1}{|F|} \sum_{f \in F} \frac{P_{i,j}^{f}(\mathrm{FR}) - \mathrm{ave}[P^{f}(\mathrm{FR})]}{\mathrm{std}[P^{f}(\mathrm{FR})]}$$

We used the C++ implementations of naive Bayesian learning and inference provided in (Huttenhower, Schroeder et al. 2008), relying on the SMILE library and GeNIe modeling environment (Druzdzel 1999) from the University of Pittsburgh Decision Systems Library for Bayesian network manipulation.

Process-specific analysis

The parameters learned by the naive classifiers in this manner yield a functional activity score (FAS) indicating the strength of the contribution of each dataset within each biological process of interest. A dataset's FAS is the sum of the change each of its possible values makes in the classifier's posterior times the prior probability of observing that value; this yields high scores for data that are both frequent and accurate. The score for dataset D within function f was thus calculated as:

$$FAS_{D,f} = \sum_{i \in D} P(D = i) \left| P(\mathrm{FR}) - P(\mathrm{FR} \mid D = i) \right|$$

Functional modules

Novel functional modules (FMs) are defined within the global process-aware FR network using an algorithm based on (Charikar 2000). We begin with a minimum initial score σ and a minimum final ratio ρ and fill a set of genes G_k and a set of excluded edges E. We repeatedly selected the most related pair of genes not being excluded. To this set, we repeatedly add the gene most related on average until this average relationship probability reaches some fraction ρ of the seed pair's original score. If no such gene can be added, the seed pair is marked as excluded; otherwise, each edge weight in the resulting set is reduced by the average connection weight, and the current G_k is output as a functional module.

Each FM is generated with two parameters: the input ratio ρ and a final average edge weight score $S(G_k)$. ρ is akin to a depth within GO. FMs generated at low ρ are larger, more general, and "higher" in the functional hierarchy; FMs generated at high ρ are smaller, more specific, and "lower" in the hierarchy. The score $S(G_k)$ is an estimated confidence in the FM such that a higher value indicates a more self-contained, certain module. In pseudocode, the algorithm is:

1. Input minimum initial score σ and minimum final ratio ρ
2. Define $S(G_k) = \frac{1}{|G_k|} \sum_{g_i, g_j \in G_k} Z_{i,j}(\text{FR}), S(g_i, G_k) = S(G_k \cup \{g_i\})$
3. Let E={}
4. Let $(g_{s1}, g_{s2}) = \arg\max_{(g_i, g_j) \notin E} Z_{i,j}(\text{FR})$
5. If $Z_{i,j}$(FR)<σ, stop
6. Let G_k={g_{s1}, g_{s2}}
7. Begin loop
8. Let $g_t = \arg\max_{g_i} S(g_i, G_k)$
9. If $S(g_t, G_k)/Z_{s1,s2}$(FR)<ρ, break
10. G_k=G_k∪{g_t}
11. If |G_k|=2
12. E=E∪{(g_{s1}, g_{s2})}
13. Go to step 4
14. Output module G_k with parameters ρ, $S(G_k)$
15. For all g_i,g_j∈G_k
16. $Z_{i,j}$(FR) = max({$Z_{i,j}$(FR) - $S(G_k)$, 0})
17. Go to step 4

To generate novel FMs, we ran this algorithm on the global process-aware human FR network with σ=0.95 and ρ∈{0.01, 0.025, 0.05, 0.075, 0.1, 0.2, ..., 0.5}, generating a set of preliminary FMs $\mathcal{M}$=$\mathcal{M}_{0.01}$∪$\mathcal{M}_{0.025}$∪...∪$\mathcal{M}_{0.5}$. To remove redundant FMs, we merged by union any pair with Jaccard

index at least 0.5, with the newly formed FM occupying the more specific depth. Specifically, for all pairs of modules M_i and M_j within module sets M_x and M_y (ρ depths x and y):

1. Until no changes occur
2. For all $\mathcal{M}_x, \mathcal{M}_y \in \mathcal{M}$
3. For all $M_i \in \mathcal{M}_x, M_j \in \mathcal{M}_y$
4. If $J(M_i, M_j) \geq 0.5$
5. $\mathcal{M}_x = \mathcal{M}_x - \{M_i\}$
6. $\mathcal{M}_y = \mathcal{M}_y - \{M_j\}$
7. $\mathcal{M}_{\max(x,y)} = \mathcal{M}_{\max(x,y)} \cup \{M_i \cup M_j\}$

To form the resulting merged FMs into a DAG similar to the structure of GO, parent/child relationships were established only from higher to lower depths when i) an indirect descendant relationship did not already exist and ii) the higher FM contained at least 2/3 of the lower FM's genes. This generated parent/child relationships $p(M_p, M_c)$:

1. For x from 0.5 to 0.01
2. For y from x to 0.01
3. For all $M_i \in \mathcal{M}_x$ and $M_j \in \mathcal{M}_y$
4. If M_j is not a descendant of M_i and $|M_i \cap M_j|/|M_j| \geq 2/3$
5. $p(M_i, M_j) = 1$

This process resulted in 17,759 FMs across the nine depth levels, 11,674 parent/child relationships, and 10 connected components in the DAG (nine singletons).

Functional mapping associations and p-values

The functional association of two gene sets quantifies the degree of specific overall relationship between their constituent genes. This score is made up of four parts. The score *between* two gene sets within a process is the average probability of all edges between them. Their *background* score in a process is the average probability of all edges incident to either set. The *baseline* score is the average probability of an edge in the process-independent network. The score *within* a single gene set is the average edge probability assuming nodes are self-connected with baseline strength, and the score *within* two gene sets is their unweighted average. The *between* and *baseline* scores are divided by the *background* and *within* scores to calculate two gene sets' functional association, which is thus increased if they are more interconnected and decreased if they are more self-connected.

This score was designed to mitigate several sources of variation and potential false positives in the networks. Known disease genes tend to be well-studied, providing them with more data and increasing their overall probability of functional relationship. Sets of genes representing genetic disorders can thus be very small and highly connected, which is normalized by the within score and its unweighted average. This and the baseline are calculated in the process-independent network, which also has lower variability than the process-specific networks. Normalizing by the *baseline* guarantees an expected value of one, and assuming self-connections with baseline weight allows the functional association score to extend seamlessly to arbitrarily small sets.

Thus, within any functional relationship network f, two gene sets G_1 and G_2 were assigned a functional association score as follows. For $f0$ the global process-independent network and n genes in the genome, let:

$$between^{f}(G_1, G_2) = \frac{1}{|G_1 \parallel G_2|} \sum_{g_i \in G_1, g_j \in G_2} P_{i,j}^{f}(\mathrm{FR})$$

$$background^{f}(G_1, G_2) = \frac{1}{n} \sum_{g_i} \left(\frac{1}{|G_1|} \sum_{g_j \in G_1} P_{i,j}^{f}(\mathrm{FR}) + \frac{1}{|G_2|} \sum_{g_j \in G_2} P_{i,j}^{f}(\mathrm{FR}) \right)$$

$$baseline = \frac{1}{n} \sum_{g_i, g_j} P_{i,j}^{f0}(\mathrm{FR})$$

$$within(G_1) = baseline\,|G_1| + \frac{2}{|G_1|(|G_1| - 1)} \sum_{g_i, g_j \in G_1} P_{i,j}^{f0}(\mathrm{FR})$$

$$within(G_1, G_2) = \frac{1}{2}\left(within(G_1) + within(G_2)\right)$$

All averages are Winsorized by 10% of their length to mitigate outliers; Winsorization is a standard robust averaging process in which the *n* largest and smallest values are replaced by copies of the *n*-1st largest and *n*-1st smallest value, respectively. This defines the functional association between two gene sets as:

$$FA^{f}(G_1, G_2) = \frac{between^{f}(G_1, G_2)}{background^{f}(G_1, G_2)} \cdot \frac{baseline}{within(G_1, G_2)}$$

This score was converted into a p-value by interpolating over a bootstrapped null distribution. For each combination of sizes 1, 2, 5, 10, 15, 20, 25, 50, 100, and 500, pairs of sets were generated randomly 62,500 times within each process and the resulting functional association score calculated. The distributions of these scores were approximately normal, and the standard deviations were asymptotic in the sizes of the two gene sets. Fitting these empirical curves with a ratio of linear polynomials allowed real-time computation of an approximate standard deviation

for any pair of gene set sizes, which then allowed the conversion of functional association scores into p-values using a normal distribution function.

Web-based interface

HEFalMp was implemented in two parts, combining a web-based front end using Ruby on Rails (37signals, Chicago, IL) with a C++ back-end for rapid data processing using the Sleipnir library (Huttenhower, Schroeder et al. 2008). For details, see http://function.princeton.edu/hefalmp.

Experimental validation

Human dermal fibroblasts were cultured in subconfluent conditions in fibroblast basal medium supplemented with FBS, insulin, and fibroblast growth factor (Lonza Group Ltd., Switzerland). Cells received fresh media every two days.

For siRNA transfection, $1.2x10^5$ fibroblasts were transiently transfected with 100nM duplex siRNA designed by the Rosetta algorithm (Sigma, St. Louis, MO) against control targets (ATG5 or luciferase) or experimental targets (RAB11A, LAMP2, VAMP7) using Oligofectamine transfection reagent (Invitrogen, Carlsbad, CA). On the day of experimentation, cells were either supplied with fresh media (not starved), or starved for amino acids for 4 hours in Kreb's Ringer Bicarbonate (KRB) solution (Sigma) at 37C.

Western blots were performed using cell lysates collected on ice by scraping each plate into RIPA buffer (50mM Tris-Cl pH 7.4, 150mM NaCl, 1% Triton X-100, 1% sodium deoxycholate and 0.1% SDS) supplemented with a protease inhibitor cocktail tablet consisting of chymotrypsin (1.5μg/mL), thermolysin (0.8μg/mL), papain (1mg/mL), pronase (1.5μg/mL), pancreatic extract (1.5μg/mL), and trypsin (0.002μg/mL) (Roche Diagnostics, Indianapolis, IN) at either 48 hours (RAB11A) or 72 hours (LAMP2 and VAMP7) post-transfection. Freeze-thawing of lysates was

avoided whenever possible, and freshly denatured samples were run on appropriate percentage SDS-PAGE gels and transferred onto PVDF membranes (Perkin Elmer, Boston, MA) using BioRad electrophoresis equipment (BioRad, Hercules, CA). Antibodies for Western blot analysis were used at the following concentrations in PBS plus BSA: rabbit anti-LC3 at 2μg/mL (Novus Biologicals, Littleton, CO), rabbit anti-RAB11A at 1mg/mL (Sigma), rabbit anti-LAMP2 at 1mg/mL (Sigma), rabbit anti-VAMP7 at 1μg/mL (Abcam Inc., Cambridge, MA).

A GFP-LC3 fusion was used as a fluorescent marker for autopaghy. We generated fibroblasts stably expressing a GFP-LC3 fusion protein by infecting subconfluent fibroblasts with a retroviral construct encoding GFP and the rat LC3 sequence (C. Thompson, University of Pennsylvania). GFP-LC3 fibroblasts transfected with siRNA against control or experimental targets were cultured in uncoated glass bottom culture dishes (MatTek Corp., Ashland, MA) and visualized either 48 hours (RAB11A) or 72 hours (LAMP2 and VAMP7) post-transfection. Transfected GFP-LC3 fibroblasts were imaged using a Zeiss LSM510 confocal microscope.

MOIRAE: Evolutionary Conservation at a Systems Level

The field of comparative genomics has expanded tremendously as the complete genome sequences of many organisms - both closely and distantly related - have become available. Comparative genomics, or the study of gene sequences, functions, and interactions as they vary across multiple organisms (Hardison 2003), provides insights not easily obtainable from other fields or experimental results; these can range from the mechanisms of molecular evolution (Dujon, Sherman et al. 2004; Durand and Hoberman 2006) to fundamental catalogs of gene composition and utilization (Boffelli, McAuliffe et al. 2003; Cliften, Sudarsanam et al. 2003) to

complex maps of regulatory networks as they change over millennia (Kellis, Patterson et al. 2003; Xie, Lu et al. 2005). In a case of scientific coevolution, the field of functional genomics (Ivakhno 2007) has developed in tandem and by utilizing much of the same data. Functional genomics seeks to define the biological roles played by individual genes (Fleischmann, Moller et al. 1999; Ashburner, Ball et al. 2000; Rost, Liu et al. 2003), groups of proteins in pathways or complexes (Kundaje, Lianoglou et al. 2007; Markowetz and Spang 2007), and, at a systems level, the ways in which functional modules within the cell are organized and coordinated (Hood, Heath et al. 2004; Sauer, Heinemann et al. 2007). The intersection of these two research areas has already led to breakthroughs ranging from the clinical (Lee, Chu et al. 2004) to the theoretical (Flannick, Novak et al. 2006).

By leveraging results from both fields, comparative functional genomics has the potential to realize one of the major goals of systems biology: a delineation of the high-level functional modules common to all organisms, their regulatory interplay, and the reasons and mechanisms underlying their organization. Functional genomics has advanced a variety of algorithms for inferring functional networks from genomic data (Lee, Date et al. 2004; Myers, Robson et al. 2005; Franke, van Bakel et al. 2006), and comparative genomics has likewise proposed multiple ways to align biological networks in a principled manner (Stuart, Segal et al. 2003; Sharan, Suthram et al. 2005; Sharan and Ideker 2006). Building on these principles, we have integrated experimental data from seven model organisms spanning over 1.6 billion years of evolution (Hedges, Blair et al. 2004) and four taxonomic kingdoms: *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *A. thaliana*, and *P. falciparum*. This represents information from over 1,700 datasets comprising approximately 25,000 publications and over 50,000 experimental conditions. From

these data, we infer and align functional networks, within which we observe the behavior and evolution of biological pathways and processes among these diverse organisms.

Specifically, we predict functional relationship networks for these seven organisms using probabilistic, context-specific Bayesian data integration (Myers and Troyanskaya 2007; Guan, Myers et al. 2008; Huttenhower, Haley et al. 2009) in 433 biological contexts. For each organism, we provide biological networks in which each edge represents a predicted functional interaction between two proteins, where functional interactions may entail any relationship by which two proteins carry out the same cellular tasks (Troyanskaya, Dolinski et al. 2003). Moreover, context-specificity results in multiple predicted networks per organism, each describing the functional interactions occurring within a specific biological process; for example, the yeast protein Cdc7p interacts specifically with Mer2p in the context of the meiotic cell cycle (Sasanuma, Hirota et al. 2008; Wan, Niu et al. 2008) and with Mcm2p during the mitotic cell cycle (Hardy, Dryga et al. 1997; Lei, Kawasaki et al. 1997), whereas it interacts with Dbf4p in both contexts (Toone, Aerne et al. 1997; Lo, Wan et al. 2008). These networks are then aligned using protein orthology mappings from the Homologene, InParanoid (Remm, Storm et al. 2001), and OrthoMCL (Li, Stoeckert et al. 2003) databases. The resulting aligned biological networks are made available online through the MOIRAE system (Multi-organism Orthologous Integrated Resolution and Alignment of Experiments) and represent a rich resource for biologists wishing to investigate genes of interest in specific model organisms and as conserved across evolution.

Next Steps: Conclusions and Future Work

In summary, modern biology is an area in which computation, machine learning, and data mining have become invaluable, and the successful integration of computer science with biology continues to enrich both fields. Not only does such interdisciplinary research push the bounds of computer science, it is vital for our understanding of molecular biology and of the mechanisms of human disease. As large collections of genomic data are made publicly available by experimentalists worldwide, the principled application of computational techniques to this vast amount of information has the potential to rapidly expand our knowledge of molecular and systems biology.

Here, we have described four major areas in which principled analyses of large genomic data collections have been particularly effective: the guidance of laboratory experiments by computational predictions (and vice versa), the development of algorithmic and software tools to address specific biological questions, tractably capturing biological systems using directed statistical models, and systems-level functional prediction and mapping by very large scale data integration. The successes of each of these areas are interrelated; our ability to discover new mitochondrial proteins in yeast relies on accurate statistical models of experimental data, our ability to manipulate very large data collections requires efficient software development, and so on. As David Botstein (Director of the Lewis-Sigler Institute for Integrative Genomics at Princeton University) has stated, "Any budding researcher needs a foundation in several fields to be able to work on the most important problems confronting scientists today," and this statement is as true of seasoned researchers - and seasoned research - as it is of undergraduate trainees. The best computational biology is also the best computation and the best biology.

Despite the tremendous growth in computational biology in the past decade, however, biology as a whole remains unsolved. The treatment of human disease, the production of optimal medicinal and food crops, and the maintenance of ecological stability are all still open problems, and each has countless facets addressable by future bioinformatic work. We anticipate three specific biological areas in which this research in understanding large scale experimental data will be particularly useful, as well as several broader concerns of the field that must be addressed in the near future. Plant genomics is currently poised to generate tremendous quantities of data and new biology, as driven by the basic science of microRNAs and alternative splicing and by the practical considerations of crop development and biofuels. Similarly, metagenomics is rapidly generating new types of data aimed at understanding microbial communities, pathogenic and symbiotic relationships in crops, and human microflora as they relate to nutrition, aging, and disease. Finally, with increased data of any type comes an increased ability to distinguish signal from noise, leading to the possibility of making more detailed and context-specific predictions than have previously been possible.

Complex and Underrepresented Organisms: Applications to Plant Genomics

At present, some 3,500 microorganisms have been sequenced, in addition to roughly 100 animals; fewer than five plants have been fully sequenced, although dozens more are currently in progress (Benson, Karsch-Mizrachi et al. 2008; Cochrane, Akhtar et al. 2008). This represents a tremendous opportunity for functional genomics, since the availability of plant genome sequences will immediately allow high-throughput data collection (e.g. microarrays), data integration,

comparative genomics, and a host of other analyses in a eukaryotic kingdom of great commercial and scientific importance. Not only would a greater understanding of plant genomics offer immediate benefits to an agricultural industry struggling to feed a global population, but basic science, pharmaceutical development, and, increasingly, bioenergy production all stand to gain immeasurably from an increased focus on plant sequencing and analysis.

The mustard weed *Arabidopsis thaliana* has long been the plant model organism of choice, due to its rapid and relatively simple cultivation in the laboratory, genetic tractability, and small genome size (Meinke, Cherry et al. 1998). It was the first plant to be fully sequenced (Arabidopsis Genome Initiative 2000), and in the subsequent eight years since its genome became available, a vibrant computational and biological community has arisen around it (Duvick, Fu et al. 2008; Swarbreck, Wilks et al. 2008). Nevertheless, a variety of practical issues have made it difficult to computationally analyze *Arabidopsis*, notably a lack of data integration and systematization (Bevan and Walsh 2005), and bioinformatic analysis has lagged as a result; for example, the journal *Bioinformatics* has to date published twice as many articles on the complex eukaryote *Drosophila melanogaster* and almost three times as many on *S. cerevisiae* than it has on *Arabidopsis*.

Perhaps even more surprising is the dearth of crop plant resources; the rice genome has been available for several years (Goff, Ricke et al. 2002; Yu, Hu et al. 2002), and the currently progressing tomato, soy, wheat, potato, and corn genomes are garnering increasing attention, but there a remarkable scarcity of genomic and computational resources are devoted to these incredibly important organisms. Unfortunately, several major causes of this delay arise from exactly the features that make these domesticated plants so important: historical selection by humans has rendered even ancestrally related genomes relatively incomparable, and not only are plants complex, multi-tissue, developmentally staged eukaryotes, their genomes are often more

repetitive and difficult to sequence than even those of mice or humans (Burke, Burger et al. 2007). These issues are in turn somewhat dwarfed by the societal hurdles to be overcome by crop plant research and the stigma of genetically modified organisms (Cockburn 2002).

The practical benefits of increased computational and integrative analysis of genomic plant data are obvious: in a world spending billions of dollars on biofuel development and struggling to feed billions of people, every advance in our understanding of the food we eat and the foundation of our agricultural economy is invaluable. The ways in which a focus on plant informatics will benefit basic science are less sweepingly dramatic but equally critical. Due in part to tens of thousands of years of human cultivation, successful Quantitative Trait Locus (QTL) mapping has flourished in plants (Abiola, Angel et al. 2003; Ashikari, Sakakibara et al. 2005; Salvi and Tuberosa 2005), allowing us to discover the genetic bases of desirable agricultural traits. microRNAs are currently believed to have a somewhat different, more easily detectable architecture in plants, leading to a recent burst of insightful research (Carrington and Ambros 2003; Mallory and Vaucheret 2006; Zhang, Pan et al. 2006). Because of its prevalence in plants, alternative splicing is being actively studied in *Arabidopsis* (Iida, Seki et al. 2004; Wang and Brendel 2006), and due to the impact of parasitic infections on crops, plants are rapidly becoming a platform for studying host-pathogen interactions at a molecular level (Simpson, Reinach et al. 2000; Bhattacharyya, Stilwagen et al. 2002; Lee, Park et al. 2005).

Our specific interests in this area begin with straightforward extensions of our previous and current work in other eukaryotes. As detailed in (Committee on the National Plant Genome Initiative 2008), there is still a need to establish the basic functions, interactions, and regulatory programs for most plant (i.e. *Arabidopsis*) genes; the abundance and richness of genomic data for *Arabidopsis* makes this a prime target for heterogeneous genomic data integration. Moreover,

with the advances discussed here regarding pathway and process understanding from large data collections, this represents an immediate opportunity to discover the molecular bases of novel plant-specific biology. Our own early results and recent work by others (Gunner 2008) already indicate that photosynthetic processes are particularly amenable to bioinformatic characterization, and with the near-availability of the tomato genome, the development of fruiting bodies will be an ideal subject for context- and tissue-specific analysis (Alba, Payton et al. 2005; Fei, Tang et al. 2006). We are particularly interested in studying the regulatory networks underlying these processes, based on our expertise in transcriptional and posttranslational regulatory modeling. Likewise, the incipient release of multiple extremely diverse fruiting plant genomes (orange, melon, berry, grape, etc.) raises the possibility of rapidly gaining an in-depth understanding of gene and protein function in these organisms by means of comparative functional genomics.

Metagenomics and Microflora

As mentioned above, the complete genome sequences of thousands of single celled organisms are currently available. The impetus for sequencing so many microorganisms has been, in part, the emerging fields of metagenomics (Raes, Foerstner et al. 2007; Warnecke and Hugenholtz 2007) and, specifically, the study of human microflora (Backhed, Ley et al. 2005; Ley, Lozupone et al. 2008; Moya, Pereto et al. 2008; Ruby 2008). The relationships between microflora and their hosts (human, plant, or otherwise) are still poorly understood, and advancing the basic science in this area will eventually be of both agricultural and clinical significance. Again, this is an area where other aspects of large scale data manipulation will play an important part, since data integration

across hundreds of species is certainly more challenging than integration within a single organism.

Current work in these areas has proceeded along three main avenues. The first, championed by the J. Craig Venter Institute, has focused on canonical metagenomics through environmental sampling, e.g. with the Global Ocean Sampling expedition (Rusch, Halpern et al. 2007; Yooseph, Sutton et al. 2007). Progress in this area has focused mainly on high-level phylogenetics: characterization of families of organisms, population distributions, protein domains and families, statistics of base pair and codon usage, and so forth. From a purely experimental point of view, this has spurred tremendous advances in high-throughput sequencing technology (Venter, Remington et al. 2004), and this has in turn made available a large body of data with strikingly novel characteristics and structure (Seshadri, Kravitz et al. 2007). However, this data has yet to be thoroughly investigated from a functional perspective, and it is likely to provide invaluable insights into microbial community interaction, photosynthetic and metabolic diversity, and, as a greater variety of environmental samples becomes available, the relationships between ecological niches and molecular systems (Whitham, Difazio et al. 2008).

The second area in which metagenomics is playing an increasing role is the microflora communities of commercial and model organisms, particularly crop plants as mentioned above. Not only do parasitic interactions (both micro- and macroscopic) destroy billions of dollars of crops annually (Granett, Walker et al. 2001; Lee, Park et al. 2005), but even healthy plants live with a remarkable diversity of soil microbes that we are only beginning to understand at a molecular level (Wardle, Bardgett et al. 2004; Johnson, Ijdo et al. 2005; van der Heijden, Bardgett et al. 2008). Fungal plant symbiotes have been recognized for over a century (Bonfante 2003; Finlay 2008), and we have barely begun to collect and analyze genomic data on mycorrhizal

systems (Martin, Aerts et al. 2008; Pain and Hertz-Fowler 2008). Bacterial interactors are equally important (Stougaard 2001; Puhler, Arlat et al. 2004) (and have, albeit in less beneficial forms, made the genetic manipulation of plants possible in the first place (Joos, Timmerman et al. 1983)), and again, we have only recently begun to develop molecular and genomic perspectives on these communities. When extended to consider fungal and bacterial organisms that actively degrade biomass (Antoni, Zverlov et al. 2007; van Zyl, Lynd et al. 2007), these studies also have immediate applications in bioethanol production.

Finally, human microflora are themselves a great priority due to their as-yet-unknown roles in disease, obesity, aging, individual variation, and basic human health (Nicholson, Holmes et al. 2005; Dethlefsen, McFall-Ngai et al. 2007). Interestingly, there are many similarities between the metagenomic studies of human and environmental microflora, since it is becoming increasingly apparent that there may be remarkable variation in composition and function of these microorganisms between hosts and between microenvironments within a host (e.g. mouth, skin, stomach, and gut in humans) (Dethlefsen, McFall-Ngai et al. 2007). This is an area in which functional genomics is particularly critical, since identifying the cellular roles of proteins in symbiotic organisms reflects directly on the impact they have on host phenotypes. This has, for example, been the source of several recent advances in the understanding of human parasitization by the malaria agent *P. falciparum* (Sakata and Winzeler 2007; Birkholtz, van Brummelen et al. 2008), and similar discoveries in the realm of symbiotic microflora could drive fundamental personalized medicine and nutrition.

In addition to identifying new symbiote proteins and protein functions, an immediate application of large-scale genomic data integration to this area lies in identifying cross-species regulatory and functional interactions. There have to date been relatively few genome-scale efforts to identify

host-pathogen regulatory networks (Musser and DeLeo 2005; Tailleux, Waddell et al. 2008), and it is vital to our understanding of the human and crop plant microbiomes that we extend these studies to symbiotic organisms as well. In human beings, as metabolic profiling becomes more widespread (Jernberg, Sullivan et al. 2005; Dumas, Barton et al. 2006; Assfalg, Bertini et al. 2008), this will also provide a key new type of data to be functionally integrated, representing a high-throughput experimental link between the unicellular and organismal worlds. Finally, complementing the many ways in which these advances can contribute to basic science, it is necessary to continue both computational and biological outreach programs in order to increase public understanding of the potential agricultural and clinical benefits of modern biotechnology.

Drilling Down: Predicting and Analyzing Specific Biological Pathways

As described above, both our work and that of others has provided a variety of ways to construct networks of protein functional relationships (Troyanskaya, Dolinski et al. 2003; Karaoz, Murali et al. 2004; Lee, Date et al. 2004; Myers, Robson et al. 2005; Huttenhower and Troyanskaya 2008). A "functional relationship" represents a very general prediction regarding proteins that may actually be directly binding, regulating, or colocalizing with each other in the cell. Biology, of course, deals not just with whether proteins are related, but how: whether they interact physically, if so at what residue, whether they are coregulated, if so by what factor, and what purposes these relationships serve in the greater context of cellular function. By simultaneously considering thousands of experimental results, it becomes possible to make specific predictions of protein-protein relationship types and to assemble them into discrete, biologically detailed pathways (Markowetz and Spang 2007; Markowetz and Troyanskaya 2007). This problem of pathway prediction consists of assigning directionality and biological specificity to previously

predicted (or known) functional relationships; it is, of course, much more computationally challenging, but when successful can be equally more biologically revealing.

Previous efforts at pathway prediction have often relied upon structural learning of graphical models (Friedman 2004; Schafer and Strimmer 2005; Werhli, Grzegorczyk et al. 2006). This is roughly akin to predicting which genes or proteins exert regulatory influences on other specific proteins, either transcriptionally, post-translationally, or indirectly. These methods tend to be computationally expensive, require large amounts of low-noise data, and offer a level of specificity not far beyond that of general functional relationships - but they are extremely statistically robust and, under the right circumstances, can produce high-quality results (Sachs, Perez et al. 2005). Similarly focused efforts have been made for transcriptional regulation networks (Segal, Shapira et al. 2003; Basso, Margolin et al. 2005; Markowetz, Bloch et al. 2005) and for predicting protein-protein interactions (Ben-Hur and Noble 2005; Shen, Zhang et al. 2007; Burger and van Nimwegen 2008) by a variety of methods, mainly kernel methods and, again, graphical models. Like (Sachs, Perez et al. 2005), these methods tend to focus on a single type of biological interaction derived from a single genomic dataset. Expanding upon these efforts to predict rich, structured models of biomolecular interactions from large, heterogeneous genomic data collections is a clear opportunity for data integration techniques.

An orthogonal way in which biological detail can be added to existing models lies in the concept of context specificity (Huttenhower, Hibbs et al. 2006; Myers and Troyanskaya 2007); biology has always known that the same protein may play different roles in different cells or at different times, and it is critical that computational approaches take this into account. In single-celled organisms such as yeast, context often differs only from pathway to pathway, whereas context in metazoan systems can include tissue and cell type, developmental stage, and long-term temporal

changes (e.g. cell cycle phases or the entry and exit from quiescence). Explicitly modeling these tissue and temporal contexts computationally will vastly expand our ability to make specific, molecular-level predictions and our ability to understand systems-level regulatory processes (and thus misregulation in disease).

To date, most approaches to biological context specificity have used one of two brute-force techniques: development of entire systems focused solely on one context (Li and Zhan 2006; Becker and Palsson 2008), or simple replication of some entire system with minimally differentiated parameters for each context (Huttenhower, Hibbs et al. 2006; Myers and Troyanskaya 2007). Many methods take experimental context specificity into account (Liu, Sivaganesan et al. 2006; Hibbs, Hess et al. 2007), which can act as a partial proxy for certain types of biological contexts (e.g. the aggregate of processes perturbed by some experimental condition), and yet more studies have delineated context by simple separation of cell or tissue type (Yu, Lin et al. 2006; Huang, Lin et al. 2007; Shlomi, Cabili et al. 2008) or by phases of temporal processes (Spellman, Sherlock et al. 1998; Tu, Kudlicki et al. 2005; Pramila, Wu et al. 2006). A few recent works (Rachlin, Cohen et al. 2006; Huttenhower and Troyanskaya 2008) have begun to take advantage of the fact that, while these various contexts are distinct, they are not unrelated; the processes of data integration, machine learning, and data visualization and exploration can all take advantage of context specificity in a broader way to offer a more complete picture of real biology.

In both cases - lack of interaction specificity and lack of context specificity - large scale integrative analyses can offer solutions. For the former, work on hierarchical classification such as (Barutcuoglu, Schapire et al. 2006) has demonstrated methods for reliably making the "most specific" prediction possible given an ontology (i.e. hierarchy) of more or less specific

possibilities. In combination with an ontology of protein interaction types (Hermjakob, Montecchi-Palazzi et al. 2004; Kushida, Takagi et al. 2006) and a sufficient body of genomic data, this may allow the inference of very specific biological interaction networks with standard machine learning techniques. Principled modeling of biological context (i.e. tissue and temporal specificity) is likely to prove more challenging, as it raises many of the same issues as does semantic understanding in language processing: in natural language, a word can appear lexicographically and semantically identical, yet have completely different meanings in different usages. A protein can "mean" the same thing in every context, in some contexts, operate very differently in different contexts, or any combination thereof, and this variation can itself vary among the thousands of proteins in a genome. This raises the interesting possibility of modeling protein and interaction "synonyms" (Fellbaum 1998) and inferring relationship networks among these rather than the underlying multifunctional proteins themselves.

Opportunities Abound

A final, broader concern that grows out of the expansion of biological data collections is the organization and integration of the research community itself. The physics community has dealt with massive data collections for decades and evolved successful solutions revolving around distributed computing, centralized storage, approximations, and just plain discarding data when necessary (Doctorow 2008). In biology, a number of difficult practical questions have arisen as intra- and inter-species data integration has become of increasing interest: how can biological data be shared securely and effectively, how can meta-information describe experimental systems, how can biological entities (genes, transcripts, proteins, etc.) be identified systematically, and how can biological knowledge be encoded so as to be both human- and machine-readable?

These questions represent important computational challenges in large scale database and ontology design that have not yet been fully overcome - particularly in human beings, where competing standards, multiple transcripts, incomplete knowledge, and clinical considerations combine to present huge practical and scientific hurdles. The organization and distribution of experimental results, computational predictions, and biological tools through publicly available databases must be improved in order to aid the progress of biologists and bioinformaticians alike.

Here, we have discussed four areas in which our research in computational biology has specifically impacted the field, and each of these in turn has its own potential to benefit future studies. Our development of a method closely integrating computational predictions of gene function with confirmatory laboratory experiments allowed us to triple the number of *S. cerevisiae* genes known to operate in mitochondrial inheritance (from 106 to 340) in less than one person year. Beyond the implications for yeast mitochondrial biology and their applications in understanding human mitochondrial disease, similar advances could be realized in any organism or biological area. By designing computational systems to be explicitly aware of biological prior knowledge and the experimental regime in which they will be applied, and by choosing that experimental regime to be appropriate to the computational system, both algorithmic and laboratory efforts are made more efficient and accurate.

Likewise, software development for specific biological purposes requires not only a comprehensive understanding of efficient algorithms but also of relevant biology. The tools discussed here, such as the Sleipnir library for computational functional genomics, advance computer science through their use of large scale data management and machine learning, but they also advance biology by providing a way to inspect and interpret experimental results.

Nearest Neighbor Networks clustering and the Graphle interface to biological networks provide novel interfaces to gene expression beyond basic coexpression, and COALESCE integrates an array of genomic data in order to reconstruct underlying regulatory networks. Like a microscope or centrifuge, bioinformatic algorithms represent tools for examining life in new ways and for allowing us to make observations that would otherwise remain inaccessible - and like these instruments, they must be carefully engineered, understood, and employed.

Statistical models of biological phenomena serve a dual purpose. From a descriptive perspective, they summarize potentially noisy, high-dimensional data along a few axes of interest, potentially revealing new avenues of interpretation in the process. By discovering genes responding consistently to growth rate based on a collection of microarray data, we provide a catalog of growth-regulated genes in yeast; by capturing the vital statistics of the yeast phosphoproteome, we open a window on the role of phosphorylation in all organisms. From a predictive perspective, however, these models allow us to generate new, concrete hypotheses that can be tested in the laboratory - they actively encourage their own proof or disproof. As with our study of yeast mitochondria, each model represents an opportunity to more quickly, easily, and accurately confirm our understanding of biology through directed experimentation.

Finally, genomic data integration is both the underlying theme of all of these results as well as a research problem in its own right. Science as a whole is a process of data integration, of realizing that the falling apple and the orbiting moon have something in common; biology just tends to be smaller, messier, and harder to understand. Systems such as MEFIT or HEFalMp represent only the first steps towards mapping the tremendous amount of data that is currently available, let alone the exponentially more complex discoveries to be made in the next years and decades. Computation represents only one tool in organizing these data, and machine learning only one

means of suggesting the next steps forward. For the first time, however, modern biotechnological techniques provide a means to observe the molecular mechanisms that allow us to develop from single cells to human beings, and the interpretation of these observations is the fundamental responsibility of computational biology.

References

Abiola, O., J. M. Angel, et al. (2003). "The nature and identification of quantitative trait loci: a community's view." Nat Rev Genet **4**(11): 911-6.

Adai, A. T., S. V. Date, et al. (2004). "LGL: creating a map of protein function with an algorithm for visualizing very large biological networks." J Mol Biol **340**(1): 179-90.

Advani, R. J., B. Yang, et al. (1999). "VAMP-7 mediates vesicular transport from endosomes to lysosomes." J Cell Biol **146**(4): 765-76.

Aebersold, R. and M. Mann (2003). "Mass spectrometry-based proteomics." Nature **422**(6928): 198-207.

Aizerman, M., E. Braverman, et al. (1964). "Theoretical foundations of the potential function method in pattern recognition learning." Automation and Remote Control **25**: 821-37.

Alba, R., P. Payton, et al. (2005). "Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development." Plant Cell **17**(11): 2954-65.

Alberts, B., A. Johnson, et al. (2007). Molecular Biology of the Cell, FIfth Edition, Garland Science.

Alfarano, C., C. E. Andrade, et al. (2005). "The Biomolecular Interaction Network Database and related tools 2005 update." Nucleic Acids Res **33**(Database issue): D418-24.

Alizadeh, A. A., M. B. Eisen, et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature **403**(6769): 503-11.

Allison, D. B., X. Cui, et al. (2006). "Microarray data analysis: from disarray to consolidation and consensus." Nat Rev Genet **7**(1): 55-65.

Alter, O., P. O. Brown, et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." Proc Natl Acad Sci U S A **97**(18): 10101-6.

Altman, D. G. and J. M. Bland (1994). "Diagnostic tests. 1: Sensitivity and specificity." BMJ **308**(6943): 1552.

Altman, R. B. and S. Raychaudhuri (2001). "Whole-genome expression analysis: challenges beyond clustering." Curr Opin Struct Biol **11**(3): 340-7.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Aluru, S., Ed. (2005). Handbook of Computational Molecular Biology. Boca Raton, FL, Chapman & Hall/CRC.

Amberg, D. C., D. J. Burke, et al. (2005). Methods in yeast genetics: a Cold Spring Harbor laboratory course manual. Cold Spring Harbor, NY, Cold Spring Harbor Press.

Andreoli, C., H. Prokisch, et al. (2004). "MitoP2, an integrated database on mitochondrial proteins in yeast and man." Nucleic Acids Res **32**(Database issue): D459-62.

Angus-Hill, M. L., A. Schlichter, et al. (2001). "A Rsc3/Rsc30 zinc cluster dimer reveals novel roles for the chromatin remodeler RSC in gene expression and cell cycle control." Mol Cell **7**(4): 741-51.

Antoni, D., V. V. Zverlov, et al. (2007). "Biofuels from microbes." Appl Microbiol Biotechnol **77**(1): 23-35.

Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature **408**(6814): 796-815.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Ashikari, M., H. Sakakibara, et al. (2005). "Cytokinin oxidase regulates rice grain production." Science **309**(5735): 741-5.

Assfalg, M., I. Bertini, et al. (2008). "Evidence of different metabolic phenotypes in humans." Proc Natl Acad Sci U S A **105**(5): 1420-4.

Attfield, P. V. (1997). "Stress tolerance: the key to effective strains of industrial baker's yeast." Nat Biotechnol **15**(13): 1351-7.

Ayscough, K. R. and D. G. Drubin (1996). "ACTIN: general principles from studies in yeast." Annu Rev Cell Dev Biol **12**: 129-60.

Babcock, M., D. de Silva, et al. (1997). "Regulation of mitochondrial iron accumulation by Yfh1p, a putative homolog of frataxin." Science **276**(5319): 1709-12.

Backhed, F., R. E. Ley, et al. (2005). "Host-bacterial mutualism in the human intestine." Science **307**(5717): 1915-20.

Bader, G. D., I. Donaldson, et al. (2001). "BIND--The Biomolecular Interaction Network Database." Nucleic Acids Res **29**(1): 242-5.

Bader, G. D. and C. W. Hogue (2002). "Analyzing yeast protein-protein interaction data obtained from different sources." Nat Biotechnol **20**(10): 991-7.

Baggerly, K. A., K. R. Coombes, et al. (2001). "Identifying differentially expressed genes in cDNA microarray experiments." J Comput Biol **8**(6): 639-59.

Baitaluk, M., M. Sedova, et al. (2006). "BiologicalNetworks: visualization and analysis tool for systems biology." Nucleic Acids Res **34**(Web Server issue): W466-71.

Baker, C. A. H., M. S. T. Carpendale, et al. (2002). GeneVis: visualization tools for genetic regulatory network dynamics. Visualization. Boston, MA, IEEE**:** 243-50.

Balakrishnan, R., K. R. Christie, et al. (2005). "Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the Saccharomyces Genome Database (SGD)." Nucleic Acids Res **33**(Database issue): D374-7.

Balguerie, A., P. Sivadon, et al. (1999). "Rvs167p, the budding yeast homolog of amphiphysin, colocalizes with actin patches." J Cell Sci **112 (Pt 15)**: 2529-37.

Ball, C. A., I. A. Awad, et al. (2005). "The Stanford Microarray Database accommodates additional microarray platforms and data formats." Nucleic Acids Res **33**(Database issue): D580-2.

Barabasi, A. L. and R. Albert (1999). "Emergence of scaling in random networks." Science **286**(5439): 509-12.

Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet **5**(2): 101-13.

Barrett, T., T. O. Suzek, et al. (2005). "NCBI GEO: mining millions of expression profiles--database and tools." Nucleic Acids Res **33**(Database issue): D562-6.

Barros, M. H., A. M. Myers, et al. (2006). "COX24 codes for a mitochondrial protein required for processing of the COX1 transcript." J Biol Chem **281**(6): 3743-51.

Baruffini, E., I. Ferrero, et al. (2007). "Mitochondrial DNA defects in Saccharomyces cerevisiae caused by functional interactions between DNA polymerase gamma mutations associated with disease in human." Biochim Biophys Acta **1772**(11-12): 1225-35.

Barutcuoglu, Z., R. E. Schapire, et al. (2006). "Hierarchical multi-label prediction of gene function." Bioinformatics **22**(7): 830-6.

Basso, K., A. A. Margolin, et al. (2005). "Reverse engineering of regulatory networks in human B cells." Nat Genet **37**(4): 382-90.

Bauerschmitt, H., S. Funes, et al. (2008). "The membrane-bound GTPase Guf1 promotes mitochondrial protein synthesis under suboptimal conditions." J Biol Chem **283**(25): 17139-46.

Baum, L. E., T. Petrie, et al. (1970). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains." The Annals of Mathematical Statistics **41**(1): 164-71.

Beadle, G. W. and E. L. Tatum (1941). "Genetic Control of Biochemical Reactions in Neurospora." Proc Natl Acad Sci U S A **27**(11): 499-506.

Beausoleil, S. A., M. Jedrychowski, et al. (2004). "Large-scale characterization of HeLa cell nuclear phosphoproteins." Proc Natl Acad Sci U S A **101**(33): 12130-5.

Becker, S. A. and B. O. Palsson (2008). "Context-specific metabolic networks are consistent with experiments." PLoS Comput Biol **4**(5): e1000082.

Beltrao, P. and L. Serrano (2007). "Specificity and evolvability in eukaryotic protein interaction networks." PLoS Comput Biol **3**(2): e25.

Ben-Dor, A., R. Shamir, et al. (1999). "Clustering gene expression patterns." J Comput Biol **6**(3-4): 281-97.

Ben-Hur, A. and W. S. Noble (2005). "Kernel methods for predicting protein-protein interactions." Bioinformatics **21 Suppl 1**: i38-46.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." J. R. Stat. Soc. Ser. B **57**(1): 289-300.

Benson, D. A., I. Karsch-Mizrachi, et al. (2008). "GenBank." Nucleic Acids Res **36**(Database issue): D25-30.

Bergmann, S., J. Ihmels, et al. (2003). "Iterative signature algorithm for the analysis of large-scale gene expression data." Phys Rev E Stat Nonlin Soft Matter Phys **67**(3 Pt 1): 031902.

Bevan, M. and S. Walsh (2005). "The Arabidopsis genome: a foundation for plant research." Genome Res **15**(12): 1632-42.
Bhattacharyya, A., S. Stilwagen, et al. (2002). "Draft sequencing and comparative genomics of Xylella fastidiosa strains reveal novel biological insights." Genome Res **12**(10): 1556-63.
Bhattacharyya, R. P., A. Remenyi, et al. (2006). "The Ste5 scaffold allosterically modulates signaling output of the yeast mating pathway." Science **311**(5762): 822-6.
Birkholtz, L., A. C. van Brummelen, et al. (2008). "Exploring functional genomics for drug target and therapeutics discovery in Plasmodia." Acta Trop **105**(2): 113-23.
Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-74.
Blencowe, B. J. (2006). "Alternative splicing: new insights from global analyses." Cell **126**(1): 37-47.
Blinkov, S. M. and I. y. I. Glezer (1968). The Human Brain in Figures and Tables. A Quantitative Handbook. New York, NY, Plenum Press.
Boer, V. M., J. H. de Winde, et al. (2003). "The genome-wide transcriptional responses of Saccharomyces cerevisiae grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur." J Biol Chem **278**(5): 3265-74.
Boffelli, D., J. McAuliffe, et al. (2003). "Phylogenetic shadowing of primate sequences to find functional regions of the human genome." Science **299**(5611): 1391-4.
Boldogh, I., N. Vojtov, et al. (1998). "Interaction between mitochondria and the actin cytoskeleton in budding yeast requires two integral mitochondrial outer membrane proteins, Mmm1p and Mdm10p." J Cell Biol **141**(6): 1371-81.
Boldogh, I. R. and L. A. Pon (2007). "Mitochondria on the move." Trends Cell Biol **17**(10): 502-10.
Bonfante, P. (2003). "Plants, mycorrhizal fungi and endobacteria: a dialog among cells and genomes." Biol Bull **204**(2): 215-20.
Bork, P., L. J. Jensen, et al. (2004). "Protein interaction networks from yeast to human." Curr Opin Struct Biol **14**(3): 292-9.
Boser, B. E., I. M. Guyon, et al. (1992). A training algorithm for optimal margin classifiers. Fifth annual workshop on Computational learning theory, Pittsburgh, PA, ACM.
Botstein, D., S. A. Chervitz, et al. (1997). "Yeast as a model organism." Science **277**(5330): 1259-60.
Boveri, T. (1902). "Über mehrpolige Mitosen als Mittel zur Analzyse des Zellkerns." Neu Folge **35**: 67-90.
Boyle, E. I., S. Weng, et al. (2004). "GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes." Bioinformatics **20**(18): 3710-5.
Bozdech, Z., J. Zhu, et al. (2003). "Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray." Genome Biol **4**(2): R9.
Brauer, M. J., C. Huttenhower, et al. (2008). "Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast." Mol Biol Cell **19**(1): 352-67.
Brauer, M. J., A. J. Saldanha, et al. (2005). "Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures." Mol Biol Cell **16**(5): 2503-17.
Breitkreutz, B. J., C. Stark, et al. (2003). "The GRID: the General Repository for Interaction Datasets." Genome Biol **4**(3): R23.
Breitkreutz, B. J., C. Stark, et al. (2003). "Osprey: a network visualization system." Genome Biol **4**(3): R22.
Brem, R. B. and L. Kruglyak (2005). "The landscape of genetic complexity across 5,700 gene expression traits in yeast." Proc Natl Acad Sci U S A **102**(5): 1572-7.
Brem, R. B., G. Yvert, et al. (2002). "Genetic dissection of transcriptional regulation in budding yeast." Science **296**(5568): 752-5.
Bro, C., B. Regenberg, et al. (2003). "Transcriptional, proteomic, and metabolic responses to lithium in galactose-grown yeast cells." J Biol Chem **278**(34): 32141-9.
Brunner, S., V. Everard-Gigot, et al. (2002). "Su e of the yeast F1Fo-ATP synthase forms homodimers." J Biol Chem **277**(50): 48484-9.

Bulik, D. A., M. Olczak, et al. (2003). "Chitin synthesis in Saccharomyces cerevisiae in response to supplementation of growth medium with glucosamine and cell wall stress." Eukaryot Cell **2**(5): 886-900.
Bullen, A. (2008). "Microscopic imaging techniques for drug discovery." Nat Rev Drug Discov **7**(1): 54-67.
Bullinger, L., K. Dohner, et al. (2004). "Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia." N Engl J Med **350**(16): 1605-16.
Burger, L. and E. van Nimwegen (2008). "Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method." Mol Syst Biol **4**: 165.
Burke, J. M., J. C. Burger, et al. (2007). "Crop evolution: from genetics to genomics." Curr Opin Genet Dev **17**(6): 525-32.
Butte, A. J. and I. S. Kohane (2006). "Creation and implications of a phenome-genome network." Nat Biotechnol **24**(1): 55-62.
Butte, A. J., P. Tamayo, et al. (2000). "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks." Proc Natl Acad Sci U S A **97**(22): 12182-6.
Calvo, S., M. Jain, et al. (2006). "Systematic identification of human mitochondrial disease genes through integrative genomics." Nat Genet **38**(5): 576-82.
Carpenter, A. E., T. R. Jones, et al. (2006). "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." Genome Biol **7**(10): R100.
Carpenter, A. E. and D. M. Sabatini (2004). "Systematic genome-wide screens of gene function." Nat Rev Genet **5**(1): 11-22.
Carrington, J. C. and V. Ambros (2003). "Role of microRNAs in plant and animal development." Science **301**(5631): 336-8.
Carter, C. D., L. E. Kitchen, et al. (2005). "Loss of SOD1 and LYS7 sensitizes Saccharomyces cerevisiae to hydroxyurea and DNA damage agents and downregulates MEC1 pathway effectors." Mol Cell Biol **25**(23): 10273-85.
Castrillo, J. I., L. A. Zeef, et al. (2007). "Growth control of the eukaryote cell: a systems biology study in yeast." J Biol **6**(2): 4.
Charikar, M. (2000). Greedy approximation algorithms for finding dense components in a graph. Third International Workshop on Approximation Algorithms for Combinatorial Optimization. Saarbrücken, Germany, Springer-Verlag.
Chatr-aryamontri, A., A. Ceol, et al. (2007). "MINT: the Molecular INTeraction database." Nucleic Acids Res **35**(Database issue): D572-4.
Chen, J. W., W. Pan, et al. (1985). "Lysosome-associated membrane proteins: characterization of LAMP-1 of macrophage P388 and mouse embryo 3T3 cultured cells." Arch Biochem Biophys **239**(2): 574-86.
Chen, X. J. and G. D. Clark-Walker (2000). "The petite mutation in yeasts: 50 years on." Int Rev Cytol **194**: 197-238.
Chen, Z., E. A. Odstrcil, et al. (2007). "Restriction of DNA replication to the reductive phase of the metabolic cycle protects genome integrity." Science **316**(5833): 1916-9.
Cheng, Y. and G. M. Church (2000). "Biclustering of expression data." Proc Int Conf Intell Syst Mol Biol **8**: 93-103.
Cherry, J. M., C. Adler, et al. (1998). "SGD: Saccharomyces Genome Database." Nucleic Acids Res **26**(1): 73-9.
Chitikila, C., K. L. Huisinga, et al. (2002). "Interplay of TBP inhibitors in global transcriptional control." Mol Cell **10**(4): 871-82.
Choi, J. K., U. Yu, et al. (2003). "Combining multiple microarray studies and modeling interstudy variation." Bioinformatics **19 Suppl 1**: i84-90.
Chu, Z., J. Li, et al. (2007). "Modulation of cell cycle-specific gene expressions at the onset of S phase arrest contributes to the robust DNA replication checkpoint response in fission yeast." Mol Biol Cell **18**(5): 1756-67.
Chung, H. J., C. H. Park, et al. (2005). "ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics." Nucleic Acids Res **33**(Web Server issue): W621-6.

Clare, A. and R. D. King (2003). "Predicting gene function in Saccharomyces cerevisiae." Bioinformatics **19 Suppl 2**: II42-II49.

Cliften, P., P. Sudarsanam, et al. (2003). "Finding functional features in Saccharomyces genomes by phylogenetic footprinting." Science **301**(5629): 71-6.

Cline, M. S., M. Smoot, et al. (2007). "Integration of biological networks and gene expression data using Cytoscape." Nat Protoc **2**(10): 2366-82.

Cochrane, G., R. Akhtar, et al. (2008). "Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database." Nucleic Acids Res **36**(Database issue): D5-12.

Cockburn, A. (2002). "Assuring the safety of genetically modified (GM) foods: the importance of an holistic, integrative approach." J Biotechnol **98**(1): 79-106.

Collins, S. R., K. M. Miller, et al. (2007). "Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map." Nature **446**(7137): 806-10.

Committee on the National Plant Genome Initiative (2008). Achievements of the National Plant Genome Initiative and New Horizons in Plant Biology. Washington, DC, National Academies Press.

Contamine, V. and M. Picard (2000). "Maintenance and integrity of the mitochondrial genome: a plethora of nuclear genes in the budding yeast." Microbiol Mol Biol Rev **64**(2): 281-315.

Cooper, G. F. (1990). "Probabilistic inference using belief networks is NP-hard." Artificial Intelligence **42**: 393-405.

Cormen, T. H., C. E. Leiserson, et al. (2001). Introduction to Algorithms. Boston, MA, MIT Press.

Cortes, C. and V. Vapnik (1995). "Support-Vector Networks " Machine Learning **20**(3): 273-97.

Cuervo, A. M. and J. F. Dice (1996). "A receptor for the selective uptake and degradation of proteins by lysosomes." Science **273**(5274): 501-3.

Cui, X. and G. A. Churchill (2003). "Statistical tests for differential expression in cDNA microarray experiments." Genome Biol **4**(4): 210.

Date, S. V. and C. J. Stoeckert, Jr. (2006). "Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale." Genome Res **16**(4): 542-9.

David, F. N. (1949). "The Moments of the Z and F Distributions." Biometrika **36**: 394-403.

Davidson, A. C. (2003). Statistical Models. Cambridge, UK, Cambridge University Press.

Davierwala, A. P., J. Haynes, et al. (2005). "The synthetic genetic interaction spectrum of essential genes." Nat Genet **37**(10): 1147-52.

de Haro, C., R. Mendez, et al. (1996). "The eIF-2alpha kinases and the control of protein synthesis." Faseb J **10**(12): 1378-87.

de Hoon, M. J., S. Imoto, et al. (2004). "Open source clustering software." Bioinformatics **20**(9): 1453-4.

De Virgilio, C. and R. Loewith (2006). "The TOR signalling network from yeast to man." Int J Biochem Cell Biol **38**(9): 1476-81.

Demeter, J., C. Beauheim, et al. (2007). "The Stanford Microarray Database: implementation of new analysis tools and open source release of software." Nucleic Acids Res **35**(Database issue): D766-70.

Dethlefsen, L., M. McFall-Ngai, et al. (2007). "An ecological and evolutionary perspective on human-microbe mutualism and disease." Nature **449**(7164): 811-8.

Detweiler, C. S., D. B. Cunanan, et al. (2001). "Host microarray analysis reveals a role for the Salmonella response regulator phoP in human macrophage cell death." Proc Natl Acad Sci U S A **98**(10): 5850-5.

Di Gesu, V., R. Giancarlo, et al. (2005). "GenClust: a genetic algorithm for clustering gene expression data." BMC Bioinformatics **6**: 289.

Diestel, R. (2005). Graph Theory. Heidelberg, Germany, Springer-Verlag.

DiMauro, S. and E. A. Schon (1998). "Nuclear power and mitochondrial disease." Nat Genet **19**(3): 214-5.

Dimmer, K. S., S. Fritz, et al. (2002). "Genetic basis of mitochondrial function and morphology in Saccharomyces cerevisiae." Mol Biol Cell **13**(3): 847-53.

Doctorow, C. (2008). "Big data: Welcome to the petacentre." Nature **455**(7209): 16-21.

Dougherty, E. R., J. Barrera, et al. (2002). "Inference from clustering with application to gene-expression microarrays." J Comput Biol **9**(1): 105-26.

Doyle, T. and D. Botstein (1996). "Movement of yeast cortical actin cytoskeleton visualized in vivo." Proc Natl Acad Sci U S A **93**(9): 3886-91.

Druzdzel, M. J. (1999). SMILE: Structural Modeling, Inference, and Learning Engine and {GeNIe}: a development environment for graphical decision-theoretic models. Sixteenth national conference on Artificial Intelligence. Orlando, Florida, American Association for Artificial Intelligence.

Duggan, D. J., M. Bittner, et al. (1999). "Expression profiling using cDNA microarrays." Nat Genet **21**(1 Suppl): 10-4.

Dujon, B., D. Sherman, et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.

Dumas, M. E., R. H. Barton, et al. (2006). "Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice." Proc Natl Acad Sci U S A **103**(33): 12511-6.

Dumeaux, V., A. Fournier, et al. (2005). "Previous oral contraceptive use and breast cancer risk according to hormone replacement therapy use among postmenopausal women." Cancer Causes Control **16**(5): 537-44.

Dunn, C. D., M. S. Lee, et al. (2006). "A genomewide screen for petite-negative yeast strains yields a new subunit of the i-AAA protease complex." Mol Biol Cell **17**(1): 213-26.

Durand, D. and R. Hoberman (2006). "Diagnosing duplications--can it be done?" Trends Genet **22**(3): 156-64.

Durbin, R., S. R. Eddy, et al. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, UK, Cambridge University Press.

Duvick, J., A. Fu, et al. (2008). "PlantGDB: a resource for comparative plant genomics." Nucleic Acids Res **36**(Database issue): D959-65.

Dwight, S. S., R. Balakrishnan, et al. (2004). "Saccharomyces genome database: underlying principles and organisation." Brief Bioinform **5**(1): 9-22.

Edwards, B. S., T. Oprea, et al. (2004). "Flow cytometry for high-throughput, high-content screening." Curr Opin Chem Biol **8**(4): 392-8.

Efron, B. (1993). An Introduction to the Bootstrap. New York, Chapman and Hall.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.

El-Samad, H., S. Prajna, et al. (2006). "Advanced Methods and Algorithms for Biological Networks Analysis." Proceedings of the IEEE **94**(4): 832-53.

Elemento, O., N. Slonim, et al. (2007). "A universal framework for regulatory element discovery across all genomes and data types." Mol Cell **28**(2): 337-50.

Epstein, C. J. (2006). "Down's syndrome: critical genes in a critical region." Nature **441**(7093): 582-3.

Erdos, P. and A. Renyi (1960). "On the evolution of random graphs." Publications of the Mathematical Institute of the Hungarian Academy of Sciences **5**: 17-61.

Ewing, B., L. Hillier, et al. (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-85.

Eyre, T. A., F. Ducluzeau, et al. (2006). "The HUGO Gene Nomenclature Database, 2006 updates." Nucleic Acids Res **34**(Database issue): D319-21.

Fan, W., K. G. Waymire, et al. (2008). "A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations." Science **319**(5865): 958-62.

Fearon, K. and T. L. Mason (1992). "Structure and function of MRP20 and MRP49, the nuclear genes for two proteins of the 54 S subunit of the yeast mitochondrial ribosome." J Biol Chem **267**(8): 5162-70.

Fedor-Chaiken, M., R. J. Deschenes, et al. (1990). "SRV2, a gene required for RAS activation of adenylate cyclase in yeast." Cell **61**(2): 329-40.

Fei, Z., X. Tang, et al. (2006). "Tomato Expression Database (TED): a suite of data presentation and analysis tools." Nucleic Acids Res **34**(Database issue): D766-70.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Boston, MA, MIT Press.

Ferreira, R. M., L. R. de Andrade, et al. (2006). "Purification and characterization of the chaperone-like Hsp26 from Saccharomyces cerevisiae." Protein Expr Purif **47**(2): 384-92.

Fields, S. and O. Song (1989). "A novel genetic system to detect protein-protein interactions." Nature **340**(6230): 245-6.

Finlay, R. D. (2008). "Ecological aspects of mycorrhizal symbiosis: with special emphasis on the functional diversity of interactions involving the extraradical mycelium." J Exp Bot **59**(5): 1115-26.
Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res **34**(Database issue): D247-51.
Flannick, J., A. Novak, et al. (2006). "Graemlin: general and robust alignment of multiple large interaction networks." Genome Res **16**(9): 1169-81.
Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.
Fleischmann, W., S. Moller, et al. (1999). "A novel method for automatic functional annotation of proteins." Bioinformatics **15**(3): 228-33.
Foury, F. (1997). "Human genetic diseases: a cross-talk between man and yeast." Gene **195**(1): 1-10.
Foury, F. and M. Kucej (2002). "Yeast mitochondrial biogenesis: a model system for humans?" Curr Opin Chem Biol **6**(1): 106-11.
Franke, L., H. van Bakel, et al. (2006). "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes." Am J Hum Genet **78**(6): 1011-25.
Freeman, T. C., L. Goldovsky, et al. (2007). "Construction, visualisation, and clustering of transcription networks from microarray expression data." PLoS Comput Biol **3**(10): 2032-42.
Friedman, N. (2004). "Inferring cellular networks using probabilistic graphical models." Science **303**(5659): 799-805.
Fu, L. D. and I. Tsamardinos (2005). "A comparison of Bayesian network learning algorithms from continuous data." AMIA Annu Symp Proc: 960.
Game, J. C., G. W. Birrell, et al. (2003). "Use of a genome-wide approach to identify new genes that control resistance of Saccharomyces cerevisiae to ionizing radiation." Radiat Res **160**(1): 14-24.
Gansner, E. R. and S. C. North (2000). "An open graph visualization system and its applications to software engineering." Software - Practice and Experience **30**(11): 1203-1233.
Gao, L. Z. and H. Innan (2004). "Very low gene duplication rate in the yeast genome." Science **306**(5700): 1367-70.
Garcia-Rodriguez, L. J., A. C. Gay, et al. (2007). "Puf3p, a Pumilio family RNA binding protein, localizes to mitochondria and regulates mitochondrial biogenesis and motility in budding yeast." J Cell Biol **176**(2): 197-207.
Garcia, B. A., J. Shabanowitz, et al. (2005). "Analysis of protein phosphorylation by mass spectrometry." Methods **35**(3): 256-64.
Gasch, A. P. and M. B. Eisen (2002). "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering." Genome Biol **3**(11): RESEARCH0059.
Gasch, A. P., P. T. Spellman, et al. (2000). "Genomic expression programs in the response of yeast cells to environmental changes." Molecular Biology of the Cell **11**(12): 4241-57.
Gates, W. H. and C. H. Papadimitriou (1979). "Bounds for Sorting by Prefix Reversal." Discrete Mathematics **27**: 47-57.
Gavin, A. C., P. Aloy, et al. (2006). "Proteome survey reveals modularity of the yeast cell machinery." Nature **440**(7084): 631-6.
Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-7.
Gelling, C., I. W. Dawes, et al. (2008). "Mitochondrial Iba57p is required for Fe/S cluster formation on aconitase and activation of radical SAM enzymes." Mol Cell Biol **28**(5): 1851-61.
Gerber, A. P., D. Herschlag, et al. (2004). "Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast." PLoS Biol **2**(3): E79.
Ghaemmaghami, S., W. K. Huh, et al. (2003). "Global analysis of protein expression in yeast." Nature **425**(6959): 737-41.
Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature **418**(6896): 387-91.

Glerum, D. M., T. J. Koerner, et al. (1995). "Cloning and characterization of COX14, whose product is required for assembly of yeast cytochrome oxidase." J Biol Chem **270**(26): 15585-90.
Goff, S. A., D. Ricke, et al. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. japonica)." Science **296**(5565): 92-100.
Goffeau, A., B. G. Barrell, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-7.
Goldstein, A. L. and J. H. McCusker (1999). "Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae." Yeast **15**(14): 1541-53.
Golub, T. R., D. K. Slonim, et al. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." Science **286**(5439): 531-7.
Goode, B. L., D. G. Drubin, et al. (1998). "Regulation of the cortical actin cytoskeleton in budding yeast by twinfilin, a ubiquitous actin monomer-sequestering protein." J Cell Biol **142**(3): 723-33.
Graf, L., M. Iwata, et al. (2002). "Gene expression profiling of the functionally distinct human bone marrow stromal cell lines HS-5 and HS-27a." Blood **100**(4): 1509-11.
Granett, J., M. A. Walker, et al. (2001). "Biology and management of grape phylloxera." Annu Rev Entomol **46**: 387-412.
Granot, D. and M. Snyder (1993). "Carbon source induces growth of stationary phase yeast cells, independent of carbon source metabolism." Yeast **9**(5): 465-79.
Greiner, R. and W. Zhou (2005). "Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers." Machine Learning Journal **59**(3): 297-322.
Gresham, D., D. M. Ruderfer, et al. (2006). "Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray." Science **311**(5769): 1932-6.
Griffith, O. L., E. D. Pleasance, et al. (2005). "Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses." Genomics **86**(4): 476-88.
Grigull, J., S. Mnaimneh, et al. (2004). "Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors." Mol Cell Biol **24**(12): 5534-47.
Gross, J. and J. Yellen (1999). Graph theory and its applications. Boca Raton, FL, CRC Press.
Grunstein, M. (1990). "Histone function in transcription." Annu Rev Cell Biol **6**: 643-78.
Guan, Y., C. L. Myers, et al. (2008). "A genomewide functional network for the laboratory mouse." PLoS Comput Biol **4**(9): e1000165.
Gunner, M. R. (2008). "Computational analysis of photosynthetic systems." Photosynth Res **97**(1): 1-3.
Hall, D. A., J. Ptacek, et al. (2007). "Protein microarray technology." Mech Ageing Dev **128**(1): 161-7.
Hamelers, I. H. and P. H. Steenbergh (2003). "Interactions between estrogen and insulin-like growth factor signaling pathways in human breast tumor cells." Endocr Relat Cancer **10**(2): 331-45.
Hamosh, A., A. F. Scott, et al. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic Acids Res **33**(Database issue): D514-7.
Hand, D. J. and K. Yu (2001). "Idiot's Bayes - not so stupid after all?" International Statistical Review **69**(3): 385-99.
Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.
Hardison, R. C. (2003). "Comparative genomics." PLoS Biol **1**(2): E58.
Hardwick, J. S., F. G. Kuruvilla, et al. (1999). "Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins." Proc Natl Acad Sci U S A **96**(26): 14866-70.
Hardy, C. F., O. Dryga, et al. (1997). "mcm5/cdc46-bob1 bypasses the requirement for the S phase activator Cdc7p." Proc Natl Acad Sci U S A **94**(7): 3151-5.
Hartwell, L. H. (2004). "Yeast and cancer." Biosci Rep **24**(4-5): 523-44.
Hartwell, L. H., J. Culotti, et al. (1974). "Genetic control of the cell division cycle in yeast." Science **183**(120): 46-51.
Hastie, T., R. Tibshirani, et al. (2001). The Elements of Statistical Learning, Springer.

Hasty, J., D. McMillen, et al. (2001). "Computational studies of gene regulatory networks: in numero molecular biology." Nat Rev Genet **2**(4): 268-79.
Haugen, A. C., R. Kelley, et al. (2004). "Integrating phenotypic and expression profiles to map arsenic-response networks." Genome Biol **5**(12): R95.
Hayes, A., N. Zhang, et al. (2002). "Hybridization array technology coupled with chemostat culture: Tools to interrogate gene expression in Saccharomyces cerevisiae." Methods **26**(3): 281-90.
Hedges, S. B. (2002). "The origin and evolution of model organisms." Nat Rev Genet **3**(11): 838-49.
Hedges, S. B., J. E. Blair, et al. (2004). "A molecular timescale of eukaryote evolution and the rise of complex multicellular life." BMC Evol Biol **4**: 2.
Helliwell, S. B., I. Howald, et al. (1998). "TOR2 is part of two related signaling pathways coordinating cell growth in Saccharomyces cerevisiae." Genetics **148**(1): 99-112.
Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). "The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data." Nat Biotechnol **22**(2): 177-83.
Heyer, L. J., S. Kruglyak, et al. (1999). "Exploring expression data: identification and analysis of coexpressed genes." Genome Res **9**(11): 1106-15.
Hibbs, M. A., D. C. Hess, et al. (2007). "Exploring the functional landscape of gene expression: directed search of large microarray compendia." Bioinformatics **23**(20): 2692-9.
Ho, Y., A. Gruhler, et al. (2002). "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." Nature **415**(6868): 180-3.
Hong, E. L., R. Balakrishnan, et al. (2008). "Gene Ontology annotations at SGD: new data sources and annotation methods." Nucleic Acids Res **36**(Database issue): D577-81.
Hood, L., J. R. Heath, et al. (2004). "Systems biology and new technologies enable predictive and preventative medicine." Science **306**(5696): 640-3.
Hu, P., C. M. Greenwood, et al. (2005). "Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models." BMC Bioinformatics **6**: 128.
Huang, T. W., C. Y. Lin, et al. (2007). "Reconstruction of human protein interolog network using evolutionary conserved network." BMC Bioinformatics **8**: 152.
Hugerat, Y., F. Spencer, et al. (1994). "A versatile method for efficient YAC transfer between any two strains." Genomics **22**(1): 108-17.
Hughes, T. R., M. J. Marton, et al. (2000). "Functional discovery via a compendium of expression profiles." Cell **102**(1): 109-26.
Hughes, T. R., M. D. Robinson, et al. (2004). "The promise of functional genomics: completing the encyclopedia of a cell." Curr Opin Microbiol **7**(5): 546-54.
Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.
Hunter, T. and G. D. Plowman (1997). "The protein kinases of budding yeast: six score and more." Trends Biochem Sci **22**(1): 18-22.
Huttenhower, C., A. I. Flamholz, et al. (2007). "Nearest Neighbor Networks: clustering expression data based on gene neighborhoods." BMC Bioinformatics **8**: 250.
Huttenhower, C., E. M. Haley, et al. (2009). "A functional map of the human genome."
Huttenhower, C., M. Hibbs, et al. (2006). "A scalable method for integration and functional analysis of multiple microarray datasets." Bioinformatics **22**(23): 2890-7.
Huttenhower, C., M. Schroeder, et al. (2008). "The Sleipnir library for computational functional genomics." Bioinformatics.
Huttenhower, C. and O. G. Troyanskaya (2006). "Bayesian data integration: a functional perspective." Comput Syst Bioinformatics Conf: 341-51.
Huttenhower, C. and O. G. Troyanskaya (2008). "Assessing the functional structure of genomic data." Bioinformatics **24**(13): i330-8.
Ideker, T., V. Thorsson, et al. (2001). "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." Science **292**(5518): 929-34.

Ideker, T., V. Thorsson, et al. (2000). "Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data." J Comput Biol **7**(6): 805-17.
Ihmels, J., S. Bergmann, et al. (2005). "Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program." PLoS Genet **1**(3): e39.
Ihmels, J., G. Friedlander, et al. (2002). "Revealing modular organization in the yeast transcriptional network." Nat Genet **31**(4): 370-7.
Iida, K., M. Seki, et al. (2004). "Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences." Nucleic Acids Res **32**(17): 5096-103.
Iragne, F., M. Nikolski, et al. (2005). "ProViz: protein interaction visualization and exploration." Bioinformatics **21**(2): 272-4.
Issel-Tarver, L., K. R. Christie, et al. (2002). "Saccharomyces Genome Database." Methods Enzymol **350**: 329-46.
Ito, T., T. Chiba, et al. (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc Natl Acad Sci U S A **98**(8): 4569-74.
Ivakhno, S. (2007). "From functional genomics to systems biology." Febs J **274**(10): 2439-48.
Jaimovich, A., G. Elidan, et al. (2006). "Towards an integrated protein-protein interaction network: a relational Markov network approach." J Comput Biol **13**(2): 145-64.
Janeway, C., P. Travers, et al. (2001). Immunobiology. New York, NY, Garland Science.
Jansen, R., D. Greenbaum, et al. (2002). "Relating whole-genome expression data with protein-protein interactions." Genome Res **12**(1): 37-46.
Jansen, R., H. Yu, et al. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science **302**(5644): 449-53.
Jasnos, L. and R. Korona (2007). "Epistatic buffering of fitness loss in yeast double deletion strains." Nat Genet **39**(4): 550-4.
Jelinsky, S. A., P. Estep, et al. (2000). "Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes." Mol Cell Biol **20**(21): 8157-67.
Jeong, H., S. P. Mason, et al. (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-2.
Jernberg, C., A. Sullivan, et al. (2005). "Monitoring of antibiotic-induced alterations in the human intestinal microflora and detection of probiotic strains by use of terminal restriction fragment length polymorphism." Appl Environ Microbiol **71**(1): 501-6.
Jin, Y. S., J. M. Laplaza, et al. (2004). "Saccharomyces cerevisiae engineered for xylose metabolism exhibits a respiratory response." Appl Environ Microbiol **70**(11): 6816-25.
Joachims, T. (1999). Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. B. Schölkopf, C. Burges and A. Smola, MIT Press.
Johnson, D., M. Ijdo, et al. (2005). "How do plants regulate the function, community structure, and diversity of mycorrhizal fungi?" J Exp Bot **56**(417): 1751-60.
Joos, H., B. Timmerman, et al. (1983). "Genetic analysis of transfer and stabilization of Agrobacterium DNA in plant cells." Embo J **2**(12): 2151-2160.
Jorgensen, P., J. L. Nishikawa, et al. (2002). "Systematic identification of pathways that couple cell growth and division in yeast." Science **297**(5580): 395-400.
Jorgensen, P., I. Rupes, et al. (2004). "A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size." Genes Dev **18**(20): 2491-505.
Joyce, A. R. and B. O. Palsson (2006). "The model organism as a system: integrating 'omics' data sets." Nat Rev Mol Cell Biol **7**(3): 198-210.
Junker, B. H. and F. Schreiber, Eds. (2008). Analysis of Biological Networks. Hoboken, NJ, John Wiley & Sons.
Jurafsky, D. and J. H. Martin (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall.
Kabeya, Y., N. Mizushima, et al. (2000). "LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing." Embo J **19**(21): 5720-8.

Kabeya, Y., N. Mizushima, et al. (2004). "LC3, GABARAP and GATE16 localize to autophagosomal membrane depending on form-II formation." J Cell Sci **117**(Pt 13): 2805-12.
Kanehisa, M., M. Araki, et al. (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(Database issue): D480-4.
Karaoz, U., T. M. Murali, et al. (2004). "Whole-genome annotation by using evidence integration in functional-linkage networks." PNAS **101**(9): 2888-93.
Karp, R. (1972). Reducibility Among Combinatorial Problems. Symposium on the Complexity of Computer Computations, Plenum Press.
Kellis, M., N. Patterson, et al. (2003). "Sequencing and comparison of yeast species to identify genes and regulatory elements." Nature **423**(6937): 241-54.
Kerrien, S., Y. Alam-Faruque, et al. (2007). "IntAct--open source resource for molecular interaction data." Nucleic Acids Res **35**(Database issue): D561-5.
Kim, J. and D. J. Klionsky (2000). "Autophagy, cytoplasm-to-vacuole targeting pathway, and pexophagy in yeast and mammalian cells." Annu Rev Biochem **69**: 303-42.
Kim, K., A. Yamashita, et al. (2004). "Capping protein binding to actin in yeast: biochemical mechanism and physiological relevance." J Cell Biol **164**(4): 567-80.
Kim, R. H., D. Wang, et al. (2000). "A novel smad nuclear interacting protein, SNIP1, suppresses p300-dependent TGF-beta signal transduction." Genes Dev **14**(13): 1605-16.
Kim, S. K., J. Lund, et al. (2001). "A gene expression map for Caenorhabditis elegans." Science **293**(5537): 2087-92.
Kishimoto, M. and S. Goto (1995). "Growth temperatures and electrophoretic karyotyping as tools for practical discrimination of Saccharomyces bayanus and Saccharomyces cerevisiae." Journal of General and Applied Microbiology **41**(3): 239-247.
Kitano, H. (2002). "Computational systems biology." Nature **420**(6912): 206-10.
Klevecz, R. R., J. Bolen, et al. (2004). "A genomewide oscillation in transcription gates DNA replication and cell cycle." Proc Natl Acad Sci U S A **101**(5): 1200-5.
Klionsky, D. J. (2007). "Autophagy: from phenomenology to molecular understanding in less than a decade." Nat Rev Mol Cell Biol **8**(11): 931-7.
Kloster, M., C. Tang, et al. (2005). "Finding regulatory modules through large-scale gene-expression data analysis." Bioinformatics **21**(7): 1172-9.
Knijnenburg, T. A., L. F. Wessels, et al. (2008). "Combinatorial influence of environmental parameters on transcription factor activity." Bioinformatics **24**(13): i172-81.
Kochetov, A. V., A. Sarai, et al. (2005). "The role of alternative translation start sites in the generation of human protein diversity." Mol Genet Genomics **273**(6): 491-6.
Kohn, K. W. (1999). "Molecular interaction map of the mammalian cell cycle control and DNA repair systems." Mol Biol Cell **10**(8): 2703-34.
Koutnikova, H., V. Campuzano, et al. (1997). "Studies of human, mouse and yeast homologues indicate a mitochondrial function for frataxin." Nat Genet **16**(4): 345-51.
Krogan, N. J., G. Cagney, et al. (2006). "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." Nature **440**(7084): 637-43.
Krogan, N. J., M. Kim, et al. (2003). "Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II." Mol Cell Biol **23**(12): 4207-18.
Krogan, N. J., W. T. Peng, et al. (2004). "High-definition macromolecular composition of yeast RNA-processing complexes." Mol Cell **13**(2): 225-39.
Kryshtafovych, A., K. Fidelis, et al. (2007). "Progress from CASP6 to CASP7." Proteins **69 Suppl 8**: 194-207.
Kulesh, D. A., D. R. Clive, et al. (1987). "Identification of interferon-modulated proliferation-related cDNA sequences." Proc Natl Acad Sci U S A **84**(23): 8453-7.
Kundaje, A., S. Lianoglou, et al. (2007). "Learning regulatory programs that accurately predict differential expression with MEDUSA." Ann N Y Acad Sci **1115**: 178-202.
Kushida, T., T. Takagi, et al. (2006). "Event ontology: a pathway-centric ontology for biological processes." Pac Symp Biocomput: 152-63.

Lambert, A. J. and M. D. Brand (2007). "Research on mitochondria and aging, 2006-2007." Aging Cell **6**(4): 417-20.
Lanckriet, G. R., M. Deng, et al. (2004). "Kernel-based data fusion and its application to protein function prediction in yeast." Pacific Symposium on Biocomputing: 300-11.
Lashkari, D. A., J. L. DeRisi, et al. (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis." Proc Natl Acad Sci U S A **94**(24): 13057-62.
Lauritzen, S. and D. Spiegelhalter (1988). "Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems." J. Royal Statistical Society **50**(2).
Lauze, E., B. Stoelcker, et al. (1995). "Yeast spindle pole body duplication gene MPS1 encodes an essential dual specificity protein kinase." Embo J **14**(8): 1655-63.
Lee, B. M., Y. J. Park, et al. (2005). "The genome sequence of Xanthomonas oryzae pathovar oryzae KACC10331, the bacterial blight pathogen of rice." Nucleic Acids Res **33**(2): 577-86.
Lee, I., S. V. Date, et al. (2004). "A probabilistic functional network of yeast genes." Science **306**(5701): 1555-8.
Lee, J. S., I. S. Chu, et al. (2004). "Application of comparative functional genomics to identify best-fit mouse models to study human cancer." Nat Genet **36**(12): 1306-11.
Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.
Lefebvre-Legendre, L., J. Vaillier, et al. (2001). "Identification of a nuclear gene (FMC1) required for the assembly/stability of yeast mitochondrial F(1)-ATPase in heat stress conditions." J Biol Chem **276**(9): 6789-96.
Lehmann, E. L. (1975). Nonparametrics: Statistical Methods Based on Ranks. San Francisco, CA, Holden-Day, Inc.
Lei, M., Y. Kawasaki, et al. (1997). "Mcm2 is a target of regulation by Cdc7-Dbf4 during the initiation of DNA synthesis." Genes Dev **11**(24): 3365-74.
Leonhardt, S. A., K. Fearson, et al. (1993). "HSP78 encodes a yeast mitochondrial heat shock protein in the Clp family of ATP-dependent proteases." Mol Cell Biol **13**(10): 6304-13.
Ley, R. E., C. A. Lozupone, et al. (2008). "Worlds within worlds: evolution of the vertebrate gut microbiota." Nat Rev Microbiol **6**(10): 776-88.
Li, H. and M. Zhan (2006). "Systematic intervention of transcription for identifying network response to disease and cellular phenotypes." Bioinformatics **22**(1): 96-102.
Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-89.
Liu, X., S. Sivaganesan, et al. (2006). "Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset." Bioinformatics **22**(14): 1737-44.
Lo, H. C., L. Wan, et al. (2008). "Cdc7-Dbf4 Regulates NDT80 Transcription as well as Reductional Segregation during Budding Yeast Meiosis." Mol Biol Cell.
Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.
Lodish, H., A. Berk, et al. (2007). Molecular Cell Biology, W. H. Freeman.
Maaloe, O. and N. O. Kjeldgaard (1966). Control of macromolecular synthesis. New York, W. A. Benjamin, Inc.
MacBeath, G. and S. L. Schreiber (2000). "Printing proteins as microarrays for high-throughput function determination." Science **289**(5485): 1760-3.
MacKay, D. J. C. (2003). Information Theory, Inference, and Learning. Cambridge, UK, Cambridge University Press.
MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observation. 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press.
Madhani, H. D. and G. R. Fink (1997). "Combinatorial control required for the specificity of yeast MAPK signaling." Science **275**(5304): 1314-7.

Mah, A. S., A. E. Elia, et al. (2005). "Substrate specificity analysis of protein kinase complex Dbf2-Mob1 by peptide library and proteome array screening." BMC Biochem **6**: 22.
Mallory, A. C. and H. Vaucheret (2006). "Functions of microRNAs and related small RNAs in plants." Nat Genet **38 Suppl**: S31-6.
Mann, H. B. and D. R. Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other." Annals of Mathematical Statistics **18**: 50-60.
Marcotte, E. M., M. Pellegrini, et al. (1999). "A combined algorithm for genome-wide prediction of protein function." Nature **402**(6757): 83-6.
Markowetz, F., J. Bloch, et al. (2005). "Non-transcriptional pathway features reconstructed from secondary effects of RNA interference." Bioinformatics **21**(21): 4026-32.
Markowetz, F. and R. Spang (2007). "Inferring cellular networks--a review." BMC Bioinformatics **8 Suppl 6**: S5.
Markowetz, F. and O. G. Troyanskaya (2007). "Computational identification of cellular networks and pathways." Mol Biosyst **3**(7): 478-82.
Martin, D. E., P. Demougin, et al. (2004). "Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data." BMC Bioinformatics **5**: 148.
Martin, F., A. Aerts, et al. (2008). "The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis." Nature **452**(7183): 88-92.
Mehrabian, M., H. Allayee, et al. (2005). "Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits." Nat Genet **37**(11): 1224-33.
Meinke, D. W., J. M. Cherry, et al. (1998). "Arabidopsis thaliana: a model plant for genome analysis." Science **282**(5389): 662, 679-82.
Mick, D. U., K. Wagner, et al. (2007). "Shy1 couples Cox1 translational regulation to cytochrome c oxidase assembly." Embo J **26**(20): 4347-58.
Middendorf, M., E. Ziv, et al. (2005). "Inferring network mechanisms: the Drosophila melanogaster protein interaction network." Proc Natl Acad Sci U S A **102**(9): 3192-7.
Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: simple building blocks of complex networks." Science **298**(5594): 824-7.
Mishra, G. R., M. Suresh, et al. (2006). "Human protein reference database--2006 update." Nucleic Acids Res **34**(Database issue): D411-4.
Mitchell, T. (1997). Machine Learning. Boston, MA, McGraw-Hill.
Mizushima, N., A. Yamamoto, et al. (2004). "In vivo analysis of autophagy in response to nutrient starvation using transgenic mice expressing a fluorescent autophagosome marker." Mol Biol Cell **15**(3): 1101-11.
Model, K., C. Meisinger, et al. (2001). "Multistep assembly of the protein import channel of the mitochondrial outer membrane." Nat Struct Biol **8**(4): 361-70.
Modesti, A., L. Bini, et al. (2001). "Expression of the small tyrosine phosphatase (Stp1) in Saccharomyces cerevisiae: a study on protein tyrosine phosphorylation." Electrophoresis **22**(3): 576-85.
Monastyrska, I. and D. J. Klionsky (2006). "Autophagy in organelle homeostasis: peroxisome turnover." Mol Aspects Med **27**(5-6): 483-94.
Moore, D. S. and G. P. McGabe (2005). Introduction to the Practice of Statistics. New York, W. H. Freeman.
Mootha, V. K., J. Bunkenborg, et al. (2003). "Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria." Cell **115**(5): 629-40.
Morange, M. (1998). A History of Molecular Biology. Boston, MA, Harvard University Press.
Moreau, Y., S. Aerts, et al. (2003). "Comparison and meta-analysis of microarray data: from the bench to the computer desk." Trends Genet **19**(10): 570-7.
Moseley, J. B. and B. L. Goode (2006). "The yeast actin cytoskeleton: from cellular function to biochemical mechanism." Microbiol Mol Biol Rev **70**(3): 605-45.
Moya, A., J. Pereto, et al. (2008). "Learning how to live together: genomic insights into prokaryote-animal symbioses." Nat Rev Genet **9**(3): 218-29.

Mozdy, A. D., J. M. McCaffery, et al. (2000). "Dnm1p GTPase-mediated mitochondrial fission is a multi-step process requiring the novel integral membrane component Fis1p." J Cell Biol **151**(2): 367-80.
Mulder, N. J., R. Apweiler, et al. (2005). "InterPro, progress and status in 2005." Nucleic Acids Res **33**(Database issue): D201-5.
Mulholland, J., D. Preuss, et al. (1994). "Ultrastructure of the yeast actin cytoskeleton and its association with the plasma membrane." J Cell Biol **125**(2): 381-91.
Mulligan, M. K., I. Ponomarev, et al. (2006). "Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis." Proc Natl Acad Sci U S A **103**(16): 6368-73.
Murali, T. M., C. J. Wu, et al. (2006). "The art of gene function prediction." Nat Biotechnol **24**(12): 1474-5; author reply 1475-6.
Murphy, K. P. (2001). "The Bayes Net Toolbox for MATLAB." Computing Science and Statistics **33**.
Musser, J. M. and F. R. DeLeo (2005). "Toward a genome-wide systems biology analysis of host-pathogen interactions in group A Streptococcus." Am J Pathol **167**(6): 1461-72.
Muta, T., D. Kang, et al. (1997). "p32 protein, a splicing factor 2-associated protein, is localized in mitochondrial matrix and is functionally important in maintaining oxidative phosphorylation." J Biol Chem **272**(39): 24363-70.
Myers, A. M., L. K. Pape, et al. (1985). "Mitochondrial protein synthesis is required for maintenance of intact mitochondrial genomes in Saccharomyces cerevisiae." Embo J **4**(8): 2087-92.
Myers, C. L., D. R. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**: 187.
Myers, C. L., D. Robson, et al. (2005). "Discovery of biological networks from diverse functional genomic data." Genome Biol **6**(13): R114.
Myers, C. L. and O. G. Troyanskaya (2007). "Context-sensitive data integration and prediction of biological networks." Bioinformatics **23**(17): 2322-30.
Nabieva, E., K. Jim, et al. (2005). "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps." Bioinformatics **21 Suppl 1**: i302-10.
Neapolitan, R. E. (2004). Learning Bayesian Networks. Chicago, Illinois, Prentice Hall.
Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-53.
Nevins, J. R. and A. Potti (2007). "Mining gene expression profiles: expression signatures as cancer phenotypes." Nat Rev Genet **8**(8): 601-9.
Ni, L. and M. Snyder (2001). "A genomic study of the bipolar bud site selection pattern in Saccharomyces cerevisiae." Mol Biol Cell **12**(7): 2147-70.
Nicholson, J. K., E. Holmes, et al. (2005). "Gut microorganisms, mammalian metabolism and personalized health care." Nat Rev Microbiol **3**(5): 431-8.
Nomura, M. (1999). "Regulation of ribosome biosynthesis in Escherichia coli and Saccharomyces cerevisiae: diversity and common principles." J Bacteriol **181**(22): 6857-64.
Nouet, C., M. Bourens, et al. (2007). "Rmd9p controls the processing/stability of mitochondrial mRNAs and its overexpression compensates for a partial deficiency of oxa1p in Saccharomyces cerevisiae." Genetics **175**(3): 1105-15.
Novick, P., B. C. Osmond, et al. (1989). "Suppressors of yeast actin mutations." Genetics **121**(4): 659-74.
Nuhse, T. S., A. Stensballe, et al. (2004). "Phosphoproteomics of the Arabidopsis plasma membrane and a new phosphorylation site database." Plant Cell **16**(9): 2394-405.
O'Rourke, S. M. and I. Herskowitz (2004). "Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis." Mol Biol Cell **15**(2): 532-42.
Obenauer, J. C., L. C. Cantley, et al. (2003). "Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs." Nucleic Acids Res **31**(13): 3635-41.
Ogur, M. and R. St John (1956). "A differential and diagnostic plating method for population studies of respiration deficiency in yeast." J Bacteriol **72**(4): 500-4.
Ogur, M., R. St. John, et al. (1957). "Tetrazolium overlay technique for population studies of respiration deficiency in yeast." Science **125**(3254): 928-9.

Oliphant, A., D. L. Barker, et al. (2002). "BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping." Biotechniques **Suppl**: 56-8, 60-1.
Olivas, W. and R. Parker (2000). "The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast." Embo J **19**(23): 6602-11.
OMIM (2008). Online Mendelian Inheritance in Man, McKusick-Nathans Institute of Genetic Medicine and National Center for Biotechnology Information.
Owen, A. B., J. Stuart, et al. (2003). "A gene recommender algorithm to identify coexpressed genes in C. elegans." Genome Res **13**(8): 1828-37.
Pain, A. and C. Hertz-Fowler (2008). "Genomic adaptation: a fungal perspective." Nat Rev Microbiol **6**(8): 572-3.
Palumbo, M. C., L. Farina, et al. (2008). "Collective behavior in gene regulation: post-transcriptional regulation and the temporal compartmentalization of cellular cycles." Febs J **275**(10): 2364-71.
Pan, X., D. S. Yuan, et al. (2004). "A robust toolkit for functional profiling of the yeast genome." Mol Cell **16**(3): 487-96.
Parkinson, H., M. Kapushesky, et al. (2007). "ArrayExpress--a public database of microarray experiments and gene expression profiles." Nucleic Acids Res **35**(Database issue): D747-50.
Pavlidis, P., J. Weston, et al. (2002). "Learning gene functional classifications from multiple data types." J Computational Biology **9**(2): 401-11.
Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann.
Pekurovsky, D., I. N. Shindyalov, et al. (2004). "A case study of high-throughput biological data processing on parallel platforms." Bioinformatics **20**(12): 1940-7.
Pena-Castillo, L. and T. R. Hughes (2007). "Why are there still over 1000 uncharacterized yeast genes?" Genetics **176**(1): 7-14.
Pena-Castillo, L., M. Tasan, et al. (2008). "A critical assessment of Mus musculus gene function prediction using integrated genomic evidence." Genome Biol **9 Suppl 1**: S2.
Penkett, C. J., J. A. Morris, et al. (2006). "YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms." Nucleic Acids Res **34**(Web Server issue): W330-4.
Perocchi, F., L. J. Jensen, et al. (2006). "Assessing systems properties of yeast mitochondria through an interaction map of the organelle." PLoS Genet **2**(10): e170.
Perocchi, F., E. Mancera, et al. (2008). "Systematic screens for human disease genes, from yeast to human and back." Mol Biosyst **4**(1): 18-29.
Peters-Golden, M. and T. G. Brock (2003). "5-lipoxygenase and FLAP." Prostaglandins Leukot Essent Fatty Acids **69**(2-3): 99-109.
Pfanner, N. and A. Geissler (2001). "Versatility of the mitochondrial protein import machinery." Nat Rev Mol Cell Biol **2**(5): 339-49.
Pierrel, F., M. L. Bestwick, et al. (2007). "Coa1 links the Mss51 post-translational function to Cox1 cofactor insertion in cytochrome c oxidase assembly." Embo J **26**(20): 4335-46.
Pinkel, D., R. Segraves, et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." Nat Genet **20**(2): 207-11.
Pitkanen, J. P., A. Torma, et al. (2004). "Excess mannose limits the growth of phosphomannose isomerase PMI40 deletion strain of Saccharomyces cerevisiae." J Biol Chem **279**(53): 55737-43.
Pramila, T., W. Wu, et al. (2006). "The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle." Genes Dev **20**(16): 2266-78.
Prekeris, R., J. Klumperman, et al. (2000). "A Rab11/Rip11 protein complex regulates apical membrane trafficking via recycling endosomes." Mol Cell **6**(6): 1437-48.
Prieto, C. and J. De Las Rivas (2006). "APID: Agile Protein Interaction DataAnalyzer." Nucleic Acids Res **34**(Web Server issue): W298-302.
Primig, M., R. M. Williams, et al. (2000). "The core meiotic transcriptome in budding yeasts." Nat Genet **26**(4): 415-23.

Pringle, J., J. Broach, et al. (1997). The Molecular and Cellular Biology of the Yeast Saccharomyces: Cell Cycle and Cell Biology, CSHL Press.
Prokisch, H., C. Andreoli, et al. (2006). "MitoP2: the mitochondrial proteome database--now including mouse data." Nucleic Acids Res **34**(Database issue): D705-11.
Prokisch, H., C. Scharfe, et al. (2004). "Integrative analysis of the mitochondrial proteome in yeast." PLoS Biol **2**(6): e160.
Ptashne, M. and A. Gann (2003). "Signal transduction. Imposing specificity on kinases." Science **299**(5609): 1025-7.
Puhler, A., M. Arlat, et al. (2004). "What can bacterial genome research teach us about bacteria-plant interactions?" Curr Opin Plant Biol **7**(2): 137-47.
Qian, J., J. Lin, et al. (2003). "Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data." Bioinformatics **19**(15): 1917-26.
Quackenbush, J. (2001). "Computational analysis of microarray data." Nat Rev Genet **2**(6): 418-27.
Quackenbush, J. (2002). "Microarray data normalization and transformation." Nat Genet **32 Suppl**: 496-501.
Rachlin, J., D. D. Cohen, et al. (2006). "Biological context networks: a mosaic view of the interactome." Mol Syst Biol **2**: 66.
Raes, J., K. U. Foerstner, et al. (2007). "Get the most out of your metagenome: computational analysis of environmental sequence data." Curr Opin Microbiol **10**(5): 490-8.
Ravasz, E., A. L. Somera, et al. (2002). "Hierarchical organization of modularity in metabolic networks." Science **297**(5586): 1551-5.
Ray, L. B., L. D. Chong, et al. (2002). "Computational biology." Sci STKE **2002**(148): EG10.
Rebane, G. and J. Pearl (1987). The Recovery of Causal Poly-trees from Statistical Data. 3rd Workshop on Uncertainty in AI, Seattle, WA.
Rebhan, M., V. Chalifa-Caspi, et al. (1997). "GeneCards: integrating information about genes, proteins and diseases." Trends Genet **13**(4): 163.
Regenberg, B., T. Grotkjaer, et al. (2006). "Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in Saccharomyces cerevisiae." Genome Biol **7**(11): R107.
Reinders, J., R. P. Zahedi, et al. (2006). "Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics." J Proteome Res **5**(7): 1543-54.
Reiss, D. J., N. S. Baliga, et al. (2006). "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks." BMC Bioinformatics **7**: 280.
Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-52.
Ren, B., F. Robert, et al. (2000). "Genome-wide location and function of DNA binding proteins." Science **290**(5500): 2306-9.
Rep, M., J. Nooy, et al. (1996). "Three genes for mitochondrial proteins suppress null-mutations in both Afg3 and Rca1 when over-expressed." Curr Genet **30**(3): 206-11.
Rhodes, D. R., S. A. Tomlins, et al. (2005). "Probabilistic model of the human protein-protein interaction network." Nat Biotechnol **23**(8): 951-9.
Rhodes, D. R., J. Yu, et al. (2004). "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression." Proc Natl Acad Sci U S A **101**(25): 9309-14.
Rieger, K. E. and G. Chu (2004). "Portrait of transcriptional responses to ultraviolet and ionizing radiation in human cells." Nucleic Acids Res **32**(16): 4786-803.
Riezman, H., A. Munn, et al. (1996). "Actin-, myosin- and ubiquitin-dependent endocytosis." Experientia **52**(12): 1033-41.
Rodriguez-Manzaneque, M. T., J. Tamarit, et al. (2002). "Grx5 is a mitochondrial glutaredoxin required for the activity of iron/sulfur enzymes." Mol Biol Cell **13**(4): 1109-21.
Rossi, V., D. K. Banfield, et al. (2004). "Longins and their longin domains: regulated SNAREs and multifunctional SNARE regulators." Trends Biochem Sci **29**(12): 682-8.
Rost, B., J. Liu, et al. (2003). "Automatic prediction of protein function." Cell Mol Life Sci **60**(12): 2637-50.

Rottensteiner, H. and F. L. Theodoulou (2006). "The ins and outs of peroxisomes: co-ordination of membrane transport and peroxisomal metabolism." Biochim Biophys Acta **1763**(12): 1527-40.
Ruby, E. G. (2008). "Symbiotic conversations are revealed under genetic interrogation." Nat Rev Microbiol **6**(10): 752-62.
Rudra, D. and J. R. Warner (2004). "What better measure than ribosome synthesis?" Genes Dev **18**(20): 2431-6.
Rudra, D., Y. Zhao, et al. (2005). "Central role of Ifh1p-Fhl1p interaction in the synthesis of yeast ribosomal proteins." Embo J **24**(3): 533-42.
Ruepp, A., A. Zollner, et al. (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Res **32**(18): 5539-45.
Rusch, D. B., A. L. Halpern, et al. (2007). "The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific." PLoS Biol **5**(3): e77.
Sachs, K., O. Perez, et al. (2005). "Causal protein-signaling networks derived from multiparameter single-cell data." Science **308**(5721): 523-9.
Saeed, A. I., V. Sharov, et al. (2003). "TM4: a free, open-source system for microarray data management and analysis." Biotechniques **34**(2): 374-8.
Sakata, T. and E. A. Winzeler (2007). "Genomics, systems biology and drug development for infectious diseases." Mol Biosyst **3**(12): 841-8.
Saldanha, A. J. (2004). "Java Treeview--extensible visualization of microarray data." Bioinformatics **20**(17): 3246-8.
Saldanha, A. J., M. J. Brauer, et al. (2004). "Nutritional homeostasis in batch and steady-state culture of yeast." Mol Biol Cell **15**(9): 4089-104.
Salvi, S. and R. Tuberosa (2005). "To clone or not to clone plant QTLs: present and future challenges." Trends Plant Sci **10**(6): 297-304.
Salwinski, L., C. S. Miller, et al. (2004). "The Database of Interacting Proteins: 2004 update." Nucleic Acids Res **32**(Database issue): D449-51.
Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." J Mol Biol **94**(3): 441-8.
Santiago, T. C. and C. B. Mamoun (2003). "Genome expression analysis in yeast reveals novel transcriptional regulation by inositol and choline and new regulatory functions for Opi1p, Ino2p, and Ino4p." J Biol Chem **278**(40): 38723-30.
Sapsford, K. E., L. Berti, et al. (2006). "Materials for fluorescence resonance energy transfer analysis: beyond traditional donor-acceptor combinations." Angew Chem Int Ed Engl **45**(28): 4562-89.
Saracco, S. A. and T. D. Fox (2002). "Cox18p is required for export of the mitochondrially encoded Saccharomyces cerevisiae Cox2p C-tail and interacts with Pnt1p and Mss2p in the inner membrane." Mol Biol Cell **13**(4): 1122-31.
Sasanuma, H., K. Hirota, et al. (2008). "Cdc7-dependent phosphorylation of Mer2 facilitates initiation of yeast meiotic recombination." Genes Dev **22**(3): 398-410.
Sauer, U., M. Heinemann, et al. (2007). "Genetics. Getting closer to the whole picture." Science **316**(5824): 550-1.
Schaefer, A. M., R. W. Taylor, et al. (2004). "The epidemiology of mitochondrial disorders--past, present and future." Biochim Biophys Acta **1659**(2-3): 115-20.
Schaefer, C. (2006). An Introduction to the NCI Pathway Interaction Database. NCI-Nature Pathway Interaction Database.
Schaerer, F., G. Morgan, et al. (2001). "Cnm67p is a spacer protein of the Saccharomyces cerevisiae spindle pole body outer plaque." Mol Biol Cell **12**(8): 2519-33.
Schafer, J. and K. Strimmer (2005). "An empirical Bayes approach to inferring large-scale gene association networks." Bioinformatics **21**(6): 754-64.
Schawalder, S. B., M. Kabani, et al. (2004). "Growth-regulated recruitment of the essential yeast ribosomal protein gene activator Ifh1." Nature **432**(7020): 1058-61.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.
Schiffman, H. R. (2001). Sensation and Perception: An Integrated Approach. New York, NY, John Wiley & Sons.
Schlossmann, J., R. Lill, et al. (1996). "Tom71, a novel homologue of the mitochondrial preprotein receptor Tom70." J Biol Chem **271**(30): 17890-5.
Schölkopf, B. and A. J. Smola (2002). Learning with Kernels. Boston, MA, MIT Press.
Schumacher, A., P. Kapranov, et al. (2006). "Microarray-based DNA methylation profiling: technology and applications." Nucleic Acids Res **34**(2): 528-42.
Schwaenen, C., M. Nessling, et al. (2004). "Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations." Proc Natl Acad Sci U S A **101**(4): 1039-44.
Schwikowski, B., P. Uetz, et al. (2000). "A network of protein-protein interactions in yeast." Nat Biotechnol **18**(12): 1257-61.
Schwimmer, C., M. Rak, et al. (2006). "Yeast models of human mitochondrial diseases: from molecular mechanisms to drug screening." Biotechnol J **1**(3): 270-81.
Sears, C. L. (2005). "A dynamic partnership: celebrating our gut flora." Anaerobe **11**(5): 247-51.
Segal, E., Y. Fondufe-Mittendorf, et al. (2006). "A genomic code for nucleosome positioning." Nature **442**(7104): 772-8.
Segal, E., M. Shapira, et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet **34**(2): 166-76.
Segal, E., B. Taskar, et al. (2001). "Rich probabilistic models for gene expression." Bioinformatics **17 Suppl 1**: S243-52.
Sekiya-Kawasaki, M., A. C. Groen, et al. (2003). "Dynamic phosphoregulation of the cortical actin cytoskeleton and endocytic machinery revealed by real-time chemical genetic analysis." J Cell Biol **162**(5): 765-72.
Sesaki, H., S. M. Southard, et al. (2003). "Mgm1p, a dynamin-related GTPase, is essential for fusion of the mitochondrial outer membrane." Mol Biol Cell **14**(6): 2342-56.
Seshadri, R., S. A. Kravitz, et al. (2007). "CAMERA: a community resource for metagenomics." PLoS Biol **5**(3): e75.
SGD. (2006). "Saccharomyces Genome Database." Retrieved May 1, 2006, from ftp://ftp.yeastgenome.org/yeast/.
Sharan, R. and T. Ideker (2006). "Modeling cellular machinery through biological network comparison." Nat Biotechnol **24**(4): 427-33.
Sharan, R., A. Maron-Katz, et al. (2003). "CLICK and EXPANDER: a system for clustering and visualizing gene expression data." Bioinformatics **19**(14): 1787-99.
Sharan, R., S. Suthram, et al. (2005). "Conserved patterns of protein interaction in multiple species." Proc Natl Acad Sci U S A **102**(6): 1974-9.
Sharan, R., I. Ulitsky, et al. (2007). "Network-based prediction of protein function." Mol Syst Biol **3**: 88.
Shen, J., J. Zhang, et al. (2007). "Predicting protein-protein interactions based only on sequences information." Proc Natl Acad Sci U S A **104**(11): 4337-41.
Sherlock, G. (2000). "Analysis of large-scale gene expression data." Curr Opin Immunol **12**(2): 201-5.
Shlomi, T., M. N. Cabili, et al. (2008). "Network-based prediction of human tissue-specific metabolism." Nat Biotechnol **26**(9): 1003-10.
Shutt, T. E. and G. S. Shadel (2007). "Expanding the mitochondrial interactome." Genome Biol **8**(2): 203.
Sickmann, A., J. Reinders, et al. (2003). "The proteome of Saccharomyces cerevisiae mitochondria." Proc Natl Acad Sci U S A **100**(23): 13207-12.
Siddiqi, S. A., J. Mahan, et al. (2006). "Vesicle-associated membrane protein 7 is expressed in intestinal ER." J Cell Sci **119**(Pt 5): 943-50.
Sieben, V. J., C. S. Debes Marun, et al. (2007). "FISH and chips: chromosomal analysis on microfluidic platforms." IET Nanobiotechnol **1**(3): 27-35.

Simpson, A. J., F. C. Reinach, et al. (2000). "The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis." Nature **406**(6792): 151-9.

Singh-Gasson, S., R. D. Green, et al. (1999). "Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array." Nat Biotechnol **17**(10): 974-8.

Sipser, M. (2005). Introduction to the Theory of Computation, Course Technology.

Smith, M. G. and M. Snyder (2006). "Yeast as a model for human disease." Curr Protoc Hum Genet **Chapter 15**: Unit 15 6.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-7.

Sokal, R. R. and C. D. Michener (1958). "A statistical method for evaluating systematic relationships." University of Kansas science bulletin **38**: 1409-1438.

Solinas-Toldo, S., S. Lampel, et al. (1997). "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances." Genes Chromosomes Cancer **20**(4): 399-407.

Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.

Sprague, B. L., R. L. Pego, et al. (2004). "Analysis of binding reactions by fluorescence recovery after photobleaching." Biophys J **86**(6): 3473-95.

St Onge, R. P., R. Mani, et al. (2007). "Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions." Nat Genet **39**(2): 199-206.

Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Res **34**(Database issue): D535-9.

Stears, R. L., T. Martinsky, et al. (2003). "Trends in microarray analysis." Nat Med **9**(1): 140-5.

Steck, H. and T. S. Jaakkola (2002). On the Dirichlet Prior and Bayesian Regularization, MIT.

Steinmetz, L. M., C. Scharfe, et al. (2002). "Systematic screen for human disease genes in yeast." Nat Genet **31**(4): 400-4.

Stougaard, J. (2001). "Genetics and genomics of root symbiosis." Curr Opin Plant Biol **4**(4): 328-35.

Strausberg, R. L., S. Levy, et al. (2008). "Emerging DNA sequencing technologies for human genomic medicine." Drug Discov Today **13**(13-14): 569-77.

Stuart, J. M., E. Segal, et al. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-55.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-50.

Suderman, M. and M. Hallett (2007). "Tools for visually exploring biological networks." Bioinformatics **23**(20): 2651-9.

Suissa, M., K. Suda, et al. (1984). "Isolation of the nuclear yeast genes for citrate synthase and fifteen other mitochondrial proteins by a new screening method." Embo J **3**(8): 1773-81.

Swarbreck, D., C. Wilks, et al. (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." Nucleic Acids Res **36**(Database issue): D1009-14.

Swayne, T. C., A. C. Gay, et al. (2007). "Visualization of mitochondria in budding yeast." Methods Cell Biol **80**: 591-626.

Swindells, M., M. Rae, et al. (2002). "Application of high-throughput computing in bioinformatics." Philos Transact A Math Phys Eng Sci **360**(1795): 1179-89.

Tailleux, L., S. J. Waddell, et al. (2008). "Probing host pathogen cross-talk by transcriptional profiling of both Mycobacterium tuberculosis and infected human dendritic cells and macrophages." PLoS ONE **3**(1): e1403.

Tanay, A., R. Sharan, et al. (2004). "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data." Proc Natl Acad Sci U S A **101**(9): 2981-6.

Tanay, A., R. Sharan, et al. (2002). "Discovering statistically significant biclusters in gene expression data." Bioinformatics **18 Suppl 1**: S136-44.

Tarjan, R. E. (1972). "Depth first search and linear graph algorithms." SIAM Journal on Computing **1**: 146-160.
Tavazoie, S., J. D. Hughes, et al. (1999). "Systematic determination of genetic network architecture." Nat Genet **22**(3): 281-5.
Thulasiraman, K. and M. N. Swamy (1992). Graphs: Theory and Algorithms, Wiley-Inter-science.
Tian, F., Z. Wang, et al. (2005). Learning TAN from Incomplete Data. Advances in Intelligent Computing. Berlin/Heidelberg, Germany, Springer**:** 495-504.
Tian, J., K. Ishibashi, et al. (2004). "The expression of native and cultured RPE grown on different matrices." Physiol Genomics **17**(2): 170-82.
Titz, B., S. Thomas, et al. (2006). "Transcriptional activators in yeast." Nucleic Acids Res **34**(3): 955-67.
Tomaska, L. (2002). "Yeast as a model for mitochondria-related human disorders." FEMS Yeast Res **2**(1): VI-IX.
Tong, A. H. and C. Boone (2006). "Synthetic genetic array analysis in Saccharomyces cerevisiae." Methods Mol Biol **313**: 171-92.
Tong, A. H., M. Evangelista, et al. (2001). "Systematic genetic analysis with ordered arrays of yeast deletion mutants." Science **294**(5550): 2364-8.
Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-13.
Toone, W. M., B. L. Aerne, et al. (1997). "Getting started: regulating the initiation of DNA replication in yeast." Annu Rev Microbiol **51**: 125-49.
Toret, C. P. and D. G. Drubin (2006). "The budding yeast endocytic pathway." J Cell Sci **119**(Pt 22): 4585-7.
Torres, E. M., T. Sokolsky, et al. (2007). "Effects of aneuploidy on cellular physiology and cell division in haploid yeast." Science **317**(5840): 916-24.
Troyanskaya, O., M. Cantor, et al. (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**(6): 520-5.
Troyanskaya, O. G. (2005). "Putting microarrays in a context: integrated analysis of diverse biological data." Brief Bioinform **6**(1): 34-43.
Troyanskaya, O. G. (2007). "Integrated analysis of microarray results." Methods Mol Biol **382**: 429-37.
Troyanskaya, O. G., K. Dolinski, et al. (2003). "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)." PNAS **100**(14): 8348-53.
Tu, B. P., A. Kudlicki, et al. (2005). "Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes." Science **310**(5751): 1152-8.
Ubersax, J. A., E. L. Woodbury, et al. (2003). "Targets of the cyclin-dependent kinase Cdk1." Nature **425**(6960): 859-64.
Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature **403**(6770): 623-7.
Valk, P. J., R. G. Verhaak, et al. (2004). "Prognostically useful gene-expression profiles in acute myeloid leukemia." N Engl J Med **350**(16): 1617-28.
van der Heijden, M. G., R. D. Bardgett, et al. (2008). "The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems." Ecol Lett **11**(3): 296-310.
van Hoof, A., P. Lennertz, et al. (2000). "Three conserved members of the RNase D family have unique and overlapping functions in the processing of 5S, 5.8S, U4, U5, RNase MRP and RNase P RNAs in yeast." Embo J **19**(6): 1357-65.
van Roermund, C. W., H. R. Waterham, et al. (2003). "Fatty acid metabolism in Saccharomyces cerevisiae." Cell Mol Life Sci **60**(9): 1838-51.
van Zyl, W. H., L. R. Lynd, et al. (2007). "Consolidated bioprocessing for bioethanol production using Saccharomyces cerevisiae." Adv Biochem Eng Biotechnol **108**: 205-35.
Vapnik, V. N. (1998). Statistical Learning Theory. New York, NY, Wiley.
Vasilescu, J. and D. Figeys (2006). "Mapping protein-protein interactions by mass spectrometry." Curr Opin Biotechnol **17**(4): 394-9.

Vastrik, I., P. D'Eustachio, et al. (2007). "Reactome: a knowledge base of biologic pathways and processes." Genome Biol **8**(3): R39.
Venter, J. C., K. Remington, et al. (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.
von Mering, C., L. J. Jensen, et al. (2007). "STRING 7--recent developments in the integration and prediction of protein interactions." Nucleic Acids Res **35**(Database issue): D358-62.
Walberg, M. W. (2000). "Applicability of yeast genetics to neurologic disease." Arch Neurol **57**(8): 1129-34.
Walhout, A. J. and M. Vidal (2001). "High-throughput yeast two-hybrid assays for large-scale protein interaction mapping." Methods **24**(3): 297-306.
Wall, M. E., A. Rechtsteiner, et al. (2003). Singular value decomposition and principal component analysis. A Practical Approach to Microarray Data Analysis. D. P. Berrar, W. Dubitzky and M. Granzow. Norwell, MA, Kluwer**:** 91-109.
Wan, L., H. Niu, et al. (2008). "Cdc28-Clb5 (CDK-S) and Cdc7-Dbf4 (DDK) collaborate to initiate meiotic recombination in yeast." Genes Dev **22**(3): 386-97.
Wanders, R. J. and H. R. Waterham (2006). "Peroxisomal disorders: the single peroxisomal enzyme deficiencies." Biochim Biophys Acta **1763**(12): 1707-20.
Wang, B. B. and V. Brendel (2006). "Genomewide comparative analysis of alternative splicing in plants." Proc Natl Acad Sci U S A **103**(18): 7175-80.
Wang, Y., M. Pierce, et al. (2004). "Ras and Gpa2 mediate one branch of a redundant glucose signaling pathway in yeast." PLoS Biol **2**(5): E128.
Ward, D. M., J. Pevsner, et al. (2000). "Syntaxin 7 and VAMP-7 are soluble N-ethylmaleimide-sensitive factor attachment protein receptors required for late endosome-lysosome and homotypic lysosome fusion in alveolar macrophages." Mol Biol Cell **11**(7): 2327-33.
Wardle, D. A., R. D. Bardgett, et al. (2004). "Ecological linkages between aboveground and belowground biota." Science **304**(5677): 1629-33.
Warnecke, F. and P. Hugenholtz (2007). "Building on basic metagenomics with complementary technologies." Genome Biol **8**(12): 231.
Warner, J. R. (1999). "The economics of ribosome biosynthesis in yeast." Trends Biochem Sci **24**(11): 437-40.
Warringer, J. and A. Blomberg (2003). "Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in Saccharomyces cerevisiae." Yeast **20**(1): 53-67.
Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." Nature **171**(4356): 737-8.
Werhli, A. V., M. Grzegorczyk, et al. (2006). "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks." Bioinformatics **22**(20): 2523-31.
Westermann, B. and W. Neupert (2000). "Mitochondria-targeted green fluorescent proteins: convenient tools for the study of organelle biogenesis in Saccharomyces cerevisiae." Yeast **16**(15): 1421-7.
Westermann, B. and W. Neupert (2003). "'Omics' of the mitochondrion." Nat Biotechnol **21**(3): 239-40.
Wheeler, D. B., A. E. Carpenter, et al. (2005). "Cell microarrays and RNA interference chip away at gene function." Nat Genet **37 Suppl**: S25-30.
Whitham, T. G., S. P. Difazio, et al. (2008). "Extending genomics to natural communities and ecosystems." Science **320**(5875): 492-5.
Wilcoxon, F. (1945). "Individual comparisons by ranking methods." Biometrics Bulletin **1**: 80-3.
Williams, R. M., M. Primig, et al. (2002). "The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast." Proc Natl Acad Sci U S A **99**(21): 13431-6.
Xiao, W., B. L. Chow, et al. (1998). "Mms4, a putative transcriptional (co)activator, protects Saccharomyces cerevisiae cells from endogenous and environmental DNA damage." Mol Gen Genet **257**(6): 614-23.
Xie, X., J. Lu, et al. (2005). "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals." Nature **434**(7031): 338-45.
Yooseph, S., G. Sutton, et al. (2007). "The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families." PLoS Biol **5**(3): e16.

Yorimitsu, T. and D. J. Klionsky (2005). "Autophagy: molecular machinery for self-eating." Cell Death Differ **12 Suppl 2**: 1542-52.

Yu, J., S. Hu, et al. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. indica)." Science **296**(5565): 79-92.

Yu, X., J. Lin, et al. (2006). "Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues." Nucleic Acids Res **34**(17): 4925-36.

Yvert, G., R. B. Brem, et al. (2003). "Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors." Nat Genet **35**(1): 57-64.

Zaman, S., S. I. Lippman, et al. (2008). "How Saccharomyces Responds to Nutrients." Annu Rev Genet.

Zeng, X., M. H. Barros, et al. (2008). "ATP25, a new nuclear gene of Saccharomyces cerevisiae required for expression and assembly of the Atp9p subunit of mitochondrial ATPase." Mol Biol Cell **19**(4): 1366-77.

Zhang, B., X. Pan, et al. (2006). "Plant microRNA: a small regulatory molecule with big impact." Dev Biol **289**(1): 3-16.

Zhao, R., M. Davey, et al. (2005). "Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone." Cell **120**(5): 715-27.

Zhao, Y., J. H. Sohn, et al. (2003). "Autoregulation in the biosynthesis of ribosomes." Mol Cell Biol **23**(2): 699-707.

Zhu, J. and M. Q. Zhang (1999). "SCPD: a promoter database of the yeast Saccharomyces cerevisiae." Bioinformatics **15**(7-8): 607-11.

Zweig, M. H. and G. Campbell (1993). "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." Clin Chem **39**(4): 561-77.

Appendix A: Supplemental Information

S. cerevisiae	*S. bayanus*	*S. pombe*	*S. cerevisiae*	*S. bayanus*	*S. pombe*
YEL040W	YEL040W	SPBC21B10.07	YGR032W	YGR032W	SPCC1840.02C, SPAC19B12.03, SPAC24C9.07C, SPBC19G7.05C
YOL014W	YOL014W		YGR236C	YGR236C	
YGL076C		SPAC664.06, SPAC3H5.07, SPBC18H10.12C	YJL161W	YJL161W	
YML056C	YML056C	SPBC2F12.14C	YMR196W	YMR196W	
YOL039W	YOL039W	SPAC1071.08, SPBC23G7.15C, SPBP8B7.06	YDR070C	YDR070C	
YOL120C	YOL120C	SPAPB17E12.13, SPBC11C11.07	YMR174C	YMR174C	
YDL081C		SPCP1E11.09C, SPBC3B9.13C, SPAC644.15	YPL280W		SPCC757.03C, SPBC947.09, SPAC5H10.02C, SPAC1F7.06, SPAC11D3.13
YOR167C	YOR167C	SPCC285.15C, SPAC25G10.06	YDL085W	YDL085W	SPBC947.15C, SPAC3A11.07
YNL301C		SPAPB17E12.13, SPBC11C11.07	YMR322C		SPCC757.03C, SPBC947.09, SPAC5H10.02C, SPAC1F7.06, SPAC11D3.13
YDL130W	YDL130W	SPCP1E11.09C, SPBC3B9.13C, SPAC644.15	YGL156W	YGL156W	SPAC513.05
YLR406C	YLR406C	SPAC890.08	YLL026W		SPBC16D10.08C
YDL014W	YDL014W	SPBC2D10.10C	YNL093W	YNL093W	SPAC6F6.15
YBR085W		SPBC530.10C	YBR169C	YBR169C	SPAC110.04C
YMR242C	YMR242C	SPAC26A3.04, SPAC3A12.10	YCR021C	YCR021C	SPCC31H12.02C
YEL026W	YEL026W	SPAC607.03C	Q0130		SPMIT.10
YBR291C	YBR291C	SPAC19G12.05	YIL136W	YIL136W	
YBL087C	YBL087C	SPCC1322.11, SPAC3G9.03	YGR256W	YGR256W	SPBC660.16
YGL031C		SPAC22E12.13C, SPCC330.14C, SPAC6G9.09C	YOL052C-A		

YHR128W	YHR128W	SPAC1B3.01C, SPAC1399.04C, SPAC1002.17C	YML128C	YML128C	SPAC23C4.05C, SPBC365.12C
YLR112W			YDR171W	YDR171W	SPCC338.06C, SPBC3E7.02C
YLR149C		SPCC4G3.03	YOR173W	YOR173W	SPBP4H10.20
YJL144W	YJL144W		YBL049W	YBL049W	SPAPJ691.02
YNL237W	YNL237W	SPBC3B8.06	YLR312C	YLR312C	
YLL067C			YPL186C	YPL186C	
YGR070W	YGR070W	SPAC1006.06, SPCC645.07	YJL116C	YJL116C	SPBC2G2.17C, SPAC1002.13C
YHR138C	YHR138C		YGL121C	YGL121C	
YJR008W	YJR008W	SPAC4H3.04C	YGR043C	YGR043C	SPCC1020.06C
YGR142W	YGR142W		YLR327C	YLR327C	
YDR379C-A	YDR379C-A	SPAC664.12C	YBL048W		
YDR258C	YDR258C	SPBC4F6.17C	YIL160C	YIL160C	
YOR338W	YOR338W	SPCC1682.13, SPAC14C4.12C	YBR116C		
YLR178C			YBR072W	YBR072W	SPCC338.06C, SPBC3E7.02C
YDL169C	YDL169C		YHR096C	YHR096C	SPCC548.06C, SPBC1348.14C, SPCC1235.13, SPCC1235.14, SPBC1683.08, SPAC1F8.01, SPBC4B4.08, SPCC548.07C
YKR046C	YKR046C				

Supplemental Table 1: *S. cerevisiae* calibration genes used for growth rate prediction in this study with *S. bayanus* and *S. pombe* orthologs. *S. cerevisiae* calibration genes were defined to have a bootstrapped p-value of growth rate response and linear fit less than 10^{-5}. *S. bayanus* orthologs were drawn from (Kellis, Patterson et al. 2003) and *S. pombe* orthologs from (Penkett, Morris et al. 2006).

Mutant	Rate	Mutant	Rate	Mutant	Rate	Mutant	Rate
ade1	0.256	hda1	0.248	sbh2	0.255	yhl013c	0.250
ade16	0.256	hdf1	0.253	sbp1	0.244	yhl029c	0.248
ade2 haploid	0.246	hes1 haploid	0.254	scs7	0.254	yhl042w	0.253
aep2	0.238	hir2	0.255	sgs1	0.248	yhl045w	0.252
afg3 haploid	0.239	his1	0.253	sgt2	0.245	yhr011w	0.241
ald5	0.256	hmg1 haploid	0.259	she4	0.236	yhr022c	0.251
anp1	0.245	hog1 haploid	0.256	sin3	0.248	yhr031c	0.247
aqy2a	0.257	hpa3	0.249	sir1	0.255	yhr034c	0.251
aqy2b	0.256	hpt1	0.254	sir2	0.252	yhr039c	0.251
ard1	0.248	hst3	0.239	sir3	0.254	yil037c haploid	0.246
are1/are2 haploid	0.252	imp2	0.246	sir4	0.268	yil117c haploid	0.252
arg5/6	0.255	imp2	0.253	sod1 haploid	0.254	yjl107c haploid	0.262
arg80	0.254	isw1	0.251	spf1	0.247	yml003w	0.255
ase1	0.254	isw1/isw2	0.251	ssn6 haploid	0.216	yml005w	0.255
ate1	0.257	isw2	0.256	sst2 haploid	0.243	yml011c	0.254
bim1	0.237	jnm1	0.252	stb4	0.255	yml018c	0.255
bni1 haploid	0.247	kim4	0.240	ste11 haploid	0.256	yml033w	0.254
bub1 haploid	0.244	kin3	0.250	ste12 haploid	0.252	yml034w	0.255
bub2	0.257	kre1	0.255	ste18 haploid	0.257	ymr009w	0.255
bub3	0.244	kss1 haploid	0.253	ste2 haploid	0.255	ymr010w	0.254
bub3 haploid	0.244	mac1	0.255	ste20	0.259	ymr014w	0.256
bul1	0.250	mad2	0.254	ste24 haploid	0.255	ymr025w	0.253
cat8	0.251	mak10	0.251	ste4 haploid	0.250	ymr029c	0.252
cbp2	0.252	mbp1	0.256	ste5 haploid	0.254	ymr030w	0.256
cem1	0.244	med2 haploid	0.248	ste7 haploid	0.253	ymr031c	0.255
cin5	0.254	mnn1	0.254	swi4	0.241	ymr031w-a	0.251
cka2	0.237	mrpl33	0.249	swi5	0.251	ymr034c	0.253
ckb2	0.245	mrt4	0.245	swi6 haploid	0.244	ymr040w	0.254
cla4 haploid	0.252	msu1	0.241	tec1 haploid	0.251	ymr041c	0.257
clb2	0.250	npr2	0.257	tom6	0.252	ymr044w	0.253
clb6	0.255	nrf1	0.257	top1 haploid	0.255	ymr140w	0.252
cmk2	0.251	nta1	0.254	top3 haploid	0.257	ymr141c	0.246
cna1/cna2 haploid	0.252	ost3	0.251	tup1 haploid	0.234	ymr145c	0.251
cnb1	0.253	pac2	0.251	ubp8	0.256	ymr147w	0.255
cue1	0.257	pau2	0.253	ubr1	0.256	ymr187c	0.256
cup5	0.247	pch1	0.255	ubr2	0.253	ymr237w	0.251
cyc2	0.255	pcl6	0.253	utr4	0.252	ymr244c-a	0.254
cyt1	0.250	pep12	0.247	vac8	0.243	ymr258c	0.253
dfr1	0.253	pet111	0.250	vma8	0.243	ymr269w	0.253
dig1	0.255	pet117	0.247	vps21	0.256	ymr285c	0.253
dig1/dig2	0.242	pet127	0.255	vps8	0.246	ymr293c	0.239
dig1/dig2 haploid	0.250	pex12	0.258	whi2	0.257	ynd1	0.253
dig2	0.252	pfd2	0.241	yaf1	0.253	yor006c	0.258
dot4	0.261	phd1 haploid	0.257	yal004w	0.255	yor009w	0.254
eca39	0.257	ppr1	0.253	yap1	0.252	yor015w	0.253
ecm1	0.251	prb1	0.252	yap3	0.252	yor021c	0.253
ecm10	0.250	qcr2 haploid	0.235	yap7	0.252	yor051c	0.254

ecm18	0.255	rad27	0.252	yar014c	0.247	yor072w	0.254
ecm29	0.254	rad57	0.244	yar030c	0.256	yor078w	0.255
ecm31	0.257	rad6 haploid	0.244	yea4	0.255	yor080w	0.237
ecm34	0.253	ras1	0.256	yel001c	0.253	ypl216w	0.252
eft2	0.259	ras1 haploid	0.253	yel008w	0.257	zds1	0.253
erd1	0.249	ras2 haploid	0.256	yel010w	0.257	AUR1 (tet)	0.248
erg2	0.238	rgt1	0.254	yel020c	0.256	CDC42 (tet)	0.244
erg3 haploid	0.245	rip1	0.242	yel028w	0.254	ERG11 (tet)	0.222
erg4 haploid	0.244	rml2	0.240	yel033w	0.260	FKS1 (tet)	0.249
erg5	0.254	rnh1	0.254	yel044w	0.242	HMG2 (tet)	0.241
erg6	0.249	rnr1 haploid	0.244	yel047c	0.254	IDI1 (tet)	0.250
erp2	0.253	rpd3 haploid	0.237	yel059w	0.254	KAR2 (tet)	0.239
erp4	0.256	rpl12a	0.252	yel067c	0.254	PMA1 (tet)	0.232
far1 haploid	0.254	rpl20a	0.255	yer002w	0.253	RHO1 (tet)	0.253
fks1 haploid	0.251	rpl27a	0.259	yer024w	0.256	YEF3 (tet)	0.257
fpr1	0.248	rpl34a	0.256	yer030w	0.262	X2 deoxy-D-glu.	0.251
fre6	0.254	rpl6b	0.256	yer033c	0.254	Calcofluor white	0.251
fus2	0.257	rpl8a	0.253	yer034w	0.255	Cycloheximide	0.261
fus3 haploid	0.252	rps24a	0.256	yer041w	0.255	Doxycycline	0.247
fus3/kss1 haploid	0.252	rps24a haploid	0.257	yer044c haploid	0.239	FR901-228	0.245
gal83	0.259	rps27b	0.264	yer050c	0.238	Glucosamine	0.256
gas1	0.240	rrp6	0.247	yer066c-a	0.254	HU	0.246
gcn4	0.254	rtg1	0.241	yer067c-a	0.256	Itraconazole	0.236
gfd1	0.253	rts1	0.247	yer071c	0.253	Lovastatin	0.246
gln3	0.257	rvs161 haploid	0.252	yer083c	0.240	MMS	0.234
gpa2	0.249	sap1	0.252	yer084w	0.255	Nikkomycin	0.252
gyp1	0.258	sap18	0.253	yer085c	0.256	Terbinafine	0.251
hat2	0.254	sap30	0.243	CMD1 (tet)	0.234	Tunicamycin	0.243

Supplemental Table 2: Predicted relative growth rates for microarray data from the (Hughes, Marton et al. 2000) deletion collection. Our predictions for the 199 mutants for which Hughes et al directly measured growth rates show significant correlation to the experimental gold standard (ρ=0.473, $p<10^{-11}$), in contrast to other single mutant fitness estimates based on growth curve analysis (e.g. (Warringer and Blomberg 2003), ρ=0.321, $p<10^{-6}$; (Jasnos and Korona 2007), ρ=0.108, p>0.2).

Scansite Group	**Phosphoproteins**
Tyrosine kinase	4
Src homology 2	3
Src homology 3	14
Basophilic	184
DNA damage	18
Acidophilic	51
Proline-dependent	59

Supplemental Table 3: Counts of phosphoproteins as detected by SCANSITE per kinase target family.

	Group Interactions		Total Interactions	
	Genetic	**Physical**	**Genetic**	**Physical**
Random mean (deviation)	49.5 (18.8)	145.1 (27.8)	989.4 (204.3)	2762.2 (227.3)
Phosphoproteins	124	238	1421	3139
p-value	<10-4	<10-3	0.035	0.098

Supplemental Table 4: Interaction counts within phosphoprotein groups (or identically sized random subsamples) and across all known interactions of the target phosphoproteins within the yeast genome. Genetic interactions were drawn from synthetic lethality data, and physical interactions were taken from yeast two-hybrid and co-immunoprecipitation studies.

First Author	Year	Pubmed ID	Title
Angus	2001	11336698	rsc3/rsc30 knockouts
Belli	2004	14722110	Oxidative stress and glutaredoxin 5-deficient mutant
Bernstein	2000	11095743	Trichostatin A treatment time course
Brauer	2005	15758028	Diauxic shift time course (Batch2)
Brem	2002	11923494	Transcriptional regulation (II)
Brem	2005	15659551	Genetic variation in gene expression among parents and progenies
Caba	2005	15878181	Genotoxic stress
Casagrande	2000	10882108	Unfolded protein response
Causton	2001	11179418	heat response
Causton	2001	11179418	Sorbitol response
Chu	1998	9784122	Sporulation time course
Cohen	2002	12006656	yap1 and yap2 knockouts with peroxide and cadmium added
Duvel	2003	12820961	post heat shock, delayed rapamycin exposure time course
Eriksson	2005	16199888	SPT10 global transcription regulator null mutant
Fleming	1999	11830665	proteasome inhibition with exposure to PS-341
Gasch	2000	11102521	Menadione exposure time course
Gasch	2000	11102521	Hydrogen peroxide response time course
Gasch	2000	11102521	Heat Shock 30C to 37C time course
Gasch	2000	11102521	Heat Shock from various temp to 37C
Gasch	2000	11102521	Carbon sources
Hardwick	2000	10611304	rapamycin exposure
Ideker	2001	11340206	GAL mutants
Iyer	2001	11206552	SBF-MBF genomic distribution (intergenic_v1.0) (I)
Jin	2004	15528549	Xylose metabolism
Keller	2001	11504737	Haa1 analysis
Lee	2005	15989963	rnt1 null mutant expression profile
Martin	2004	15476558	TOR2-controlled transcription
Ogawa	2000	11102525	Phosphate-regulated pathway (I)
Orlandi	2004	14623890	Deubiquitinating enzyme UBP10 inactivation
Primig	2000	11101837	Sporulation of two strains
Roberts	2000	10657304	Pheremone response
Rudra	2005	15692568	fhl1 and ifh1 deletion mutants
Saldanha	2004	15240820	limitation by Uracil
Saldanha	2004	15240820	limitation by Leucine
Schawalder	2004	15616569	IFH1 overexpression: time course
Segal	2003	12740579	Stationary phase, ypl230w mutant
Spellman	1998	9843569	Cell cycle, alpha-factor block-release
Spellman	1998	9843569	Cell cycle, elutriation
Tai	2005	15496405	Nutrient limitation under aerobic and anaerobic conditions
Williams	2002	12370439	Ume6 regulon (Ye6100subB)
Yamamoto	2005	15647283	Heat shock transcription factor 1 mutant response to heat stress
Yoshimoto	2002	12058033	Na(+) exposure
Yvert	2003	12897782	Trans-acting regulatory variation

Supplemental Table 5: Microarray data integrated by the MEFIT system.

Term ID	Description	Term ID	Description
GO:0000067	DNA replication and chromosome cycle	GO:0009266	response to temperature
GO:0000074	regulation of cell cycle	GO:0009268	response to pH
GO:0000160	two-component signal transduction system (phosphorelay)	GO:0009302	snoRNA transcription
GO:0000278	mitotic cell cycle	GO:0009303	rRNA transcription
GO:0000279	M phase	GO:0009305	protein amino acid biotinylation
GO:0000280	nuclear division	GO:0009306	protein secretion
GO:0000338	protein deneddylation	GO:0009308	amine metabolism
GO:0000746	conjugation	GO:0009314	response to radiation
GO:0000902	cellular morphogenesis	GO:0009410	response to xenobiotic stimulus
GO:0001101	response to acid	GO:0009415	response to water
GO:0001510	RNA methylation	GO:0009452	RNA capping
GO:0001522	pseudouridine synthesis	GO:0009581	detection of external stimulus
GO:0005975	carbohydrate metabolism	GO:0009636	response to toxin
GO:0006033	chitin localization	GO:0009743	response to carbohydrate stimulus
GO:0006056	mannoprotein metabolism	GO:0009847	spore germination
GO:0006066	alcohol metabolism	GO:0009966	regulation of signal transduction
GO:0006081	aldehyde metabolism	GO:0010035	response to inorganic substance
GO:0006082	organic acid metabolism	GO:0015791	polyol transport
GO:0006112	energy reserve metabolism	GO:0015833	peptide transport
GO:0006113	fermentation	GO:0015837	amine transport
GO:0006118	electron transport	GO:0015849	organic acid transport
GO:0006260	DNA replication	GO:0015891	siderophore transport
GO:0006265	DNA topological change	GO:0015893	drug transport
GO:0006266	DNA ligation	GO:0015931	nucleobase, nucleoside, nucleotide and nucleic acid transport
GO:0006276	plasmid maintenance	GO:0015976	carbon utilization
GO:0006280	mutagenesis	GO:0016032	viral life cycle
GO:0006308	DNA catabolism	GO:0016050	vesicle organization and biogenesis
GO:0006310	DNA recombination	GO:0016071	mRNA metabolism
GO:0006323	DNA packaging	GO:0016072	rRNA metabolism
GO:0006352	transcription initiation	GO:0016073	snRNA metabolism
GO:0006353	transcription termination	GO:0016074	snoRNA metabolism
GO:0006354	RNA elongation	GO:0016192	vesicle-mediated transport
GO:0006360	transcription from RNA polymerase I promoter	GO:0016458	gene silencing
GO:0006366	transcription from RNA polymerase II promoter	GO:0016481	negative regulation of transcription
GO:0006383	transcription from RNA polymerase III promoter	GO:0016485	protein processing
GO:0006390	transcription from mitochondrial promoter	GO:0016925	protein sumoylation
GO:0006399	tRNA metabolism	GO:0016926	protein desumoylation
GO:0006401	RNA catabolism	GO:0016998	cell wall catabolism
GO:0006417	regulation of protein biosynthesis	GO:0017006	protein-tetrapyrrole linkage
GO:0006457	protein folding	GO:0018065	protein-cofactor linkage
GO:0006461	protein complex assembly	GO:0018193	peptidyl-amino acid modification
GO:0006473	protein amino acid acetylation	GO:0018410	peptide or protein carboxyl-terminal blocking

GO:0006476	protein amino acid deacetylation
GO:0006508	proteolysis and peptidolysis
GO:0006512	ubiquitin cycle
GO:0006518	peptide metabolism
GO:0006519	amino acid and derivative metabolism
GO:0006629	lipid metabolism
GO:0006662	glycerol ether metabolism
GO:0006725	aromatic compound metabolism
GO:0006730	one-carbon compound metabolism
GO:0006766	vitamin metabolism
GO:0006790	sulfur metabolism
GO:0006793	phosphorus metabolism
GO:0006800	oxygen and reactive oxygen species metabolism
GO:0006807	nitrogen compound metabolism
GO:0006811	ion transport
GO:0006818	hydrogen transport
GO:0006839	mitochondrial transport
GO:0006858	extracellular transport
GO:0006869	lipid transport
GO:0006913	nucleocytoplasmic transport
GO:0006914	autophagy
GO:0006944	membrane fusion
GO:0006970	response to osmotic stress
GO:0006974	response to DNA damage stimulus
GO:0006984	ER-nuclear signaling pathway
GO:0006986	response to unfolded protein
GO:0006997	nuclear organization and biogenesis
GO:0007001	chromosome organization and biogenesis (sensu Eukaryota)
GO:0007005	mitochondrion organization and biogenesis
GO:0007009	plasma membrane organization and biogenesis
GO:0007010	cytoskeleton organization and biogenesis
GO:0007029	ER organization and biogenesis
GO:0007030	Golgi organization and biogenesis
GO:0007031	peroxisome organization and biogenesis
GO:0007032	endosome organization and biogenesis
GO:0007033	vacuole organization and biogenesis
GO:0007034	vacuolar transport
GO:0007039	vacuolar protein catabolism
GO:0007046	ribosome biogenesis
GO:0007047	cell wall organization and biogenesis
GO:0007059	chromosome segregation
GO:0018987	osmoregulation
GO:0019236	response to pheromone
GO:0019748	secondary metabolism
GO:0019932	second-messenger-mediated signaling
GO:0019953	sexual reproduction
GO:0019954	asexual reproduction
GO:0030261	chromosome condensation
GO:0030397	membrane disassembly
GO:0030435	sporulation
GO:0030447	filamentous growth
GO:0030705	cytoskeleton-dependent intracellular transport
GO:0031023	microtubule organizing center organization and biogenesis
GO:0031123	RNA 3'-end processing
GO:0040029	regulation of gene expression, epigenetic
GO:0042044	fluid transport
GO:0042157	lipoprotein metabolism
GO:0042176	regulation of protein catabolism
GO:0042180	ketone metabolism
GO:0042214	terpene metabolism
GO:0042594	response to starvation
GO:0043037	translation
GO:0043094	metabolic compound salvage
GO:0043284	biopolymer biosynthesis
GO:0044419	interaction between organisms
GO:0045116	protein neddylation
GO:0045184	establishment of protein localization
GO:0045185	maintenance of protein localization
GO:0045333	cellular respiration
GO:0045454	cell redox homeostasis
GO:0045471	response to ethanol
GO:0045595	regulation of cell differentiation
GO:0045941	positive regulation of transcription
GO:0046483	heterocycle metabolism
GO:0046677	response to antibiotic
GO:0046713	boron transport
GO:0048284	organelle fusion
GO:0048285	organelle fission
GO:0048308	organelle inheritance
GO:0050790	regulation of enzyme activity
GO:0050801	ion homeostasis
GO:0050821	protein stabilization

GO:0007155	cell adhesion	GO:0050874	organismal physiological process
GO:0007166	cell surface receptor linked signal transduction	GO:0051049	regulation of transport
GO:0007243	protein kinase cascade	GO:0051052	regulation of DNA metabolism
GO:0007264	small GTPase mediated signal transduction	GO:0051129	negative regulation of cell organization and biogenesis
GO:0007530	sex determination	GO:0051169	nuclear transport
GO:0007568	aging	GO:0051180	vitamin transport
GO:0007584	response to nutrients	GO:0051181	cofactor transport
GO:0007624	ultradian rhythm	GO:0051186	cofactor metabolism
GO:0008213	protein amino acid alkylation	GO:0051236	establishment of RNA localization
GO:0008219	cell death	GO:0051238	sequestering of metal ion
GO:0008298	intracellular mRNA localization	GO:0051248	negative regulation of protein metabolism
GO:0008380	RNA splicing	GO:0051252	regulation of RNA metabolism
GO:0008643	carbohydrate transport	GO:0051258	protein polymerization
GO:0009100	glycoprotein metabolism	GO:0051261	protein depolymerization
GO:0009116	nucleoside metabolism	GO:0051301	cell division
GO:0009117	nucleotide metabolism	GO:0051321	meiotic cell cycle
GO:0009225	nucleotide-sugar metabolism	GO:0051325	interphase

Supplemental Table 6: Gene Ontology terms deemed to be experimentally informative and used for positive gold standard generation and evaluation in MEFIT.

Term ID	**Description**	**Term ID**	**Description**
GO:0000280	nuclear division	GO:0009100	glycoprotein metabolism
GO:0005975	carbohydrate metabolism	GO:0009116	nucleoside metabolism
GO:0006056	mannoprotein metabolism	GO:0009117	nucleotide metabolism
GO:0006066	alcohol metabolism	GO:0009225	nucleotide-sugar metabolism
GO:0006081	aldehyde metabolism	GO:0009308	amine metabolism
GO:0006082	organic acid metabolism	GO:0015791	polyol transport
GO:0006091	generation of precursor metabolites and energy	GO:0015833	peptide transport
GO:0006259	DNA metabolism	GO:0015837	amine transport
GO:0006276	plasmid maintenance	GO:0015849	organic acid transport
GO:0006350	transcription	GO:0015891	siderophore transport
GO:0006403	RNA localization	GO:0015931	nucleobase, nucleoside, nucleotide and nucleic acid transport
GO:0006457	protein folding	GO:0015976	carbon utilization
GO:0006461	protein complex assembly	GO:0016032	viral life cycle
GO:0006464	protein modification	GO:0016044	membrane organization and biogenesis
GO:0006518	peptide metabolism	GO:0016050	vesicle organization and biogenesis
GO:0006519	amino acid and derivative metabolism	GO:0016070	RNA metabolism
GO:0006629	lipid metabolism	GO:0016192	vesicle-mediated transport
GO:0006662	glycerol ether metabolism	GO:0016265	death
GO:0006725	aromatic compound metabolism	GO:0016458	gene silencing
GO:0006730	one-carbon compound metabolism	GO:0019748	secondary metabolism
GO:0006766	vitamin metabolism	GO:0031023	microtubule organizing center organization and biogenesis

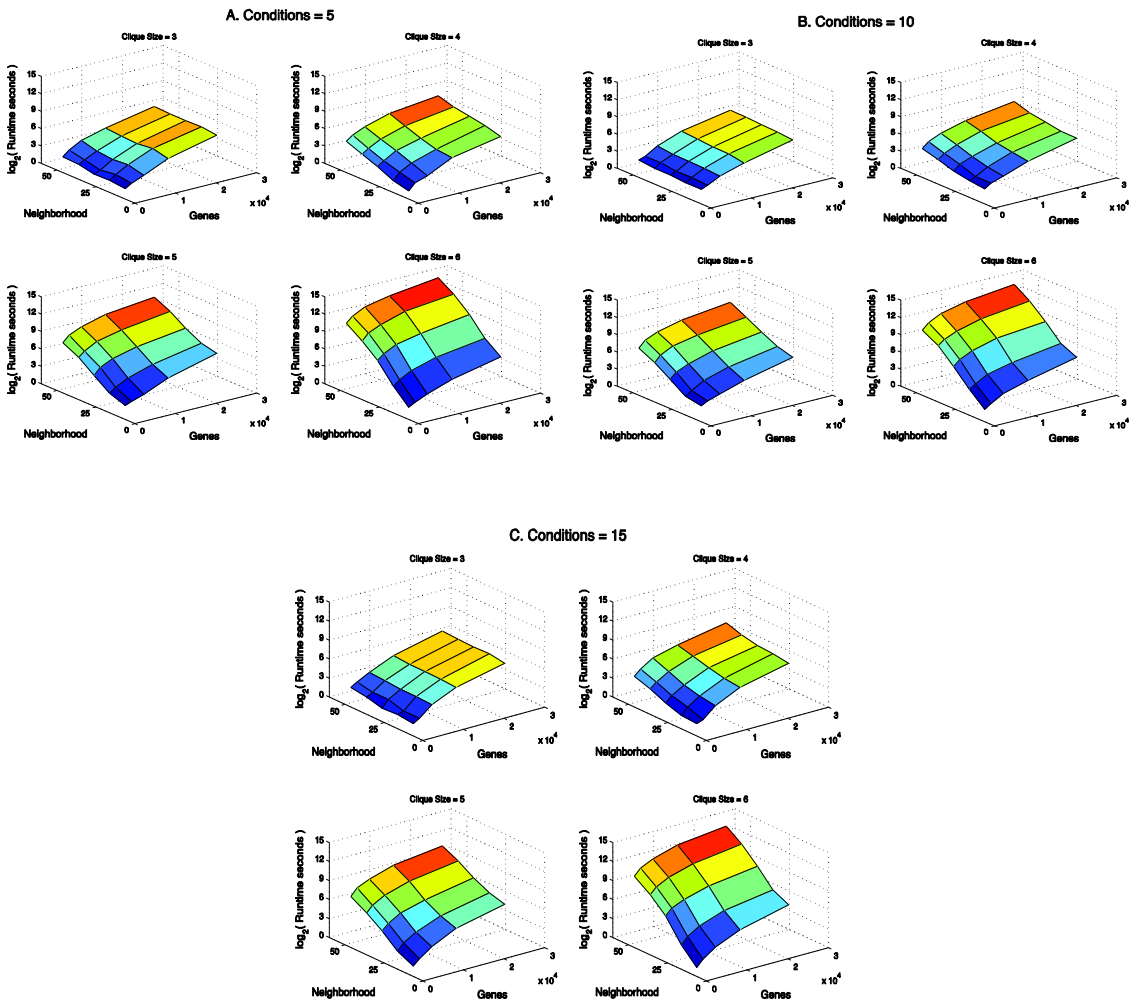
GO:0006790	sulfur metabolism	GO:0042044	fluid transport
GO:0006793	phosphorus metabolism	GO:0042157	lipoprotein metabolism
GO:0006800	oxygen and reactive oxygen species metabolism	GO:0042180	ketone metabolism
GO:0006807	nitrogen compound metabolism	GO:0042592	homeostasis
GO:0006811	ion transport	GO:0043037	translation
GO:0006818	hydrogen transport	GO:0043094	metabolic compound salvage
GO:0006869	lipid transport	GO:0043241	protein complex disassembly
GO:0006914	autophagy	GO:0043284	biopolymer biosynthesis
GO:0006944	membrane fusion	GO:0045229	external encapsulating structure organization and biogenesis
GO:0006997	nuclear organization and biogenesis	GO:0046483	heterocycle metabolism
GO:0007005	mitochondrion organization and biogenesis	GO:0046903	secretion
GO:0007010	cytoskeleton organization and biogenesis	GO:0046907	intracellular transport
GO:0007028	cytoplasm organization and biogenesis	GO:0048284	organelle fusion
GO:0007029	ER organization and biogenesis	GO:0048285	organelle fission
GO:0007031	peroxisome organization and biogenesis	GO:0048308	organelle inheritance
GO:0007032	endosome organization and biogenesis	GO:0050789	regulation of biological process
GO:0007033	vacuole organization and biogenesis	GO:0050874	organismal physiological process
GO:0007049	cell cycle	GO:0050896	response to stimulus
GO:0007059	chromosome segregation	GO:0051180	vitamin transport
GO:0007154	cell communication	GO:0051181	cofactor transport
GO:0007275	development	GO:0051186	cofactor metabolism
GO:0008104	protein localization	GO:0051235	maintenance of localization
GO:0008283	cell proliferation	GO:0051261	protein depolymerization
GO:0008643	carbohydrate transport	GO:0051276	chromosome organization and biogenesis
GO:0009056	catabolism	GO:0051301	cell division

Supplemental Table 7: Gene Ontology terms to which less than 10% of the yeast genome is annotated, used for negative gold standard generation by MEFIT.

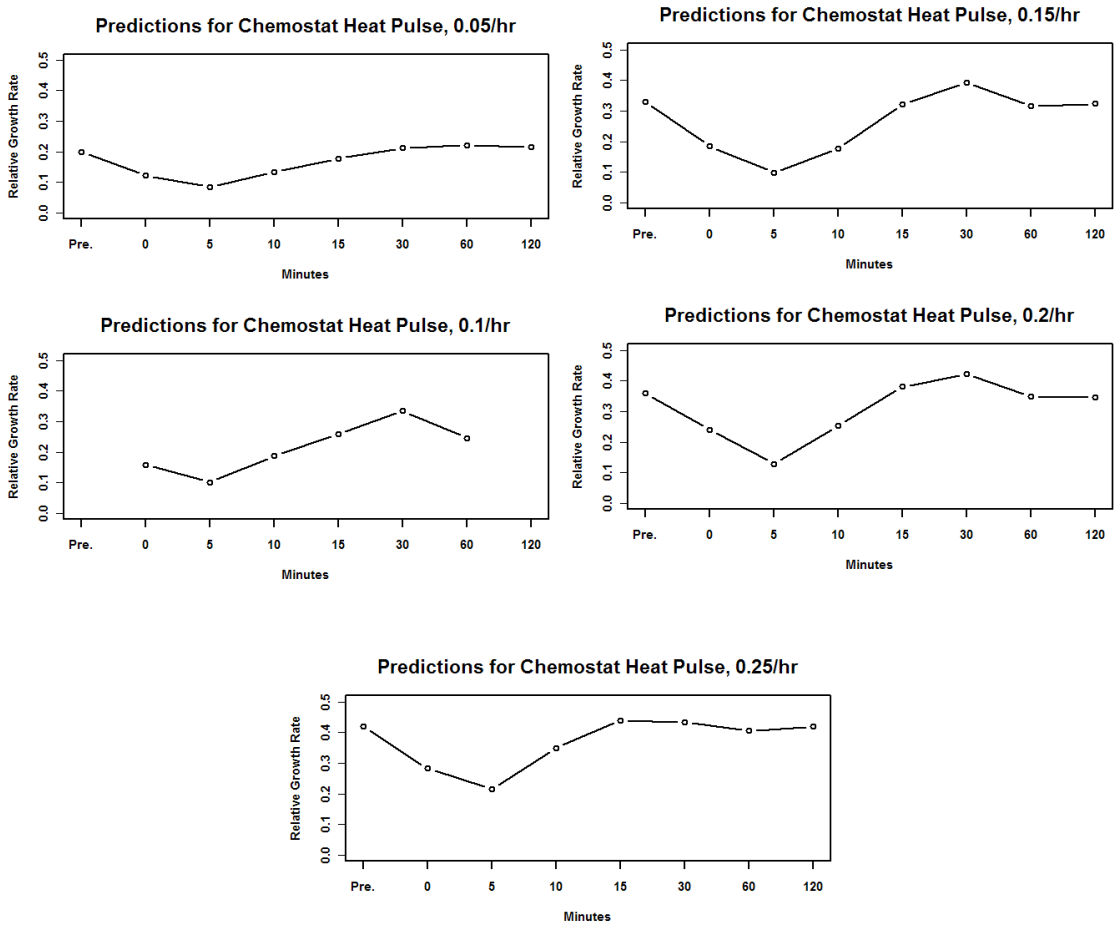
Term	Assoc. Unch. Genes	Term	Assoc. Unch. Genes
carbohydrate metabolism	972.14	RNA splicing	336.31
phosphorus metabolism	895.33	transcription from RNA polymerase III promoter	333.74
reproductive physiological process	863.52	nucleobase, nucleoside, nucleotide and nucleic acid transport	329.01
establishment of protein localization	862.03	response to pheromone	326.68
sporulation	832.73	ribosome biogenesis	322.36
autophagy	797.55	membrane fusion	314.22
one carbon compound metabolism	794.90	glycoprotein metabolism	307.35
cell wall organization and biogenesis	788.22	regulation of protein biosynthesis	307.31
chromosome organization and biogenesis (sensu Eukaryota)	773.14	establishment of RNA localization	301.47
cofactor metabolism	743.81	metabolic compound salvage	301.07
vesicle mediated transport	742.26	lipoprotein metabolism	300.53
RNA editing	734.47	translation	293.39
M phase	733.32	sulfur metabolism	280.19
cell division	732.04	transcription from RNA polymerase I promoter	270.30
biopolymer biosynthesis	731.06	vacuole organization and biogenesis	270.08
vacuolar transport	706.45	carbohydrate transport	269.92
proteolysis	692.12	mitochondrial transport	265.94
cell morphogenesis	691.10	RNA elongation	263.94
alcohol metabolism	690.93	transcription initiation	258.83
vitamin metabolism	667.45	nuclear organization and biogenesis	254.96
meiotic cell cycle	665.35	second messenger mediated signaling	250.00
lipid metabolism	658.82	regulation of DNA metabolism	246.22
positive regulation of transcription	637.66	snoRNA metabolism	241.68
energy reserve metabolism	630.33	asexual reproduction	217.43
protein complex assembly	619.45	organic acid transport	212.69
cytoskeleton organization and biogenesis	617.33	cell redox homeostasis	212.18
transcription from RNA polymerase II promoter	613.66	nucleoside metabolism	207.18
regulation of progression through cell cycle	609.89	RNA 3' end processing	202.24
ubiquitin cycle	601.92	small GTPase mediated signal transduction	200.57
response to DNA damage stimulus	594.10	peroxisome organization and biogenesis	196.31
mitochondrion organization and biogenesis	592.31	amine transport	192.19
organic acid metabolism	590.30	protein kinase cascade	178.71
pseudouridine synthesis	575.34	protein amino acid alkylation	175.79
DNA recombination	573.36	electron transport	175.23
ion transport	571.40	regulation of RNA metabolism	159.27
mitotic cell cycle	563.44	cell surface receptor linked signal transduction	152.45
negative regulation of transcription	561.83	protein amino acid acetylation	150.95
ion homeostasis	552.71	response to inorganic substance	150.53
response to osmotic stress	545.04	sex determination	136.88
mRNA metabolism	520.03	regulation of catalytic activity	134.23
response to temperature stimulus	516.43	lipid transport	128.82
nucleotide metabolism	516.34	hydrogen transport	125.50

protein processing	499.81	peptide metabolism	124.92
cellular respiration	496.70	negative regulation of protein metabolism	123.77
filamentous growth	489.90	DNA catabolism	116.25
nitrogen compound metabolism	485.11	maintenance of protein localization	111.07
tRNA metabolism	480.44	protein amino acid deacetylation	90.98
DNA packaging	480.25	secondary metabolism	87.43
RNA catabolism	472.10	fermentation	85.13
response to toxin	448.72	maintenance of cellular localization	84.47
regulation of gene expression, epigenetic	446.79	response to unfolded protein	76.93
aldehyde metabolism	444.22	organelle fusion	70.21
gene silencing	432.75	regulation of signal transduction	57.83
aging	429.83	transcription termination	55.28
nucleocytoplasmic transport	408.31	establishment of nucleus localization	52.87
nuclear transport	408.31	vesicle organization and biogenesis	51.59
heterocycle metabolism	401.12	chromosome condensation	43.30
interphase	396.30	plasmid maintenance	40.65
protein folding	387.26	mannoprotein metabolism	35.48
cell death	379.45	response to nutrient	34.81
amino acid and derivative metabolism	379.01	peroxisomal transport	33.31
rRNA metabolism	373.56	cytoskeleton dependent intracellular transport	27.98
chromosome segregation	371.42	drug transport	26.53
sexual reproduction	368.60	protein depolymerization	24.35
conjugation	368.60	response to acid	23.21
vacuolar protein catabolism	363.47	microtubule organizing center organization and biogenesis	22.50
DNA replication	360.78	intracellular mRNA localization	19.44
organelle inheritance	356.23	regulation of transport	13.75
response to starvation	349.83	cofactor transport	9.71
peptidyl amino acid modification	340.43	protein sumoylation	2.15
aromatic compound metabolism	336.93		

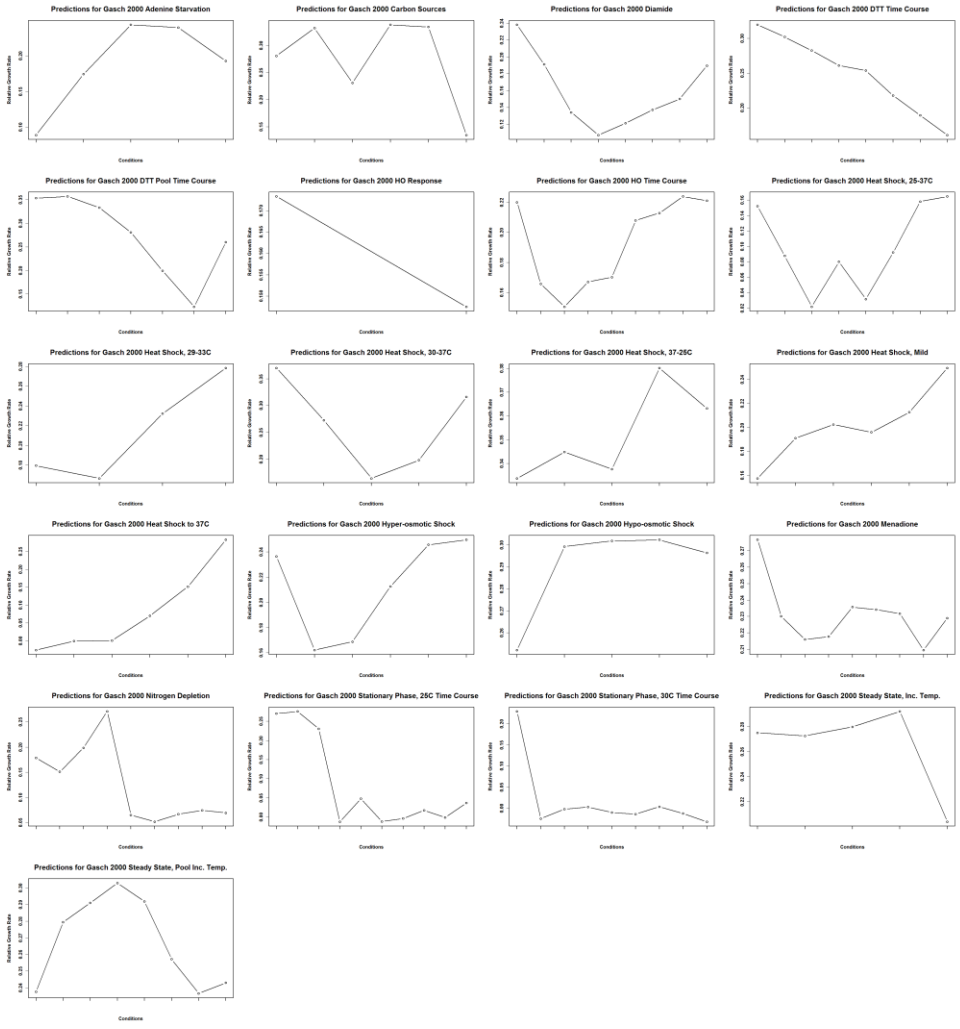
Supplemental Table 8: Association of each biological process of interest with the ~1,500 uncharacterized genes of the yeast genome. Each score represents the ratio of the average predicted probability of functional relationship between the uncharacterized genes and the set of genes known to participate in each biological area, normalized by that process's cohesiveness.



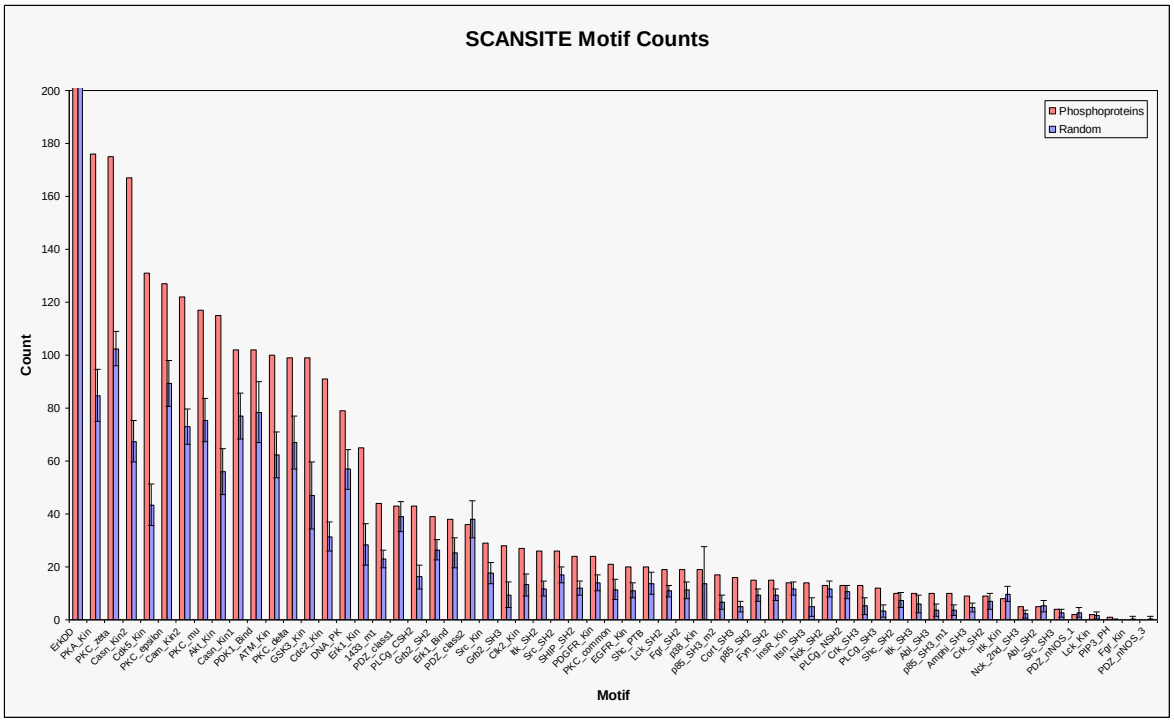
Supplemental Figure 1: Effect of clique size on NNN performance. The concatenated dataset clustered using NNN with clique size g = 3, 4, 5, and 6, ranging over n from one to 40 by increments of three. Performance varies relatively little as g is varied, with lower clique sizes trading a small amount of recall for increased precision and runtime.



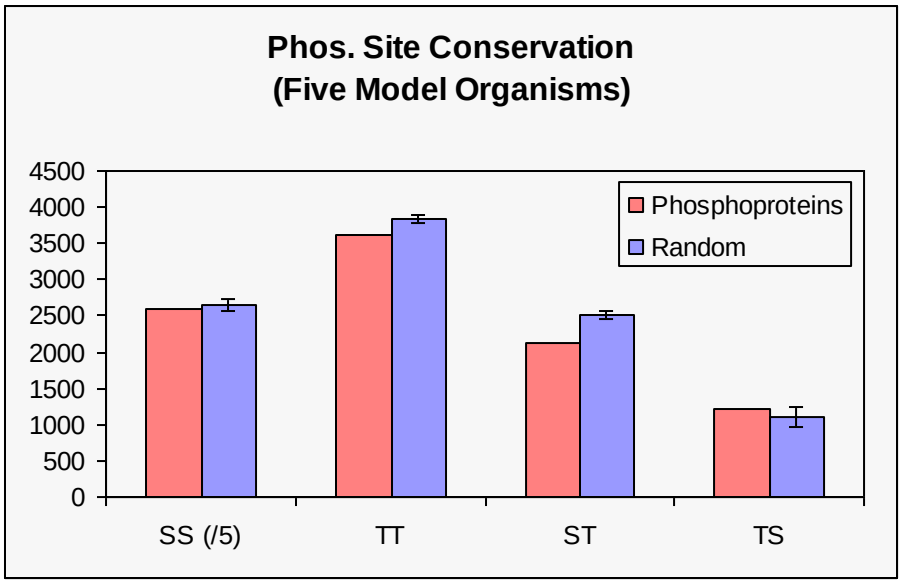
Supplemental Figure 2: Growth rate predictions for chemostat cultures subjected to a brief heat pulse at various flow rates. Microarray time courses were taken for a collection of chemostats at increasing growth rates, each subjected to a brief (<30s) heat pulse at time zero. Predicted growth rates show an immediate departure from steady state as the heat pulse is administered immediately before time zero, followed by a gradual return to steady state and regulatory overshoot. This behavior is consistent across growth rates, with the lowest growth rates potentially showing a lesser shock response due to stress tolerance.

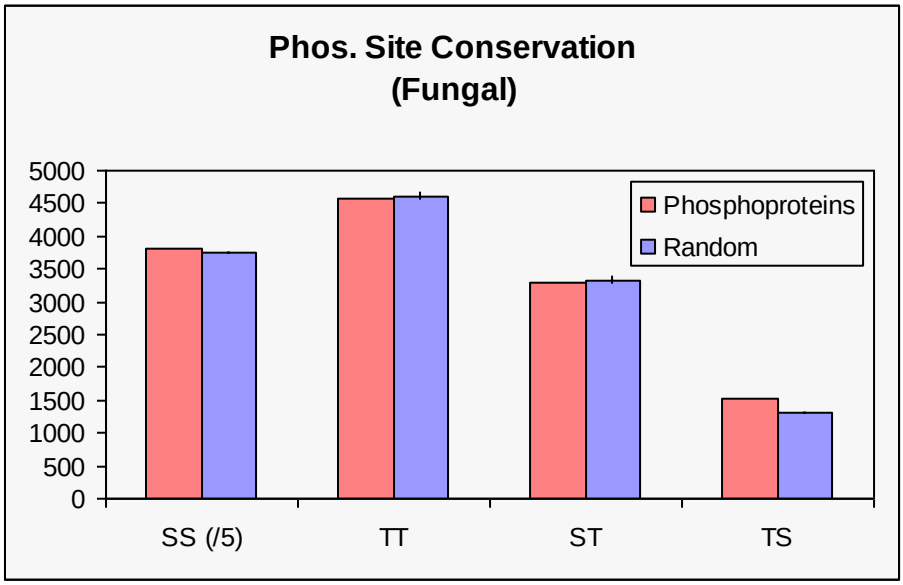


Supplemental Figure 3: Growth rate predictions for all conditions in the (Gasch, Spellman et al. 2000) stress response microarrays. These predictions are generally consistent with known yeast biology and agree with expected growth behavior; most shock time courses, including all heat shocks, peroxide, diamide, and hyper-osmotic stress, provoke an initial sharp decrease in growth rate followed by a return to initial or near-initial rate. Shorter shocks, such as DTT, menadione, and peroxide responses, capture only the rate decrease. Batch growth proceeds at a fairly constant rate until nutrients become depleted, at which point the rate decreases sharply; this pattern is also seen in intentional nitrogen depletion. Growth rates across varying temperatures peak as expected at 25C, falling off at lower and higher temperatures. Response to varying carbon sources is also as expected, with ethanol inducing the slowest growth and fructose, sucrose, and glucose allowing the most rapid. The model's inference of growth rate from microarray data alone thus allows both post hoc growth analysis (e.g. years after the original experiment) and an estimation of growth rates for cultures where it would be difficult or time consuming to measure directly.

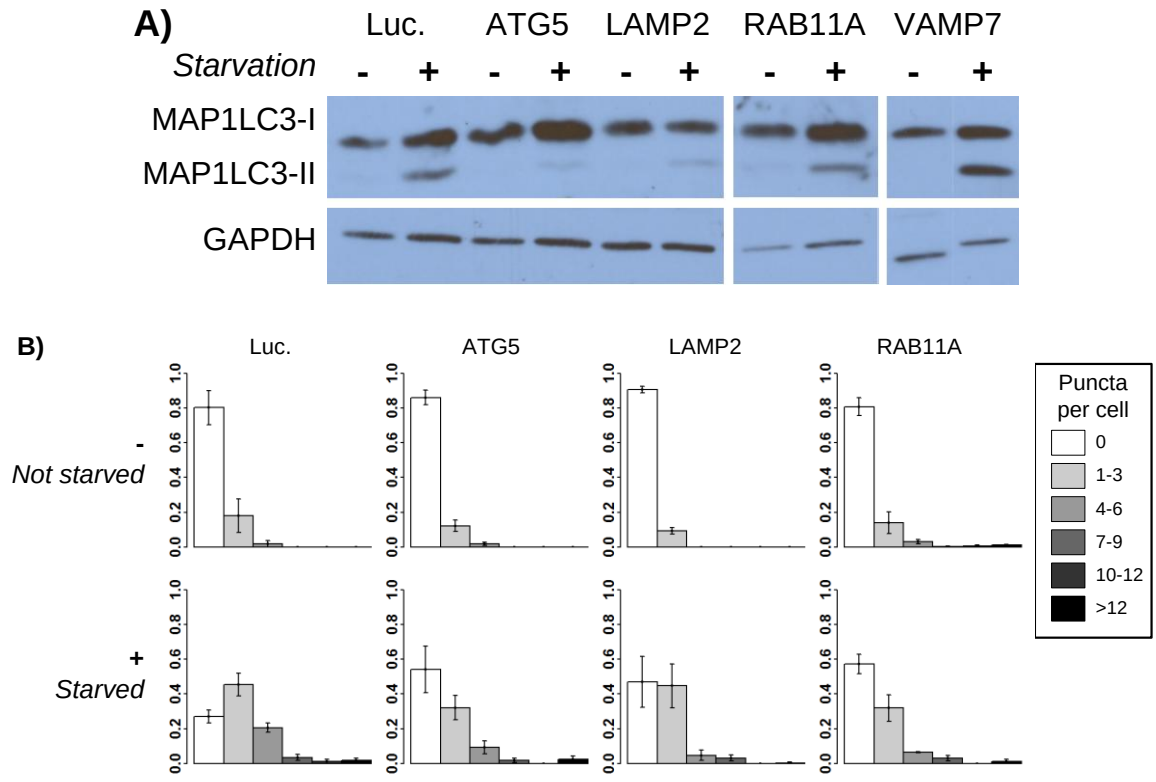


Supplemental Figure 4: Counts of each SCANSITE motif recovered in phosphopeptides relative to those drawn from identically sized random genomic samples.

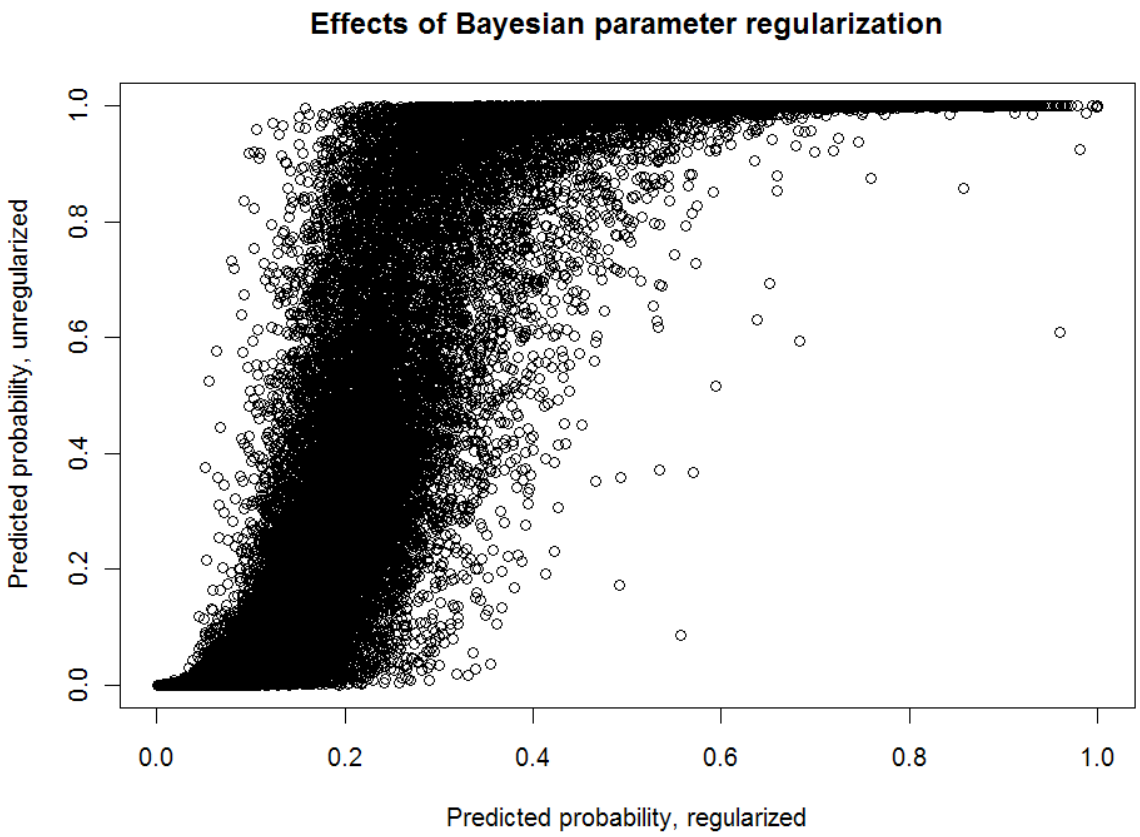




Supplemental Figure 5: Conservation of phosphorylation sites relative to identically sized random samples. SS and TT represent conserved serine and threonine sites, while ST and TS represent serines in yeast converted to threonines in another organism and vice versa. The five model organisms used were *A. gossypii*, *C. elegans*, *D. melanogaster*, *H. sapiens*, and *A. thaliana*. The fungal scan covered all genomes covered by BLASTP through the Saccharomyces Genome Database.



Supplemental Figure 6: Quantification of autophagosome formation in starved cells. MAP1LC3 is typically diffuse throughout the cytoplasm in non-starved cells. Under normal conditions, starved cells will initiate autophagy, process MAP1LC3 to the MAP1LC3-II isoform, and form punctate autophagosomes to which it is localized. We measured the degree to which this was impaired by luciferase (negative control), ATG5 (positive control), LAMP2, RAB11A, and VAMP7 siRNA depletions using immunoblotting and manual inspection of ten images for each condition (totaling 80 images). While VAMP7 knockdowns showed no effect (see Discussion), siRNA knockdowns of LAMP2 and RAB11A inhibited normal autophagy. A) Measurement of the autophagosome-bound MAP1LC3-II isoform by immunoblotting. Under a control condition (luciferase siRNA), starvation (+) induces autophagy in human fibroblasts and upregulates the autophagy marker MAP1LC3-II; this upregulation is inhibited by knockdown of some proteins required for autophagy, e.g. ATG5. B) Manual inspection of the number of puncta per cell shows decreased autophagosome formation when autophagy is impaired. Error bars indicate standard error over counts by three independent investigators viewing randomized, unlabeled images. The number of puncta increases when cells are starved under the luciferase control condition, but this increase is substantially impaired in ATG5 (positive control), LAMP2, and RAB11A siRNA conditions.



Supplemental Figure 7: Bayesian parameter regularization prevents overconfident probability estimation in the presence of many datasets. While naive Bayesian classifiers provide an accurate and efficient way to integrate hundreds of genomic datasets, they assume complete independence between all data. Violations of this assumption, which occur due to shared biological and technical signals between datasets, become increasingly problematic as the number of integrated datasets increases. We use Bayesian parameter regularization to combine each dataset's probability distribution with a uniform prior, mixing this prior in with weight proportional to the amount of information shared by each dataset. Intuitively, this results in datasets with strong, unique signals being upweighted during the integration process, while groups of datasets sharing most of the same information will be downweighted. Without regularization, a low-confidence datum contributed by many datasets can inappropriately result in a high-confidence prediction of functional relationship. Regularization downweights such shared data and results in a more biologically realistic distribution of low- and high- probability functional relationships.