INSIGHTS FROM SYSTEMATICALLY ANALYZING

MICROBIAL PHENOTYPIC PROFILES

A Dissertation

by

I-FAN WU

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Margret Glasner
Committee Members,	Deborah A. Siegele
	Rodolfo Aramayo
	Michael Polymenis
	Sing-Hoi Sze
Head of Department,	Josh Wand

August 2021

Major Subject: Biochemistry

ABSTRACT

Following classical genetic approaches to understanding gene function, high-throughput phenotyping methods have emerged as a new way of studying gene functions, especially in microorganisms, which are highly amenable to high-throughput experimental design. As more high-throughput microbial phenotype data as well as the low-throughput data become available, systematically managing, displaying, and analyzing these data become a pivotal part in discovering unknown functions for genes. In this work, I have curated some datasets for high-throughput microbial phenotype data that contain genomic-scale phenotypes from *E. coli* tested under hundreds of conditions. Next, I conducted systematic and unbiased statistical analysis of these phenotype datasets and showed that the phenotypic profiles within these datasets are highly correlated with various functional annotations. The phenotype-function correlation has also been seen when a curated cell-cycle related phenotypic profile of *S. cerevisiae* is used with Gene Ontology annotations. Furthermore, I have displayed the preliminary results of using machine learning techniques to predict gene functions using high-throughput phenotype data of complete annotations, given more functional annotations as labels. Lastly, I describe a software package written in R that is potentially useful in analyzing high-throughput microbial phenotype data.

ACKNOWLEDGEMENTS

I sincerely appreciate all the support, help and company of people around me during my almost 5-year stay at Texas A&M University, be it academic, technical, financial or mental. First of all, I would like to thank Dr. Jim Hu, who took me when I only showed strong interest in becoming a data-driven bioinformatician without too much background and expertise. He had patiently guided me throughout the process of scientific thinking and allowed much freedom for me to explore different possible directions where my dissertation can go. Although this great guy suddenly passed away on 1/23/2020 from illness, his wisdom and amazing character will always remain in my brain. Next, I would like to thank Dr. Deborah Siegele, Jim's wife, who took me, supported me financially as well as helping with everything I needed to finish my Ph.D. She is always warm, kind and willing to listen to my scientific inquiries as well as providing invaluable opinions. In my mind, she is really a wonder woman that is courageous, unbreakable, intelligent and supportive. During her hardest time, in addition to the unwarned swarm of COVID-19 pandemic, she really stood out in coordinating with everything I needed. I really appreciate Dr. Ry Young for lecturing the critical analysis class when I was in my first year, as well as nominating me as a Texas A&M Heep Fellowship winner. His generous support really helped me focus on my graduate research projects. I would like to thank our dear past lab members Curtis Ross, Sandra LaBonte, Suzanne Aleksander and Jolene Ramsey. Your presence and kindness made Hu/Siegele lab a vibrant research environment.

Throughout my stay at Texas A&M, I always had 5 committee members. I have benefited much because of additional support from this strong group of people. I appreciate the help from Dr.

Margaret Glasner, Dr. Roldolfo Aramayo, Dr. Craig Kaplan, Dr. Michael Polymenis and Dr. Sing-Hoi Sze in pushing to finish this dissertation. I would like to thank Dr. Mumita Karmakar in the department of statistics and Dr. Bobak Mortazavi in the department of computer science for providing useful suggestions regarding potential biostatistics/machine learning methods.

I appreciate the help and company from friends in BGA, including but not limited to Ting-Yi Wang, Rafael Almanzar, Tingfeng Guo, Laith Harb, Ran Meng, Menqui Jiang. These folks have made my Ph.D. study much more colorful. I also thank Jason Lin, Jack Guo from outside of our department for extra support. Finally, I would like to thank my family members and friends in Taiwan for essential mental support: My mom, Min-Hui Lai, my dad, Wen-Hsiang Wu, my aunt, Amy Lai. They all provided me shelter when I felt vulnerable. My cute girlfriend, Yen-Ting Lu, has been accompanying me during some very hard times of my career; My college buddies, Robert Hsieh and Chun-Hao Chang, who have amazing characters and have been accompanying me for more than two decades.

Sincerely & Thank you all,

Peter I-Fan Wu

CONTRIBUTORS AND FUNDING SOURCES

Contributors

The work presented in this dissertation was supervised by the committee consisting of Drs. Jim Hu, Deborah Siegele, Margret Glasner, Aramayo Rodolfo, Michael Polymenis, Sing-Hoi Sze and Craig Kaplan.

Curtis Ross helped conduct the systematic analysis of metabolic pathways in chapter 2, while Dr. Jim Hu and Dr. Siegele help the revision of the manuscript. In chapter 3, Dr. Siegele help the revision of the manuscript. In chapter 4, Dr. Michael Polymenis and Dr. Rosa M. Bermudez performed have done most of the result. In chapter 6, Dr. Deborah Siegele helped test the usability of the software.

Other work not listed above was completed independently by the student I-Fan Wu.

Funding Sources

This work was generously supported by **NIH R01-GM089636-04** from National Institute of Health, and the Heep fellowship from Texas A&M University

Table of Contents

Table of figures and tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

CHAPTER 1. INTRODUCTION

MICROBIAL PHENOTYPIC PROFILING

Microbial genetics is useful in understanding complicated biological processes and molecular functions in higher organisms including human disease models. To understand the functions of genes: genetic, molecular biological and biochemical approaches are commonly performed. In parallel with these low-throughput approaches, it is also argued that the large amount of phenotype data generations by high-throughput experiments can be used to elucidate the functions of genes (Bochner, 2009). Phenotypes are observable traits given a defined genotype and a defined environment. They are dynamic properties that change with the alteration of the environment, just like the color change of a chameleon in order to camouflage itself or to display aggression. In microorganisms, phenotypes can simply be the growth rate in the presence of antibiotics, various environmental stresses or different nutrients. They can be behavioral, such as forming a biofilm, swarming using flagella, exhibiting chemotaxis, being predatory, or having different lifestyles like free living or attached lifestyle. They can also be morphological, including cell shape, length, width, volume and so on. In addition, some molecular functions or observations at the molecular level can also be viewed as phenotypes, for example, the presence of transporter activity, DNA breaks or gene silencing events.

Given the large number of possible phenotypes, it may be difficult to exhaustively define the landscape of all possible phenotypes of an organism. However, if many phenotypes of a given organism can be measured simultaneously using high-throughput approaches, the possibility for gaining insight into gene functions increases, because not only do the individual phenotypes provide information related to functions, but the co-occurrence of multiple phenotypes in the same environment may help connect the observed phenotypes to underlying cellular functions.

To better exploit large-scale phenotype data for informative insights, using strategic ways to record phenotypes is also important, since it permits comparison not only within a study, but will also allow comparison of results from different experiments for the same organism, or even comparison of results for different organisms. For example, recording that a mutant strain has a cell length of 6 micron might indicate abnormal growth for *Escherichia coli*, but be within the normal size range for the budding yeast *Saccharomyces cerevisiae*. However, if the phenotype record also includes a field that contains the property 'increased cell length,' it will be easier to make comparisons between organisms.

Recording information about the assays used for detecting phenotypes also adds value to the phenotype records. For example, the Evidence Code Ontology (ECO) can be used to identify both the type of evidence used and whether a person or a computer has made the annotation (Giglio et al., 2019). For instance, if a phenotype annotation includes the evidence code "high-throughput mutant phenotype evidence used in manual assertion" (HMP), we understand that the annotation was based on manual review of results from a large-scale experiment and the phenotype probably needs to be validated by low-throughput approaches.

Among the rising number of Omics technologies, the “Phenome“, in contrast to genome, proteome or metabolome, consists of all the phenotypes expressed by an organism (Thessen, Walls, et al., 2020). As “Phenotypic profiles” can be defined as a series of observed phenotypes associated with a given organism in different environments, “Phenotypic profiling” refers to the methods used to generate and measure the phenotypes. By systematically collecting phenotype data in large scale, the “Guilt by Association” rationale can be used to bridge phenotypes and gene functions. In other words, if a mutant of an unknown gene shows a similar pattern across hundreds or thousands of phenotypes compared to mutants of genes of known function, it is highly probable that the uncharacterized genes share some level of functional similarity to those genes that have been well characterized.

Phenomics, the collective phenotypic expression pattern of an organism, can serve as a natural complement to genome sequencing (Houle et al., 2010). In Acin-Albiac et al., (Acin-Albiac et al., 2020) Phenomics was mentioned as one of the irreplaceable Omic sciences, due to the fact that phenotypes cannot be fully explained by genomics and transcriptomics. Understanding phenomic data is key to understanding growth, fitness, development and disease models. In recent years, there has been an increase in the number of high-throughput bacterial phenotypic profiling studies (Nichols et al., 2011; Price et al., 2018; Thompson et al., 2019). The many examples presented in these studies show that phenotypic profiling is an effective tool to understand the functions of genes. However, additional systematic and unbiased analyses of these phenotypic datasets may be able to provide additional insights.

The usefulness of high-throughput phenotypic profiling is enhanced by biocuration efforts that compile information from published papers and put the data in a form that is computable, as well as making it more accessible to researchers. The availability of high-quality functional annotations makes it much easier to generate hypotheses about gene function based on results from high-throughput phenotype studies. Examples of biocuration effort that benefits the biological scientific community includes the Gene Ontology Consortium (The Gene Ontology Consortium, 2017), the Monarch Initiative (Shefchek et al., 2020), the UniProt group (UniProt, 2019), the Ontology of Microbial Phenotypes (OMP) group (Chibucos et al., 2014) and many more. These databases are constantly evolving by accruing more data and developing new analysis tools.

Despite the increasing amount of microbial phenotype data available, especially data from high-throughput experiments, systematic curation of these datasets has not been performed. In addition, there is a need for statistical or analytical methods tailored to handle these large datasets. It may be possible to adapt some of the computational and statistical tools developed for omics technologies, such as whole genome sequencing or metabolomics, which generate large amounts of data, because many omics datasets share a common two-dimensional matrix-like data structure.

APPLICATIONS OF MICROBIAL PHENOTYPIC PROFILING

In addition to the power of predicting new gene functions, high-throughput microbial phenotype data can directly contribute to many other applications. For example: analyzing the phenotypic responses of mutant strains to a series of antimicrobial chemicals can help identify which compounds that target the same cellular process (Nichols et al., 2011). Studying the phenotypes associated with members of the human microbiome can contribute to our understanding of the mechanisms underlying a disease (Ha et al., 2020). Analyzing growth rates of mutant strains of *Mycobacterium tuberculosis* in different media identified new genes required for fitness in a low-iron environment (Dragset et al., 2019). Ethanologenic strains of yeast that may be useful for industrial applications were identified by screening the phenotypes of wild strain of yeast. (Farooq et al., 2018). Identifying key differences between the phenotypic profiles of wild and domesticated strains of the bacterium *Caulobacter* led to the conclusion that wild environmental conditions result in more variable phenotypic profiles (Hentchel et al., 2019). Analyzing the phenotypes of photosynthetic bacteria will potentially lead to better biomass production (Abernathy et al., 2017; Alfred et al., 2012). Results from analyzing the carbon-utilization phenotypes of *Lactobacillus* may lead to improvement of its food and probiotic applications (Ceapa et al., 2015). Recently, phenotypic profiles of strains with point mutations rather than knock-out alleles were used to reconstruct protein 3D structures (Braberg et al., 2020; Wang, 2020). In general, if there is a specific target function or biological process, collecting a spectrum of phenotypes has the potential to identify unique phenotypic patterns and improve our understanding of the function or process that is occurring.

In summary, there are many important applications of microbial phenotypic profiling. The usefulness of the profiling not only depends on using the right analytical methods, but also on the scalability and reliability of the phenotype data. Hopefully, with sophisticated versions of molecular tools like Next Generation Sequencing, CRISPR/Cas systems (Tarasava et al., 2018), and Multiplex Automated Genome Evolution (MAGE) (Wang et al., 2009; Wrighton, 2018), more phenotype data for larger numbers of mutants tested in many more conditions will become available.

BIOCURATION IS AN IMPORTANT FIELD TO GET STRUCTURED BIOLOGICAL DATA

With the soaring growth of biological data, biologists spend increasingly more time searching for information relevant to their studies. Better data management would reduce the time needed for data or information retrieval and minimize the chances of experiments being done multiple times in an unnecessary manner. Biocuration is an important part of the solution to this problem. Biocuration can be defined as "the activity of organizing, representing and making biological information accessible to both humans and computers" (Howe et al., 2008). Salimi *et al.* (Salimi & Vita, 2006) describes a biocurator as: A person who is able to comprehend scientific data and annotate it following curation guidelines while maintaining the integrity of the data. However, greater awareness of the importance of biocuration is still needed (Biocuration, 2018). In addition, more effort in curating more data is needed in order to best represent the up-to-date biological knowledge from literature resources.

There are many widely used databases, which are powerful resources for the biological sciences, that were built by and continue to depend on professional biocurators. Some of these databases are listed in table 1.1.

Database	About	Website	Reference
The Gene Ontology Resource	Functional annotations for multiple species are available as downloads	http://geneontology.org/docs/download-go-annotations/	(Ashburner et al., 2000; Gene Ontology, 2021)
Ensembl	Genome annotations source mainly on vertebrates	www.ensembl.org	(Yates et al., 2020)
Catalogue of Somatic Mutations in Cancer (COSMIC)	Expert-curated database of somatic mutations in human cancers	cancer.sanger.ac.uk/cosmic	(Tate et al., 2019)
EcoCyc	A comprehensive database for *E. coli* K-12, with primarily manually curated data	www.ecocyc.org	(Keseler et al., 2017)
Subtiwiki	A wiki-based collaborative resource for the *Bacillus* community	www.subtiwiki.uni-goettingen.de	(Zhu & Stulke, 2018)
Saccharomyces Genome database (SGD)	Professionally curated model organism database for *Saccharomyces cerevisiae*	www.yeastgenome.org	(Cherry et al., 2012)
PomBase	Professionally curated model organism database for *S. pombe.*	www.pombase.org	(Lock et al., 2019)
DictyBase	A comprehensive database for D. *discoideum* (slime mold)	www.dictybase.org	(Fey et al., 2019)
WormBase	A professionally-curated model organism database for the nematode *Caenorhabditis elegans*	www.wormbase.org	(Harris et al., 2020)
Zebrafish Information Network (ZFin)	A comprehensive database for zebrafish *D. rerio*	www.zfin.org	(Howe et al., 2021)
FlyBase	A comprehensive database for fly *D. melanogaster*	www.flybase.org	(Larkin et al., 2021)

Mouse Genome Informatics (MGI)	A comprehensive database for mouse *M. musculus*	www.informatics.jax.org	(Bult et al., 2019) (Smith et al., 2019) (Krupke et al., 2017)
Monarch Initiative	The largest phenotype-genotype annotation database for human and other mammals I think the latest version incorporates information from flybase, wormbase, SGD, and others.	www.monarchinitiative.org	(Shefchek et al., 2020)

Table 1.1. List of widely used biocuration databases.

Who serves as the main contributors in making biological annotations? Usually, well trained, Ph.D. level biocurators with many years of wet-bench experience are able to generate accurate annotations in a very efficient manner. Community annotation, or crowdsourcing, is also a good alternative (Hanauer et al., 2017; Thessen, Grondin, et al., 2020), although it is still under development in many areas of biocuration, due to the time and effort and expert knowledge needed to make a complete, quality annotation. This being said, what are the common strategies for making large quantities of annotations? Annotations are made to genes, gene products, or mutant strains, etc, and in addition to an annotation term with a unique identifier, annotations typically include a reference (the publication or where the raw data come from) and an "evidence code" that describes the type of experiment an annotation is based on and gives annotations different levels of confidence. When high-throughput experimental approaches are used, additional effort may be needed to retrieve the original data and metadata. In addition, mapping the raw data or inferences from original publications to the appropriate annotation terms is nontrivial. Furthermore, maintaining and updating annotations is another challenge: In order for these annotations to benefit the general scientific community, they need to be organized in ways people can easily browse and get useful information from.

Making annotations that can be easily compared with other annotations is important. In many cases, biological annotations are made using terms that come from an ontology (Smith et al., 2007). In modern day Information Science, an ontology is a structured vocabulary whose terms represent a domain of knowledge. An annotation made from an ontology is amenable to computational reasoning - that is, it is understandable by modern computers - because its terms are connected by logical relationships, such as “is a” or “part of”. Ontologies are often Directed-Acyclic Graphs (DAGs). In these graphs, terms representing concepts are implemented as nodes that link to each other in a hierarchical order, allowing traversal from the most detailed levels of knowledge upwards to the top of the graph (the root), where the most generic concepts reside. As the number of annotations increases, the need to use computer-based reasoners to find connections between the objects being annotated also increases. For example, it would be easy for a person to gauge the relationship between two gene products: one that is annotated with the term “transcription regulator activity” and the other with the term “positive regulation of DNA-binding transcription factor activity,” since the latter is obviously a descendent of the former. However, as the number of annotations made to each gene grows, and as the number of genes being compared becomes large, computers can outcompete humans by delivering results in a timely manner. Table 1.2 lists some representative projects or model organism databases that make use of biological ontologies. As described in Smith et al., 2007 (Smith et al., 2007), the proliferation of biological ontologies can cause problems, because terms within one ontology are not necessarily interoperable (easily related to) with terms in other ontologies. The Open Biological and Biomedical Ontology (OBO) foundry (Smith et al., 2007) was developed to overcome this problem by providing rules and best practices to help those biological ontologies stay interoperable.

Ontology	Database where the ontology is used	Database website	Reference for the ontology
Gene Ontology	AmiGO 2	www.amigo.geneontology.org	(Gene Ontology, 2021)
Phenotype Ontology (HPO)	Monarch Initiative	www.monarchinitiative.org	(Kohler et al., 2019)
Ascomycete Phenotype ontology (APO)	Saccharomyces Genome database (SGD)	www.yeastgenome.org	N/A
Fission Yeast Phenotype Ontology (FYPO)	PomBase	www.pombase.org	(Harris et al., 2013)
Ontology of Microbial Phenotypes (OMP)	Microbial Phenotypes Wiki	www.microbialphenotypes.org	(Chibucos et al., 2014)

Table 1.2. List of ontologies used in different biological databases.

Is biocuration worth the cost? Using *E. coli* biocuration as an example, Karp et al. (Karp, 2016) estimated that over a 5-year period of time, the EcoCyc database costs less than 1% of the overall cost of the research projects that had generated the experimental results, which was estimated to be one-tenth of the coffee break money for researchers carrying out the research. The International Society for Biocuration (International Society for Biocuration, 2018), pointed out that direct operational fees of European Bioinformatics Institute (EMBL-EBI) benefits the users and funders more than 20 times compared to the original cost.

In conclusion, the databases mentioned above represent only a portion of the biocuration circle. As more data are generated thanks to high-throughput experimental approaches, the lower cost of next-generation sequencing, and booming growth of biotechnological industry as a whole, more biocuration is needed, and thus, expansion of the impactful public databases that will be freely available are expected to occur. In terms of managing these valuable curated data, management of metadata also plays an important role. In 2016, Wilkinson *et al*. (Wilkinson et al., 2016) proposed the a data principle to leverage the scholarly data in general, pointing out that making data Findable, Accessible, Interoperable and Reusable (FAIR) can serve as a good standard for handling digital data, which can possibly become the long-term, high-end guideline for the biocuration field as a whole.

IMPORTANT RESOURCES FOR PHENOTYPE ANNOTATIONS

Since phenotypes are produced directly or indirectly from gene functions, the large-scale collection of phenotypes can be a powerful tool for inferring functions. Due to the explosive speed at which low and high-throughput phenotype data are appearing, many biocuration groups have been actively gathering these data in a structured manner. For humans and other animals, there is the Monarch Initiative (Shefchek et al., 2020). For fungi, there are (SGD) (Cherry et al., 2012), and Pombase (Lock et al., 2019). For plants, there is work done by Oellrich *et al.* (Oellrich et al., 2015) and Cooper *et al*. (Cooper et al., 2018). For bacteria, there are the

Ontology for Microbial Phenotype (OMP) group (Chibucos et al., 2014) deposited in the Microbial Phenotypes wiki (https://microbialphenotypes.org), BacDive (Reimer et al., 2019), Subtiwiki (Zhu & Stulke, 2018) and others. There is also VEuPathDB that stores host-eukaryotic pathogen related phenotypes (https://veupathdb.org). The resources above, and others that have not been mentioned, make both raw data and curated data available.

COMPUTATIONAL AND STATISTICAL APPROACHES FOR HIGH-THROUGHPUT MICROBIAL PHENOTYPE DATA

High-quality microbial phenotype data are not insightful until appropriate computational processing or statistical analysis allow interpretation of the results. While there is no single best method for extracting functional insights from these data, there are many existing tools that can help to extract insights (Grys et al., 2017; Xu & Jackson, 2019). One general approach is "guilt by association", in which mutant genes are assigned new functions because they show similar phenotypic profiles to mutants of genes whose functions are relatively well known. Usually, high-throughput phenotype data come in the form of a two-dimensional matrix, where one dimension is the mutated genes, and the other dimension is the phenotype observed for each mutant in one or many growth conditions. The phenotypic profile, or the "phenotypic signature", is the series of phenotypes measured for that mutant. Since phenotypes are direct/indirect consequences of gene functions, the more phenotypic variables/features a study is able to incorporate, the richer the extractable functional information is expected to be.

One common technique used for “guilt by association” is to calculate the pairwise phenotypic similarity of all the mutants, using similarity or distance metrics, such as Pearson Correlation Coefficient (PCC) (Nichols et al., 2011), Spearman’s Rank Correlation Coefficient (SRCC), and Mutual Information (MI) (Priness et al., 2007). Each similarity metric has its own strength in capturing different information from the phenotype data: PCC best measures linear relationships between two phenotypic profiles; SRCC measures whether a phenotype profile decreases or increases monotonically with others, and is not prone to outliers like PCC is; MI measures phenotypic profile similarity based on entropy, which is based on the probability of occurrence and co-occurrence of particular phenotypes.

More sophisticated computational/statistical approaches for using phenotypes to predict functions involve machine learning methods (Grys et al., 2017; Xu & Jackson, 2019), which consist of supervised, unsupervised, and less commonly, semi-supervised learning. The central concept of these machine learning methods, which differ from calculating pairwise similarity, is that they aim to extract common patterns from subgroups of the input data and assign new roles for the unlabeled data points, instead of being limited to comparisons within pairs.

For supervised learning, there are many different methods to perform classification of functions. The simplest method is Logistic Regression, which is no different from linear regression except the output goes through an exponential function that transforms it into a range from 0 to 1, which can be interpreted as an odds ratio (Anderson et al., 2003). Extensions of Logistic Regression include three models: Ridge Regression (Marquardt & Snee, 1975), which adds an L2 penalty to regularize the fit; Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), which adds an L1 penalty to not only regularize but also tries to shrink number of variables; and Elastic Net, which includes both L1 and L2 penalty to efficiently deal with situations of having too many variables, where many of them might be correlated (Zou & Hastie, 2005).

The decision tree method is a nonparametric (there is no fixed-sized number of parameters), hierarchical series of binary classification sets that can be easily interpreted. A decision tree uses the variables that are most effective in separating data into desired groups. It is built by determining the variables that can effectively separate the target classes, usually by calculating the information gain via Entropy or by Gini Index, and assign the "decision variables" hierarchically (Song & Lu, 2015).

The Support Vector Machine method uses a decision boundary to separate data points that can be more easily separated when a kernel function brings them into a higher dimension. A decision boundary is a hyperplane, which is defined as a line; a plane of two, three or higher dimension that is one degree lower than the sample space (Noble, 2006). Unlike logistic regression and its extensions, Support Vector Machine doesn't require a function to calculate the outcome for classification of each data points. Rather, it only requires the optimal hyperplane to be determined in order to separate data points into different classes.

There are supervised learning methods that belong to the Ensemble Learning category; these methods aggregate multiple models to make final decisions (Rokach, 2019; Tan & Gilbert, 2003;

Wang, 2006). These methods include Random Forest (Breiman, 2001; Fabris et al., 2018) and Boosting (Babajide Mustapha & Saeed, 2016; Schapire Robert & Freund, 2013). Random Forest is a method built from many decision trees. First, each decision tree is built by randomly picking some variables and samples. Second, classification is done by majority voting the decisions from these trees. Compared with a single decision tree, the Random Forest considers variables that are weaker classifiers. By incorporating many weak classifiers, it forms a much stronger classification system than a single decision tree. Like Random Forest, Boosting also uses many decision trees, but instead of taking a majority vote from the many decision trees at the end, it sequentially links many trees so that the output from one decision tree gives the input for the next tree. The final decisions are made by passing the improved residuals of many trees in order. Empirically, by using trees that are usually of shallow depth in boosting, a much more robust model is formed compared to decision trees.

Convolutional Neuro Network (CNN), a type of deep neural network (Albawi et al., 2017), has recently emerged as a very popular method for image recognition in both academia and industry. The CNN is typically built by arbitrarily linking many layers of perceptrons (nodes) by linear regression and certain activation functions, for example, Rectified Linear Unit (RelU) function, and trained with many epochs – the number of rounds for the training process. Although it works particularly well for image recognition, it is also a general machine learning method where the input can range from simple one-dimensional data to multidimensional data. The major weakness of this method is that it can reach very high prediction accuracy but without simple interpretability.

For unsupervised learning, there are many methods that aim to categorize unlabeled data in order to reduce the number of dimensions or to cluster data points sharing similar attributes or patterns. Principle Component Analysis (PCA) (Pearson, 1901) provides a way to reduce the number of significantly variable dimensions (here the dimensions are the phenotypes) by condensing most of the variation of all variables within the first couple of new transformed variables. Similarly, when variables (phenotypes) are categorical rather than continuous, Multiple Correspondence Analysis (MCA) can be used as the "categorical" version of PCA (Abdi & Williams, 2010);

For high-dimensional data not easily separable by a given hyperplane, t-Distributed Stochastic Neighbor embedding (t-SNE) (Hinton, 2008) assigns pairwise probabilities based on pairwise distances of points in high dimensional space, then projects the points onto lower dimensional space, and transforms the data into clusters that are easy to visualize. Similar to t-SNE, Self-Organizing maps (SOM) reduces data of high dimension to lower dimension, but with the help of an artificial neuro network, in which the elements of the neuro network compete against each other for the opportunity to respond to the input (Akman et al., 2019; Kohonen, 1990).

K-means Clustering (Kanungo et al., 2002) is a clustering method based on a pre-specified number of clusters. It is often used when there is prior knowledge that indicates that there are some distinct groups. K-means clustering finds the optimal solution by minimizing the within-cluster variances.

Hierarchical Clustering (Ward, 1963) is a widely used method for comparing gene expression profiles and constructing phylogenetic trees. First, a pairwise distance matrix (or equivalently, similarity matrix) is obtained. Second, the tree grows by iteratively picking and merging the least distant pairs until everything merges into one branch, a so-called agglomerative, or bottom-up approach. There is also a type of Hierarchical Clustering that generates a tree by a divisive approach (Rousseeuw, 1990).

For clustering data where some data points might be involved in multiple clusters, in other words, are not mutually exclusive when being grouped, Gaussian Mixture Models (GMM) (Reynolds, 1995) can be applied. This method assumes that each cluster is a multivariate normal distribution. Based on this assumption, it tries to estimate the optimal means and standard deviation for these distributions.

I described the usage of several machine learning methods in chapter 5. For supervised learning, I tested Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, CNN. For unsupervised learning, I have tested PCA, t-SNA and Self-Organizing Maps.

RESOURCES FOR HIGH-THROUGHPUT MICROBIAL PHENOTYPES

High-throughput microbial phenotype data are good sources to extract functional inferences from, since collecting data from microorganisms are often more scalable and less prone to moral issues compared to higher species like multicellular plants and animals. Often, the process of gathering these phenotype data is accompanied with sophisticated quality control methods to reduce noise and demonstrate significance (Collins et al., 2006). Recently, there have been many high-throughput phenotypic screens from bacteria and fungi, including the model organisms *Escherichia coli, Bacillus. subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe*. To systematically collect functional information for the genes of model organisms, single gene knockouts, knockdowns, or overexpression strains are constructed (Koo et al., 2017; Lian et al., 2019). While construction and testing of double gene mutant libraries can be used to detect genetic interactions. (Koo et al., 2017). I have described some of these approaches in detail below.

In generating phenotypic profiles of *E. coli*, the Keio collection is widely used. The Keio collection was constructed by replacing all non-essential genes with a kanamycin resistant cassette (Kan^R), resulting in 3985 single gene mutants (Baba et al., 2006). Another method that generates single knockouts efficiently is RB-TnSeq, where a random transposon insertion is used to disrupt a gene (Wetmore et al., 2015). Recently some methods were described to create knockdowns instead of knockouts using CRISPR interference (CRISPRi), which is based on a truncated CRISPR-Cas9 system where Cas9 is changed to dCas9 that lacks endonuclease activity. The CRISPR-dCas9 can block transcription elongation to knock down expression of genes in the same operon (Larson et al., 2013). As opposed to knockout/knockdown approaches that aim to study loss of functions, dual-barcoded shotgun expression library sequencing (Dub-Seq) provides a platform to “knock in” genes of interest for studying gains of function (Mutalik et al., 2019).

In *B. subtilis*, homologous recombination was used to construct genome-wide single gene knockout strains where the deleted gene is replaced with either a Kan^R or an erythromycin (Em^R) cassette (Koo et al., 2017). A method to generate double mutants was also described (Koo et al., 2017). The use of CRISPRi to knock down expression of essential genes in *B. subtilis* was described recently (Peters et al., 2016).

In *S. cerevisiae*, genome-wide single-gene knockouts were constructed using homologous recombination to replace each gene with a *KanMX* gene cassette (Giaever et al., 2004; Winzeler et al., 1999). The single-gene knockouts were made in both haploid and diploid strains.

Many of the phenotypic profile data generated using knockout libraries may contain rich information about gene functions that is worth further investigation, because many of them come in high-throughput and are structured data. I have identified many of the major high-throughput phenotype resources that contain large numbers of mutant phenotypes for microorganisms as shown in table 1.3:

Organism	Types of phenotype	Mutants tested	Conditions	Reference
E. coli	Fitness scores by imaging colony sizes	3,979 single knockouts from Keio collection (Baba et al., 2006)	Nutrients, chemicals and stress giving 324 conditions	(Nichols et al., 2011)
E. coli and 31 other bacteria	Fitness scores by RB-TnSeq	Genome-wide single knockouts. For E. coli there are 3,789 single mutants	Nutrients, chemicals and stress giving 173-194 conditions for each bacterium	(Price et al., 2018)
E. coli	Cell morphological features from image data	3,979 single knockouts from Keio collection (Baba et al., 2006)	21 morphological features including cell length, cell width	(Campos et al., 2018)

E. coli	Phage-host interaction phenotypes	Genome-wide single knockouts using CRISPRi and RB-TnSeq; Genome-wide single knockins using Dub-seq	Infections from 14 phages	(Mutalik et al., 2020)
E. coli	Fitness scores from Biolog phenotype microarray (Bochner et al., 2001)	3,796 single knockouts from Keio collection (Baba et al., 2006)	30 different carbon sources	(Tong et al., 2020)
B. subtilis	Fitness scores by imaging colony sizes	Knock-down library of 258 essential genes based on CRISPRi	93 different chemical conditions	(Peters et al., 2016)
B. subtilis	Fitness, Competence, Sporulation…etc	2 genome-wide single knockouts that are marked with Em^{R} and Kan^{R}	Carbon or nitrogen sources, cold condition…etc	(Koo et al., 2017)
S. cerevisiae	Fitness scores by tag hybridization from competitive growth	Genome-wide single knockouts	Rich medium, different nutrient availability conditions, antifungal nystatin, other stresses.	(Giaever et al., 2002)
S. cerevisiae	146,129 Literature based phenotype annotations			Saccharomyces Genome Database (SGD) (Cherry et al., 2012)
S. cerevisiae	Fitness scores by tag hybridization from competitive growth	Genome-wide single knockouts	726 Chemical conditions applied to homozygotes and 418 to heterozygotes	(Hillenmeyer et al., 2008)
S. cerevisiae	Fitness scores by tag hybridization from competitive growth	1,095 single knockouts for essential genes and 4,810 homozygous single knockouts	Conditions treated with one of the 3250 small molecules	(Lee et al., 2014)
S. cerevisiae	Literature based phenotype annotations, where 42% are growth phenotypes and 53% are expression phenotypes			Yeast Phenome Database (www.yeastphenome.org)
S. Pombe	80,781 Literature based phenotype annotations			(Lock et al., 2019)

Table 1.3. List of high-throughput microbial phenotype datasets/resource hubs

AIMS

As described in the introduction, the process of collecting, organizing, curating and analyzing phenotype data is indispensable for answering questions about gene functions that might be unanswerable using genomics, proteomics or metabolomics methodology. It also has the potential to provide a positive feedback loop between classic biochemical/genetic experiments and computational approaches that make functional prediction. Therefore, in this work, I aim to curate recent high-throughput microbial phenotype data and develop methods that help to systematically analyze microbial phenotypes in order to draw insights for genes of unknown functions. In Chapter 2, I present the results of systematically re-analyzing the data from Nichols *et al*. (Nichols et al., 2011), and highlight interesting functional insights based on unbiased statistical approaches (P. I.-F. Wu et al., 2021). To follow up on this work, Chapter 3 describes the analysis of two additional high-throughput *E. coli* phenotypic profile datasets, and the integration of all three datasets. I have used similar statistical approaches as described in Wu *et al.* (P. I.-F. Wu et al., 2021) and some ontology-based analytical methods to systematically draw functional insights. In Chapter 4, I discuss my contribution to the work that identified phenotypic associations among cell-cycle related genes in *S. cerevisiae* using functional annotations made with the Gene Ontology (Bermudez et al., 2020). In the work described in Chapter 5, I show that using well-annotated genes with mutually exclusive labels can effectively train models to predict functions, with the help of machine learning. In the last results chapter of this dissertation, Chapter 6, I describe a software package that is potentially useful for analyzing microbial phenotypes using my developed analytical pipeline. Finally, in Chapter 7 I discuss the overall advantages, pitfalls and possible future directions for using phenotypic profiling to predict gene function.

CHAPTER 2. INSIGHTS FROM THE REANALYSIS OF HIGH-THROUGHPUT CHEMICAL GENOMICS DATA FOR ESCHERICHIA COLI K-12[1]

ABSTRACT

Despite the demonstrated success of genome-wide genetic screens and chemical genomics studies at predicting functions for genes of unknown function or predicting new functions for well-characterized genes, their potential to provide insights into gene function hasn't been fully explored. We systematically reanalyzed a published high-throughput phenotypic dataset for the model Gram-negative bacterium *Escherichia coli* K-12. The availability of high-quality annotation sets allowed us to compare the power of different metrics for measuring phenotypic profile similarity to correctly infer gene function. We conclude that there is no single best method; the three metrics tested gave comparable results for most gene pairs. We also assessed how converting quantitative phenotypes to discrete, qualitative phenotypes affected the association between phenotype and function. Our results indicate that this approach may allow phenotypic data from different studies to be combined to produce a larger dataset that may reveal functional connections between genes not detected in individual studies.

[1] *This is an open access article reprinted from “Insights from the reanalysis of high-throughput chemical genomics data for Escherichia coli K-12” by Wu, P. I-F., Ross, C., Siegele, D.A. and Hu, J.C. G3: Genes|Genomes|Genetics, Volume 11, Issue 1, Pages 1-13 under the terms of the Creative Commons CC BY license.

INTRODUCTION

Genome-wide genetic screens and chemical genomic studies, pioneered in yeast (Giaever & Nislow, 2014), are now widely used to study gene function in many model organisms, including the bacterium *Escherichia coli* (Campos et al., 2018; Nichols et al., 2011; Price et al., 2018). Based on the same principle that underlies the interpretation of forward genetic studies — that mutations that cause similar phenotypes are likely to affect the same biological process(es) — these high-throughput approaches have led to insights into the biology of a variety of organisms (Arnoldo et al., 2014; Hillenmeyer et al., 2010; Shefchek et al., 2020). It has been concluded that the collective phenotypic expression pattern of an organism can serve as a key to understand growth, fitness, development, and diseases (Bochner, 2009; Houle et al., 2010).

Despite the demonstrated success of high-throughput phenotypic studies at predicting functions for genes of unknown function or predicting new functions for well-characterized genes, their potential to provide insights into gene function hasn't been fully explored. There does not seem to have been a systematic comparison of different metrics for measuring the similarity of phenotypic profiles. Further, while the likely benefits of combining information from high throughput phenotypic studies from different laboratories have been recognized, very few methods of doing this have been described (Hoehndorf et al., 2013; Shefchek et al., 2020).

Here, we report reanalysis of the data from a published high-throughput phenotypic study of *Escherichia coli* K-12 (Nichols et al., 2011). *E. coli* is one of the best-studied bacterial organisms, and the availability of high-quality, abundant annotation sets with information on gene function and regulation allowed us to compare the ability of different metrics for measuring phenotypic profile similarity to correctly infer gene function. We conclude that there is no single best method for comparing phenotypic profiles. Overall, the three metrics we tested gave comparable results for most gene pairs. However, there were instances where the metrics behaved differently from one another. We also assessed how converting quantitative phenotypes to discrete, qualitative phenotypes affected associations between phenotype and function. Our results indicate that this may be a viable approach for combining phenotypic data from different studies, creating a larger dataset that may reveal functional associations not detected by individual studies alone.

MATERIALS & METHODS

Sources of data

The high-throughput phenotypic profiling data as normalized fitness scores were downloaded from supplemental Table S2 of the original paper (Nichols et al., 2011). Missing values (0.17% of total fitness scores) were replaced with population mean as an imputation method. In Table S2, fitness scores were associated with the relevant mutant gene with ECK identifiers. In order to map functional annotations to these genes, the ECK identifiers were verified, corrected, and mapped to b numbers and EcoCyc gene identifiers using information in the genes.dat file from EcoCyc version 21.0. This and other EcoCyc files were downloaded from their website (https://biocyc.org/download.shtml).

The six annotation sets were obtained from various sources. EcoCyc pathway annotations were mapped to each gene using information in the pathways.col file (EcoCyc version 21.0). EcoCyc protein complex annotations were mapped to each gene using information in the protcplxs.col file (EcoCyc version 21.1) after removal of homomeric protein complexes. KEGG module annotations were obtained and mapped by retrieving module name and b numbers from the KEGG website (https://www.kegg.jp). Operon and regulon annotations were obtained and mapped to each gene using a download of Regulon DB version 9.4 (http://regulondb.ccg.unam.mx). The operon.txt file was the source of operon annotations. The object_synonym.txt file was used to map ECK12 gene identifiers to ECK gene identifiers. RegulonDB annotations were then obtained from the file regulon_d_tmp.txt and mapped to ECK identifiers. GO biological process annotations were obtained from the Ecocyc gene_association.ecocyc file (EcoCyc 21.1) and mapped to each gene to produce the file 2017_05_ECgene_association.ecocyc.csv. UniProt IDs retrieved from the Bioconductor package UniProt.ws were used to associate GO annotations from proteins to genes. The annotation sets, the number of genes annotated by each annotation set, and the total number of annotations are summarized in Table 1.

Statistical analysis and software

The statistical programming language R was used throughout the study. Phenotypic profile similarity was calculated using Pearson Correlation Coefficient (|PCC|), Spearman's Rank Correlation Coefficient (|SRCC|), Mutual Information (MI), and semantic similarity. Pearson and Spearman's Rank Correlation Coefficients were calculated using the cor() function, with the metric argument specified by either "pearson" or "spearman". Different implementations are needed to calculate Mutual Information for continuous, quantitative data and discretized, qualitative data. Mutual Information for quantitative data was calculated using the cminjk() function provided in the mpmi package, while Mutual Information for discretized data was calculated using the mutinformation() function provided in the infotheo package. Both packages are available from CRAN (https://cran.r-project.org/web/packages/mpmi/index.html). For the plots of precision versus ranking based on phenotypic profile similarity (Fig.2, 3, 4, and 6), the negative control is precision calculated for randomly-ordered gene pairs that were generated using the R function sample() to permute the rankings of all possible gene pairs. For precision-recall curves (Figures S5, S6, and S7), the negative control is precision calculated for 5,000 gene pairs selected randomly without replacement from the set of all possible gene pairs using the R function sample(). For all negative controls, the number of co-annotated gene pairs present in the set of all possible gene pairs differed depending on which annotation set or combination of annotation sets was used to identify co-annotated gene pairs, except Figure 2, where only the negative control using the union of annotation sets 1 through 5 is shown.

The semantic similarity of GO biological process annotations was calculated using a graph-based method (Wang et al., 2007). Calculations were performed using the GOSemSim package (Yu et al., 2010) from Bioconductor. For the Mann-Whitney U test, wilcox.test() function was used.

For violin plots, geom_violin() was used to plot the kernel density plot and geom_box() was used for the boxplot. Both functions are from the ggplot2 package (Wickham, 2016). In the box plot associated with each violin plot, the middle line in the box represents the median; the whiskers indicate the 1.5 interquartile range (IQR) away from either Q1 (lower box boundary) or Q3 (upper box boundary). For the violin plots that display the distribution of MI values for gene pair profile similarity determined using discretized, ternary fitness scores (Figures 7A and 7B), the MI values were log transformed after addition of a constant (1×10^{-6}) to eliminate zero values.

For each pathway and protein complex in Figures S1 and S2, a permutation-based p-value was calculated by randomly sampling the same number of phenotypic profiles as the number of genes contained in each pathway or protein complex, calculating the mean pairwise profile similarity based on |PCC|, repeating 1,000 times, and then calculating the fraction of these mean |PCC| values that has a higher mean |PCC| than the actual |PCC| value for that pathway or protein complex.

Data Availability Statement

The code and data files used for calculations and reproducing the results are available on GitHub: https://github.com/peterwu19881230/Systematic-analyses-ecoli-phenotypes. Supplemental material (Tables S1 and S2 and Figures S1-S9) can be downloaded from https://gsajournals.figshare.com/.

RESULTS

Phenotypic profiles and the functional annotation sets used

We start with descriptions of the phenotype data and functional annotation sets that were used for our analysis. The phenotypic profiles come from a high-throughput chemical genomics study of *E. coli* K-12 (Nichols et al., 2011). Growth phenotypes for 3,979 mutant strains, which were primarily single-gene deletions of non-essential genes, were based on sizes of spot colonies grown under 324 conditions, which represented 114 unique stresses. For each of the growth conditions, fitness scores were obtained and scaled to a standard normal distribution. Positive scores indicate increased fitness and negative scores indicate decreased fitness.

Six annotation sets were used as sources of information about gene function. The number of genes annotated in each annotation set and the total number of annotations for each annotation set are shown in Table 1. Annotations of *E. coli* genes to metabolic and signaling transduction pathways (annotation set 1) and to heteromeric protein complexes (annotation set 2) were obtained from EcoCyc (Keseler et al., 2017); annotation of genes to operons (annotation set 3) and to regulons (annotation set 4) were extracted from EcoCyc and RegulonDB (Gama-Castro et al., 2016); and annotations of genes to KEGG modules (annotation set 5), which associate genes to metabolic pathways, molecular complexes, and also to phenotypic groups, such as pathogenesis or drug resistance, were obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016). For these five annotation sets, genes were scored as co-annotated if they shared the same annotation(s) from one or more of the annotation sets, for example, being annotated to the same pathway or protein complex, etc.

The annotations of *E. coli* genes with Gene Ontology (GO) biological process terms (annotation set 6) (The Gene Ontology Consortium, 2017) were obtained from EcoCyc. The GO biological process annotations of *E. coli* genes were treated separately from the other five annotation sets because GO's directed-acyclic graph structure allows semantic similarity rather than co-annotation to be used for assessing functional similarity (Pesquita, 2017). Simply looking for co-annotations with the same GO term(s) will include co-annotations to high-level terms, such as ‘GO:0044237 cellular metabolic process’ or ‘GO:0051716 cellular response to stimulus’, terms that don’t provide very specific information about function. Also, co-annotations won't capture instances where two genes are annotated with related, but not identical, terms. These limitations can be overcome by using semantic similarity rather than co-annotation to estimate functional similarity from GO annotations. The method for determining the semantic similarity of two GO terms developed by Wang *et al.* (Wang et al., 2007) takes into account the locations of the terms in the GO graph, as well as incorporating the different semantic contributions that a shared ancestral term may make to the two terms, based on the logical relationship, such as 'is_a' or 'part_of', that connect the term to the shared ancestor. In addition, when calculating functional similarity, the Wang method includes both identical GO terms and semantically similar GO terms associated with the two genes being compared.

Functional connections between genes enriched for higher phenotypic profile similarity

The association between phenotypic profiles and functional annotations was examined from two perspectives: First, are gene pairs that share the same annotation(s), i.e. co-annotated gene pairs, more likely to have higher phenotypic profile similarity? Second, are gene pairs with higher phenotypic profile similarity more likely to be co-annotated?

To address whether co-annotated gene pairs have higher phenotypic profile similarity, we used Pearson Correlation Coefficient (PCC) to assess the phenotypic profile similarity. This metric was chosen because it is probably the most widely used metric to assess phenotypic profile similarity and was the metric used in the original paper for comparing phenotypic profiles (Nichols et al., 2011). To visualize the results, the distributions of the absolute value of PCC (|PCC|) for gene pairs were plotted as violin plots for various combinations of annotation sets (Figure 2.1). The first violin plot shows the distribution of |PCC| values for all possible gene pairs (mean |PCC| = 0.09). The majority have a |PCC| value <0.25 and only 0.16% have a |PCC| value >0.75 (an arbitrarily chosen cut-off based on Hinkle *et al.* (Hinkle et al., 2002). When only gene pairs that are co-annotated to the same EcoCyc pathway were considered (second violin plot), there was a statistically significant increase in the mean |PCC| value (0.16), and the percentage of gene pairs with |PCC| >0.75 increased twenty-fold. Similar results were seen for gene pairs that are co-annotated to the same heteromeric protein complex (third violin plot, mean |PCC| = 0.22). When considering only gene pairs that are co-annotated to more than one annotation set (fourth and fifth violin plots), even higher phenotypic profile similarity was observed (mean |PCC| = 0.39, 0.54, respectively), supporting the expectation that gene pairs with stronger functional associations will have more similar phenotypic profiles. The trend of there being a higher fraction of gene pairs with |PCC| >0.75 as functional associations increased also continued; this fraction increased from 0.16% for all gene pairs, to 3.2% for gene pairs in the same EcoCyc pathways, to 4.9% for gene pairs in the same heteromeric protein complexes, to 19% for gene pairs in the same EcoCyc pathways and heteromeric protein complexes, and to 30% for gene pairs that are co-annotated in annotation sets 1 through 5 (the union of EcoCyc pathways, heteromeric protein complexes, operons, regulons and KEGG modules).

A more detailed analysis within the EcoCyc pathway or heteromeric protein complex annotations was conducted by examining all pairwise combinations of gene pairs within pathways or protein complexes that contain two or more gene products. Supplemental Figures S1 and S2 show the distributions of |PCC| values for all pairwise combinations of genes in each pathway or protein complex. For 70% of the pathways and 67% of the protein complexes analyzed the average |PCC| value is significantly higher than random expectation (|PCC| = 0.09).

Phenotypic profile similarity is explained by functional annotations

To address the second question, which is whether gene pairs with higher phenotypic profile similarity are more likely to be co-annotated, we ranked gene pairs based on phenotypic profile similarity and then calculated precision based on whether or not gene pairs are co-annotated (Figure 2.2). Precision is the fraction of results that a test identifies as positive that represent true positives. Mathematically, precision, also known as the positive predictive value, is the number of True Positives divided by True Positives plus False Positives, or TP/(TP+FP). After ranking gene pairs based on phenotypic profile similarity expressed as |PCC| values, precision for each position *n* in the ranking was calculated considering gene pairs ranked at or above position *n* to be TPs if they are co-annotated or FPs if they are not co-annotated. For example, for the 100th gene pair in the ranking, precision is calculated for gene pairs 1 through 100. Figure 2 shows the plots of precision versus ranking for the top-ranking 500 gene pairs computed for single annotation sets or combinations of annotation sets. For gene pairs co-annotated to the same pathway(s), precision started at zero, because the highest ranked gene pair was not co-annotated, but then increased to ~0.8 before gradually declining and leveling off at approximately 0.2.

Surprisingly, for gene pairs co-annotated to the same protein complex, precision was very low and not significantly different from the precision values computed for randomly ordered gene pairs. Combining the annotation sets for pathways and protein complexes, brought a slight increase in precision. When operon, regulon, and KEGG modules were also included to define the broadest set of co-annotations, precision increased dramatically.

The Pearson Correlation Coefficient is sensitive to the extreme fitness scores on minimal media

To try to understand why precision was so low for protein complex annotations (Figure 2.2), we inspected the gene pairs and saw that 98 of the 100 top-ranking gene pairs consisted of genes coding for biosynthetic enzymes, and, in 84 of these 98 gene pairs, the genes were annotated to different biosynthetic pathways. For example, the top-ranked gene pair (|PCC| = 0.96) contained the genes *ilvC* and *argB*, which encode enzymes required for isoleucine-valine and arginine biosynthesis, respectively. Mutant strains lacking any of these biosynthetic genes would be auxotrophs and share the phenotype of little or no growth on unsupplemented minimal media. To test whether the |PCC|-based measure of phenotypic profile similarity was dominated by the large negative fitness scores associated with growth of auxotrophic mutants on minimal media, we excluded the fitness scores for the growth conditions that involved minimal media (10 out of 324 total conditions) and reassessed the relationship between precision and phenotypic profile similarity. As shown in Figure 2.3, even though only a small fraction of conditions was excluded, this change resulted in dramatically higher precision overall, not only for gene-pairs co-annotated to heteromeric protein complexes but also for gene-pairs co-annotated to either

EcoCyc pathways, the union of EcoCyc pathways and heteromeric protein complexes, or the union of annotation sets 1 through 5. In addition, when strains known to have auxotrophic phenotypes were excluded from the analysis, little difference in precision was seen whether growth conditions involving minimal media were included or excluded (Figure 2.S3).

Alternative metrics for measuring phenotypic profile similarity

There are other methods, besides the Pearson Correlation Coefficient, that can be used to assess phenotypic profile similarity. We chose the absolute value of Spearman's Rank Correlation Coefficient (|SRCC|) or mutual information (MI), which were implemented as described in the methods, to measure similarity, and used the union of annotation sets 1 through 5 to score co-annotation. Violin plots of the distributions of phenotypic profile similarity obtained using these alternative metrics were not significantly different from the distributions seen using |PCC| as the metric (results not shown). In contrast, as shown in Figure 2.4A, the correlation between phenotypic profile similarity and precision was dramatically higher for |SRCC| and MI compared to |PCC|. For both |SRCC| and MI, precision was >0.9 for the top 100 ranked gene pairs and remained >0.5 for approximately the top 500 pairs. This result suggests that determining phenotypic profile similarity using Spearman’s Rank Correlation Coefficient or Mutual Information is less sensitive to the presence of a relatively small number of extreme phenotype scores than using the Pearson Correlation Coefficient, at least for this phenotypic dataset. If we recalculate precision for all three metrics after excluding the 10 growth conditions where auxotrophic mutants don’t grow, we see very little change in precision for gene-pairs ranked based on |SRCC| or |MI| (compare Figures 2.4A and 2.4B). There is now very little difference in precision for the three metrics (Figure 2.4B). In addition, we calculated precision after removing the strains known to have an auxotrophic phenotype (Figure 2.S4). This result is consistent with Figure 2.4B in that all three metrics have similar precision.

Simplified phenotypic profiles preserve biological meanings

Combining phenotypic information from different studies is expected to increase the likelihood of finding associations between genes and functions. However, the ability to combine datasets can be limited by differences in how quantitative phenotypes are scored in different studies. In addition, there is a need for methods to incorporate qualitative phenotypes, such as changes in cell or colony morphology, which are inherently qualitative, as well as changes in phenotype that are reported in a qualitative way, such as increased or decreased growth rate or increased or decreased resistance to a chemical. To address both of these issues we took the approach of converting quantitative phenotypes to qualitative phenotypes. We chose this approach because, if successful, it would allow a larger number of datasets to be combined. It would also allow us to utilize microbial phenotype information that has been collected and annotated with qualitative phenotype ontology terms in databases such as PomBase (Harris et al., 2013), SGD (Cherry et al., 2012), and OMP (Chibucos et al., 2014).

The quantitative fitness scores in the phenotypic dataset were discretized to create a qualitative dataset with the fitness scores converted to 1, 0, or -1, where 1 stands for increased fitness, -1 for decreased fitness, and 0 for no difference in fitness compared to the mean fitness for all strains in a particular growth condition. The |PCC| cutoffs used to separate the quantitative fitness scores into discretized, ternary bins were based on the 5% false discovery rate (FDR) for each growth condition, which was the cutoff used to identify significant phenotypes in the original study (Nichols et al., 2011). Because the majority of strains have no significant phenotype in the growth conditions used (Nichols et al., 2011), after discretizing the data the majority of strains will have fitness scores of 0. Therefore, the Pearson Correlation Coefficient was no longer suitable for measuring phenotypic profile similarity. Instead, mutual information (MI) (Priness et al., 2007) was used as the scoring metric. The distribution of MI values for gene pairs were plotted as violin plots, after addition of a constant (1x10^{-6}) to eliminate zero values followed by log transformation of the data. The first violin plot in Figure 5A shows the distribution of MI values for all possible gene pairs, followed by, from left to right, the distribution of MI values for gene pairs co-annotated to either the same EcoCyc pathway; the same heteromeric protein complex; to both an EcoCyc pathway and a heteromeric protein complex; or are co-annotated to the same EcoCyc pathway, heteromeric protein complex, operon, regulon, and KEGG module. As was seen for the mean |PCC| values in the analysis of the quantitative data (Figure 2.1), the mean MI values increased as the functional associations for a given gene pair increased (Figure 2.5A).

Another complication that can arise when trying to combine phenotype information from different studies is variation in the conditions used. For example, different studies may look at the effects of the same chemical but use different concentrations. To determine how removing concentration information affects phenotypic profile similarity, we reduced the original 324 growth conditions to 114 unique stresses. When different concentrations of a chemical were tested, for each strain only the concentration with the most significant fitness score was included and assigned a value of 1 or -1, as appropriate, or a score of 0 if no significant phenotype was seen for that treatment. The violin plots in Figure 2.5B show the distribution of MI values (after log transformation as described above) for all gene pairs and for different annotation sets or combinations of annotation sets for the reduced set of conditions. As seen for the full qualitative dataset, the mean MI values for co-annotated gene pairs in the reduced dataset were significantly higher than the mean MI value for all possible gene pairs (Figure 2.5B). In addition, when the distributions of gene-pairs in the same co-annotation group are compared between Figures 2.5A and 2.5B, significant differences of the means were observed for every co-annotated group (p-value <0.001). Overall, these results indicate that useful inferences about gene function can still be made after the conversion of quantitative phenotypes to qualitative phenotypes and even after collapsing the number of phenotypes for each chemical treatment.

We expected loss of information after converting quantitative phenotype scores to discretized, ternary fitness scores. To compare how many functional associations could still be retrieved using the qualitative scores, gene pairs were sorted based on their MI values determined using either quantitative phenotype scores, the qualitative ternary fitness scores, or the qualitative ternary fitness scores for the reduced set of conditions. Precision was then calculated, as described earlier, and was plotted versus ranking. As can be seen in Figure 2.6, precision is comparable for the top 100 gene pairs for both quantitative and for discretized, qualitative fitness scores. After this point, precision drops more quickly for the qualitative data than for the quantitative data. When precision for the reduced set of conditions is compared to precision for either of the other datasets, we see that precision drops off sooner and decreases more rapidly. Yet, precision is still much higher than for randomly ordered gene pairs, which indicates that functional associations can still be identified when qualitative, discretized fitness scores are used.

Semantic similarity of GO annotations increased for gene pairs with shared functional annotations and with higher phenotypic profile similarity

Another way to assess whether two genes are likely to have similar functions is to compare the semantic similarity of the GO terms annotated to each gene. In the dataset from Nichols *et al*., 66% (2,609 out of 3,979) of the strains used have mutations of genes that are annotated with GO biological process terms, which seemed a sufficient number to justify using this approach. Semantic similarity was computed using the method described by Wang *et al. (Wang et al., 2007)*, and the distribution of semantic similarity scores for all gene pairs where both members of the pair are annotated with at least one GO biological process term was compared to the distributions for subsets of gene pairs that have similar functions based on being co-annotated in one or more of the non-GO annotation sets. As shown in Figure 2.7A, semantic similarity increased when only co-annotated gene pairs were considered. The mean pairwise semantic similarity increased from 0.22 for all genes with GO biological process annotations (first violin plot), to 0.54 for gene pairs co-annotated to the same EcoCyc pathway (second violin plot), and to 0.80 for gene pairs co-annotated to the same heteromeric protein complex (third violin plot). Mean profile similarity was even higher for gene pairs that are co-annotated to both pathways and heteromeric protein complexes (mean=0.90) as well as for gene pairs that are co-annotated in annotation sets 1 through 5 (mean=0.89), as shown in the fourth and fifth violin plots, respectively. These results show that co-annotated gene pairs are also enriched for functional similarity based on GO biological process annotations.

To test whether gene pairs that have higher phenotypic profile similarity are more likely to have similar functions based on GO biological process annotations, we compared the distributions of semantic similarity values for all gene pairs annotated with GO biological process terms and for subsets of these gene pairs that have high phenotypic profile similarity based on |PCC| or MI. A cutoff of |PCC| >0.75 for the second violin plot was chosen arbitrarily to represent a moderate to high correlation (Hinkle et al., 2002), while the cutoffs of MI>0.15 and >0.32 for the third and fourth violin plots, respectively, were chosen so that all three subsets of gene pairs would contain the same number (~1,200) of gene pairs. Comparison of the first two violin plots in Figure 2.7B shows that semantic similarity increased significantly for gene pairs with |PCC| >0.75 (mean semantic similarity=0.61) compared to all gene pairs with GO biological process annotations (mean=0.22). Enrichment for higher semantic similarity was also seen when phenotypic profile similarity was determined using discretized, ternary fitness scores either for all growth conditions (third violin plot, MI>0.15, mean=0.59) or for the collapsed set of 114 growth conditions (fourth violin plot, MI>0.32, mean=0.58). These results are consistent with those in Figure 2.1, which show higher phenotypic profile similarity for co-annotated gene pairs.

DISCUSSION

We systematically reanalyzed a published high-throughput phenotypic profile dataset for the model Gram-negative bacterium *E. coli* comparing different metrics for measuring phenotypic profile similarity, and assessing the effect of converting quantitative fitness scores to qualitative fitness on measurements of phenotypic profile similarity. We re-examined the *E. coli* phenotypic profiles in a pairwise fashion with the help of existing functional annotations. Overall, we found that gene pairs with functional associations are enriched for phenotypic profile similarity and that gene pairs with high phenotypic similarity scores tend to have functional associations.

Six high-quality annotations sets were used as sources of functional information. The gene annotations in EcoCyc, RegulonDB, KEGG, and GO come primarily from expert manual curation (Gama-Castro et al., 2016; Kanehisa et al., 2016; Keseler et al., 2017; Keseler et al., 2014; The Gene Ontology Consortium, 2017). The GO biological process annotations include ~1,200 annotations (21%) that are inferred from electronic annotation without additional human review. We decided to include the electronic annotations in our analysis because most of them come from the transfer of annotations from orthologous gene products or are based on mappings from external sources, such as InterPro2GO or EC2GO, which have been shown to be very accurate (Camon et al., 2005; Hill et al., 2001; Holliday et al., 2017). Indeed, no significant difference was found in the semantic similarity of gene pairs whether electronic annotations were included (Figure 2.7B) or excluded (Figure 2.S5).

One aim of this study was to determine whether different metrics for determining phenotypic profile similarity differed in their ability to identify gene pairs with functional similarity. Comparison of the profile similarity scores for the top-ranked gene-pairs showed that the three metrics used, |PCC|, |SRCC|, and MI, produced comparable results for most, although not all, gene pairs (data not shown). A more quantitative way to compare the performance of the metrics is by introducing precision: the fraction of positive results that are true positives. Gene pairs with phenotypic profile similarity above a specified cutoff were considered as positive results, and true positives were defined as gene pairs that are co-annotated in at least one of annotation sets 1 through 5. We chose to use precision rather than accuracy, which is the fraction of correct results, because the co-annotated and non-co-annotated gene pairs constitute a highly imbalanced dataset (Saito & Rehmsmeier, 2015). Since the number of non-co-annotated gene pairs is much larger than the number of co-annotated gene pairs, high accuracy could be achieved by classifying all gene pairs as true negatives without being informative.

We chose to plot precision versus ranked gene pairs because when the data are graphed in this way, precision represents the fraction of gene pairs whose profile similarity is above a specified cutoff value that have already been co-annotated. This presentation seemed the most useful for choosing for future study non-co-annotated gene pairs that are likely to have a functional association. We also plotted the data in a more standard way as precision-recall curves. Recall, also known as sensitivity, is the fraction of real positives that a test identifies. It is equal to TP/(TP+FN), where True Positives + False Negatives is the number of real positives. We scored as True Positives gene pairs that are co-annotated in one or more annotation sets and whose profile similarity was above a specified cut-off value. Co-annotated gene pairs whose profile similarity was below the specified cutoff were scored as False Negatives. Precision and recall were calculated for the 5,000 top-ranked gene pairs for each similarity metric. This cutoff was chosen because the low correlation values seen for gene pairs below the top 5,000 are expected to be less useful in identifying functional associations. Figure S7 shows precision-recall curves for gene pairs ranked based on either |PCC|, |SRCC|, or MI after minimal media conditions were excluded. This corresponds to the precision versus ranking graphs presented in Figure 2.4B. Both representations of the data show that highly correlated gene pairs were enriched for functional associations.

Precision-recall curves were also made that correspond to the precision versus ranking graphs shown in Figures 2.3 and 2.6. These are Figures 2.S6 and 2.S8, respectively. The conclusions from these precision-recall curves are consistent with the conclusions made from the graphs of precision versus ranking.

Based on the precision scores for the top 500 ranked gene pairs, it initially appeared that |SRCC| and MI outperformed |PCC| (Figure 2.4A). However, when phenotypic profile similarity was recalculated after removing conditions involving growth on minimal media, the precision for gene pairs ranked based on |PCC| increased significantly, and there was now little difference in the performance of |PCC|, |SRCC| or MI (compare Figures 4A and 4B). We suggest that this observed increase in precision for gene pairs ranked by |PCC| might be due to the sensitivity of the Pearson Correlation Coefficient to outliers in the data (Schober, 2018). We realized that the collection of strains used by Nichols *et al.* contains many mutants that have little or no growth on minimal media because the gene for a biosynthetic enzyme is deleted. Precision was low when minimal media growth conditions were included because so many combinations of genes from different biosynthetic pathways shared large, negative fitness scores on the 10 conditions involving minimal media but did not share a functional annotation in the annotation sets used. In general, the auxotrophic mutants didn't have a significant phenotype in most of the other 314 growth conditions tested, which used rich media, so the large negative fitness scores on minimal media were essentially outliers. When these outliers were excluded, precision increased for the gene-pairs ranked based on |PCC|. We suggest that when high-throughput phenotype studies include conditions that involve defined media, such as testing for utilization of carbon or nitrogen sources, it would be useful to supplement the base minimal media with amino acids, nucleosides, and enzyme co-factors to reduce the phenotypic clustering of mutant strains unable to synthesize these compounds.

The results presented in Figures 2.4A and 2.4B show that when gene-pairs are ranked by similarity calculated using the metrics |SRCC| or |MI|, precision didn't change very much when conditions involving minimal media were excluded. While this observation might indicate that |SRCC| or MI are more useful for determining phenotypic profile similarity in high-throughput studies, we think it is premature to draw this conclusion based on analysis of only one phenotypic dataset. Moreover, for gene pairs ranked by |PCC|, many of the gene pairs that were excluded by eliminating the minimal media growth conditions would have been recognized as true positives if the annotation sets included annotations to cellular processes such as amino acid biosynthesis or nucleotide biosynthesis in addition to the annotations to metabolic pathways for individual compounds.

We conclude that there is no single best way to measure phenotypic profile similarity, and suggest it may be advantageous to use more than one correlation metric to look for functional associations. When we compared the 10,000 top-ranked gene pairs identified using either |PCC| or |SRCC| with minimal media conditions excluded, we found that each metric identified gene pairs not identified by the other. There were 204 gene pairs with |PCC| ≥ 0.5000 that weren't present among the top 10,000 gene pairs ranked based on Spearman ranked correlation, and 87 gene pairs with |SRCC| ≥ 0.5000 that weren't present among the top 10,000 gene pairs ranked based on Pearson correlation.

We also found differences among the highly ranked gene pairs when we compared gene pairs ranked by |PCC| when minimal media growth conditions were included or excluded. For most gene pairs that didn't include an auxotrophic mutant, the phenotypic profile similarity based on |PCC| changed very little when minimal media conditions were removed (data not shown). However, there were a few gene pairs where a possible functional association could have been missed if the minimal media conditions were not removed. We illustrate this with a gene pair where the functions of the gene products are known to have a functional association. The *exbD* and *fepA* genes are both needed for transport of ferric iron-enterobactin across the outer membrane (Noinaj et al., 2010). When profile similarity was calculated using the fitness scores for all conditions, |PCC| = 0.4773. After minimal media conditions were removed, |PCC| increased to 0.6204, a high enough correlation that this gene pair would be a reasonable candidate for future experiments to test the prediction.

To make it easier to compare results for the different similarity metrics, we have made the dataset from Nichols *et al.* available in a searchable, interactive format that allows queries for strains, conditions, and phenotypic profile similarity of gene pairs determined by |PCC| with all conditions, |PCC| with minimal media conditions excluded, |SRCC|, MI, and semantic similarity (https://microbialphenotypes.org/wiki/index.php?title=Special:Ecolispecialpage).

The relationship between precision and ranking based on profile similarity shown in Figure 2.4B suggests that a shared function is known for most of the highly correlated gene pairs. To test this idea, we used a cutoff of |PCC| >0.75 to define highly correlated gene pairs and then manually examined the non-co-annotated gene pairs. If fitness scores for the growth conditions involving minimal media were excluded, there were only 10 non-co-annotated gene pairs (summarized in Table 2). We found functional associations that could explain the observed phenotypic profile similarity for 7 of the 10 gene pairs. In one case, the two genes (*dsbB* and *dsbA*) showed up as non-co-annotated because they are in a pathway that wasn’t yet included in EcoCyc version 21.0. The other six gene pairs highlight some of the challenges of creating (and using) annotation, such as deciding where pathways start and end and determining appropriate levels of granularity. For example, the gene pairs *rfaF*(*waaF*)-*rfaE*(*hldE*) and *rfaF*(*waaF*)-*lpcA* (*gmhA*) are non-co-annotated, even though all three genes are required for synthesis of the lipid A-core oligosaccharide component of outer membrane lipopolysaccharide. The explanation is that *rfaF*(*waaF*) is annotated to the central assembly pathway for building the lipid-core oligosaccharide moiety, while *rfaE*(*hldE*) and *lpcA*(*gmhA*) are annotated to a branch pathway that builds one of the saccharide subunits of the core (Raetz & Whitfield, 2002). The functional association between the three genes would have been revealed if we had included GO annotations, since all three genes are annotated to the GO term for the lipopolysaccharide core region biosynthetic process (GO:0009244).

We did not find a shared function for the last three non-co-annotated gene pairs. Given that so many of the other highly correlated gene pairs do share a function, it is possible that future experiments will uncover a shared function for these three gene pairs. However, it also possible that the observed phenotypic profile similarity is fortuitous, as we saw for mutants with an auxotrophic phenotype or mutants with increased sensitivity to DNA damage. For example, this may be the most likely explanation for the phenotypic similarity of the *mnmE* and *apaH* genes. Both are required for growth at pH 4.5 (Nichols et al., 2011; Vivijs et al., 2016), but appear to function independently. MnmE, partnered with MnmG, modifies 2-thiouridine residues in the wobble position of tRNA anticodons (Elseviers et al., 1984), while ApaH is a diadenosine tetraphosphatase (Guranowski et al., 1983) and mRNA decapping enzyme (Luciano et al., 2019). Both MnmE and ApaH are proposed to affect resistance to pH and other stresses through their effects on gene expression (Dedon & Begley, 2014; Luciano et al., 2019; Vivijs et al., 2016).

A significant conclusion from this study is that functional associations can still be inferred from phenotypic profiles after quantitative fitness scores are converted to discretized, ternary scores. While some information was lost compared to using the original quantitative fitness scores, the precision based on the ternary fitness scores was much greater than for randomly ordered gene pairs (Figure 2.6). This result suggests that discretized, ternary scores could be used to combine quantitative phenotype information from different studies. Using discretized scores might also allow qualitative phenotype information, such as aspects of cell morphology, to be incorporated into phenotypic profiles along with discretized quantitative phenotype information. This approach would also allow information from phenotype annotations, available from databases such as PomBase, SGD, or Microbial Phenotypes Wiki, to be incorporated into phenotypic profiles. The phenotype annotations typically capture information in a discretized fashion and have previously been shown to be useful for inferring gene function (Ascensao et al., 2014; Hoehndorf et al., 2013).

The precision of the discretized data could be increased by partitioning the quantitative scores into a larger number of bins, as shown in Figure 2.S9. Precision increased incrementally as the number of bins was increased from 3 to 5 bins, from 5 to 7 bins and from 7 to 9 bins. However, because the results from many phenotypic studies are not amenable to being partitioned into a larger number of bins, we believe that using ternary scores will maximize the number of datasets that can be combined and allow more inferences about gene function to be made from phenotypic information.

ACKNOWLEDGEMENTS:

We thank Michelle Giglio, Matthew Sachs, Yen-Ting Lu and the reviewers for helpful comments on the manuscript. This work was supported by a grant from the National Institutes of Health (R01GM089636) to JCH.

AUTHOR CONTRIBUTIONS:

JH and DS conceptualized the project. JH, PW, and DS designed the experiments and the analytical pipeline. PW implemented the experiments and analyzed the data. CR helped with the implementation of experiments. PW, DS, and JH wrote the manuscript.

COMPETING INTERESTS:

The authors declare no competing interests.

Table 2.1. sources of the gene annotations used in this study.

Annotation set (source)	**Number of annotated genes**[a]	**Total number of gene annotations**[b]
1) EcoCyc pathways (EcoCyc)	885	2,317
2) Heteromeric protein complexes (EcoCyc)[c]	688	871
3) Operons (RegulonDB)	3,858	5,349
4) Regulons (RegulonDB)	1,572	3,886
5) Modules (KEGG)	333	524
6) GO biological process annotations	2,609	5,775
7) Annotation to both EcoCyc pathways and heteromeric	188	818[d]

protein complexes (intersection of annotation sets 1 and 2)		
8) Annotation in each of annotation sets 1 through 5 (intersection of annotation sets 1 through 5)	77	922[e]
9) Annotation to either EcoCyc pathways or heteromeric protein complexes (union of annotation sets 1 and 2)	1,385	3,269
10) Annotation in any of annotation sets 1-5 (union of annotation sets 1 through 5)	3,866	12,937

[a] Number of annotated genes that were deleted or otherwise mutated in the set of strains used in the original study (Nichols et al., 2011).

[b] Total number of annotations associated with the genes in the first column.

[c] We have excluded genes annotated to EcoCyc protein complexes that are homomeric complexes.

[d] This is the number of annotations associated with any of the 188 genes that are annotated to both annotation sets.

[e] This is the number of annotations associated with any of the 77 genes that are annotated in each of annotation sets 1 through 5.

Table 2.2. non-co-annotated gene pairs with |PCC| >0.75

Gene pair[a]	**Known or predicted functional association**
ECK0730-*pal*_ECK0725-*ybgC*[b]	Tol-Pal cell envelope complex (CPLX0-2201)
ECK0768-*uvrB*_ ECK2563-*recO*	DNA repair: recombinational repair (RECFOR-CPLX) and nucleotide excision repair (UVRABC-CPLX)
ECK1912-*uvrC*_ECK2563-*recO*	DNA repair: recombinational repair (RECFOR-CPLX and nucleotide excision repair (UVRABC-CPLX)
ECK2901-*visC*(*ubiI*)_ECK3033-*yqiC*(*ubiK*)[c]	ubiquinol-8 biosynthesis (PWY-6708)
ECK3610-*rfaF*(*waaF*)_ECK3042-*rfaE*(*hldE*)[d]	superpathway of lipopolysaccharide biosynthesis (LPSSYN-PWY)
ECK3610-*rfaF*(*waaF*)_ECK0223-*lpcA*[d]	super pathway of lipopolysaccharide biosynthesis (LPSSYN-PWY)
ECK3852-*dsbA*_ECK1173-*dsbB*	periplasmic disulfide bond formation (PWY0-1599)[e]
ECK1544-*gnsB*_ECK2394-*gltX*	unknown
ECK2066-*yegK*(*pphC*)_ECK0345-*mhpB*	unknown
ECK3699-*mnmE*_ECK0050-*apaH*	unknown

[a] The strain names are from supplemental Table S2 of (Nichols et al., 2011). Where the gene name has changed, the new gene name is included in parentheses.

[b] *ybgC* is in an operon that also includes the genes for three of the protein components of the Tol-Pal cell envelope complex

[c] *ubiK* codes for an accessory protein required for efficient synthesis of ubiquinol-8 under aerobic conditions, but is not annotated as part the ubiquinol-8 biosynthesis pathway

[d] *rfaE*(*hldE*) and *lpcA* are not annotated to the super pathway of lipopolysaccharide biosynthesis (LPSSYN-PWY)

[e] PWY0-1599 was not present in EcoCyc version 21.0

FIGURES

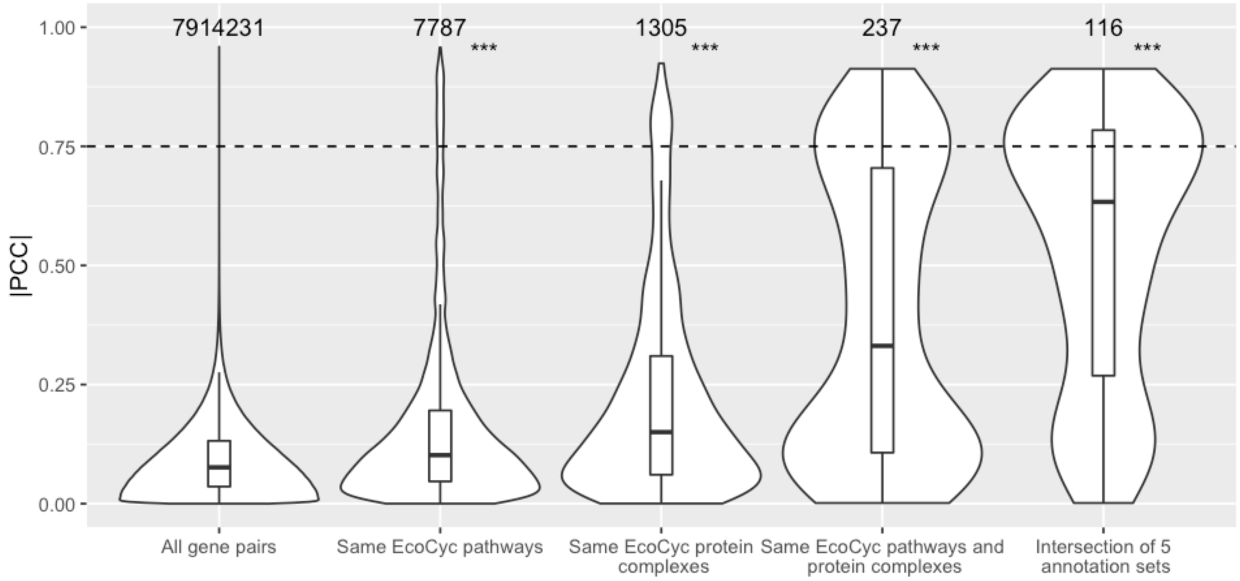
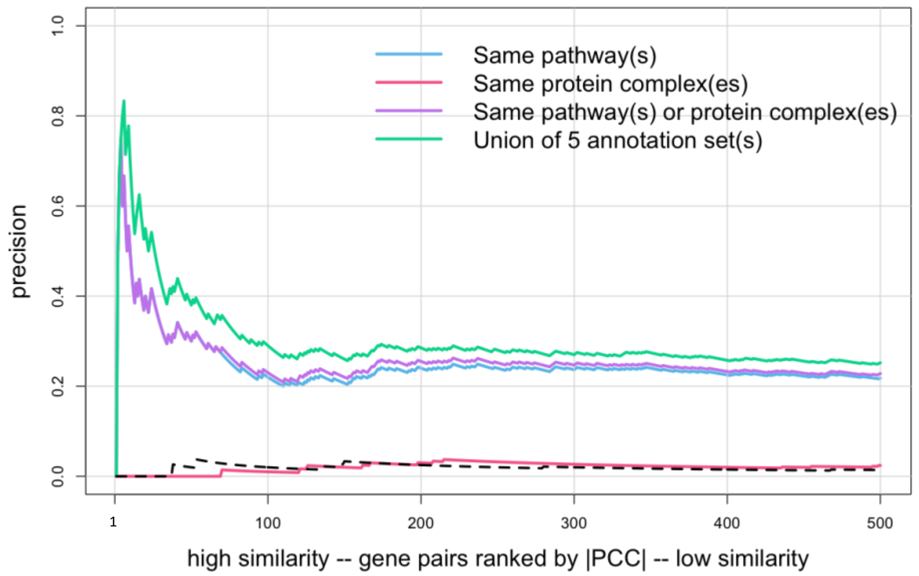


FIGURE 2.1. Higher phenotypic similarity was found for co-annotated gene pairs. Violin plots of the distributions of |PCC| values for, from left to right, all possible gene pairs, gene pairs annotated to the same EcoCyc pathway, gene pairs annotated to the same heteromeric protein complex, gene pairs annotated to the same EcoCyc pathway and heteromeric protein complex, and gene pairs that are co-annotated in annotation sets 1 through 5 (the intersection of EcoCyc pathways, heteromeric protein complexes, operon, regulon, and KEGG module). Numbers above each violin plot indicate the number of gene pairs in each plot. ***: p-value <0.001 was determined by 1-sided Mann-Whitney U test, compared to all gene pairs. The dashed line indicates |PCC| = 0.75, which was chosen as an arbitrary cutoff.



Ranking / Similarity	1st	100th	200th	300th	400th	500th
|PCC|	0.96	0.92	0.90	0.89	0.87	0.86

FIGURE 2.2. Increased co-annotation was found for gene pairs with higher phenotypic profile similarity. Gene pairs were ranked from high to low similarity based on |PCC| values and plotted versus precision, which was calculated as described in the text (only the first 500 gene pairs are shown). The different colored lines indicate either gene pairs that are annotated to the same EcoCyc pathway (blue), to the same heteromeric protein complex (pink), to either the same EcoCyc pathway or protein complex (purple), or are co-annotated in any of annotation sets 1 through 5 (the union of EcoCyc pathways, heteromeric protein complexes, operon, regulon, and KEGG module). Note that for the first few gene pairs the lines overlap, except the line for protein complexes. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control). The correspondence between ranking and |PCC| is shown below the graph.

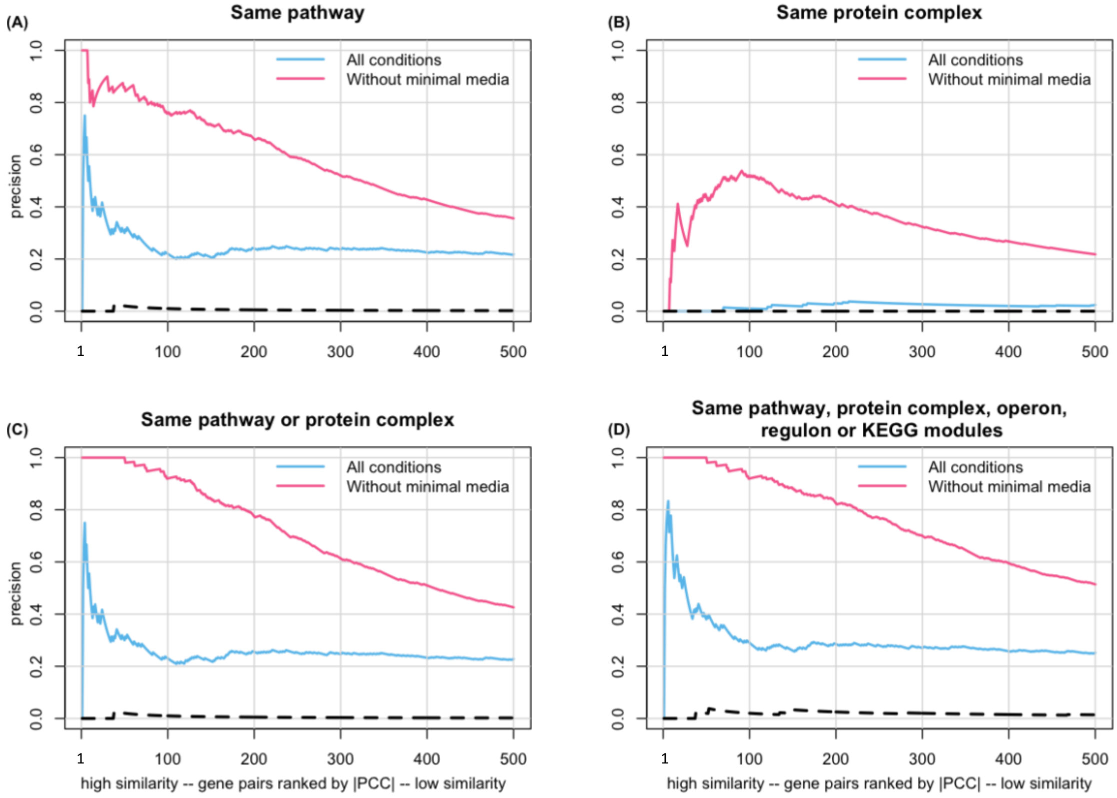
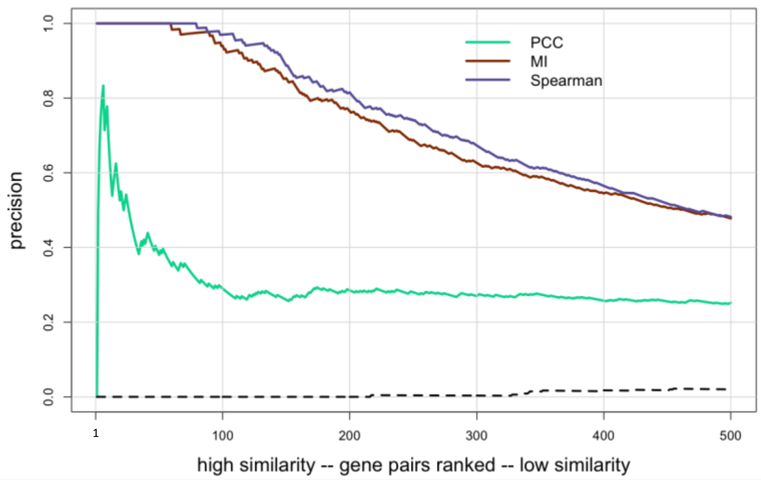


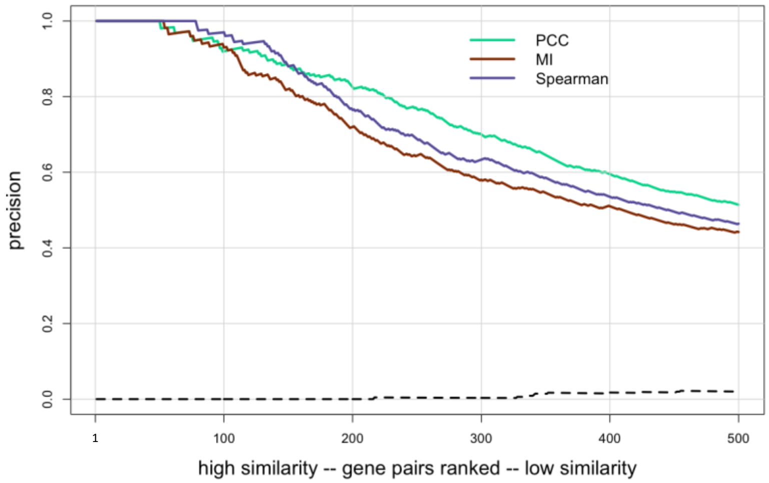
FIGURE 2.3. Precision increased when minimal media conditions were excluded. Gene pairs were ranked from high to low similarity based on |PCC| and plotted versus precision, calculated as described in the text (only the first 500 gene pairs are shown). The four panels show (A) gene pairs annotated to the same EcoCyc pathway, (B) gene pairs annotated to the same heteromeric protein complex, (C) gene pairs annotated to either the same EcoCyc pathway or protein complex, and (D) gene pairs co-annotated in any of annotation sets 1 through 5. The dashed lines show precision for randomly ordered gene pairs generated as described in the Methods (negative control). The correspondence between ranking and |PCC| is the same as in Figure 2.

(a)



Ranking Similarity	1st	100th	200th	300th	400th	500th
|PCC|	0.96	0.92	0.90	0.89	0.87	0.86
MI	1.20	0.60	0.47	0.42	0.39	0.37
|Spearman|	0.94	0.76	0.66	0.63	0.61	0.59

(b)



Ranking Similarity	1st	100th	200th	300th	400th	500th
|PCC|	0.96	0.77	0.68	0.64	0.62	0.61
MI	1.68	0.83	0.65	0.58	0.55	0.52
|Spearman|	0.94	0.75	0.66	0.63	0.61	0.60

FIGURE 2.4. Precision versus ranking when different metrics are used to measure phenotypic profile similarity. Gene pairs were ranked from high to low similarity determined using either |PCC|, MI, or |SRCC| and plotted versus precision, using the union of annotation sets 1 through 5 to identify co-annotated gene pairs. Only the first 500 gene pairs are shown. Phenotypic profile similarity was assessed using either (A) all growth conditions or (B) excluding growth conditions with minimal media. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control). The correspondence between ranking and similarity scores is shown below each graph.

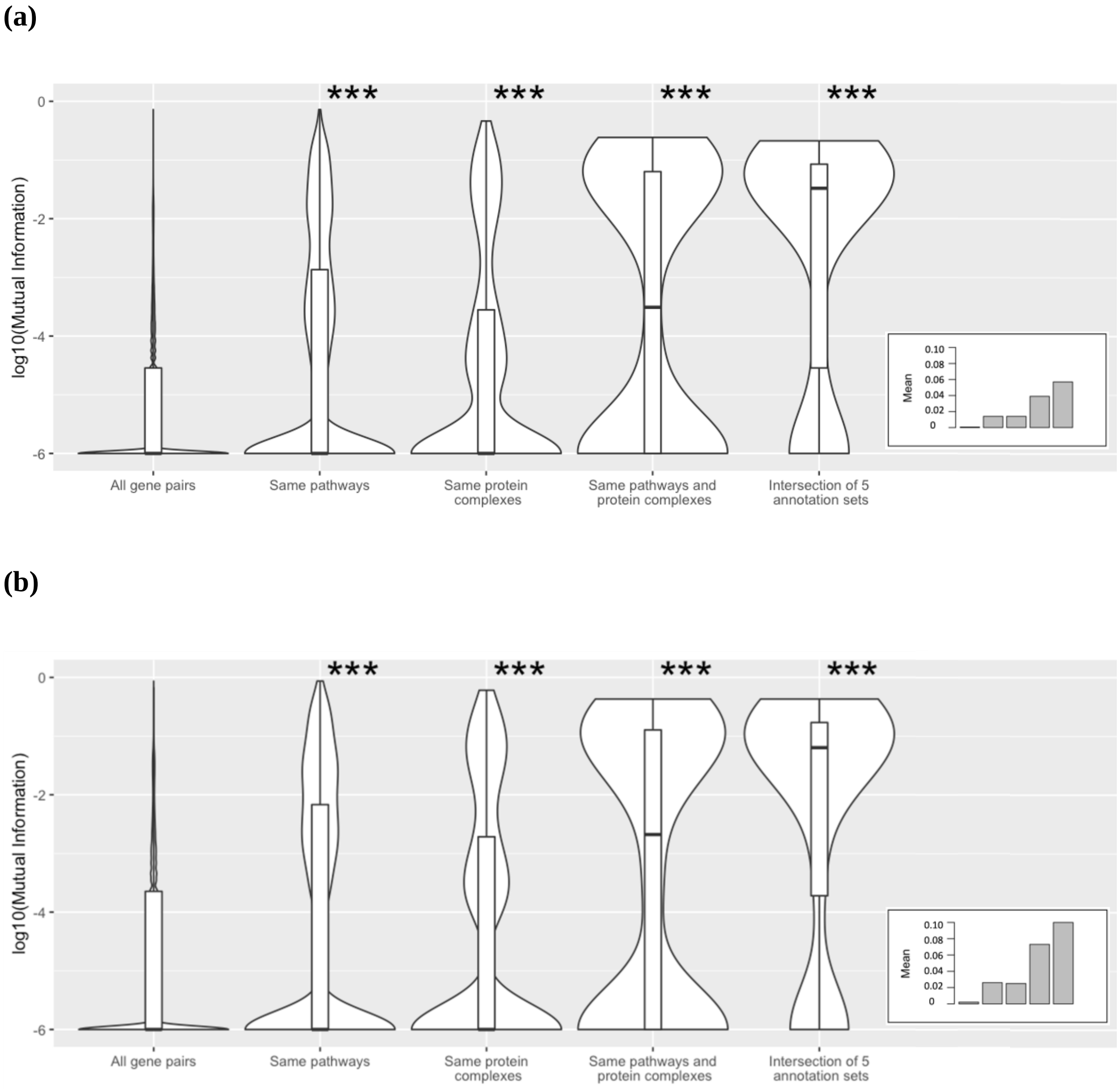
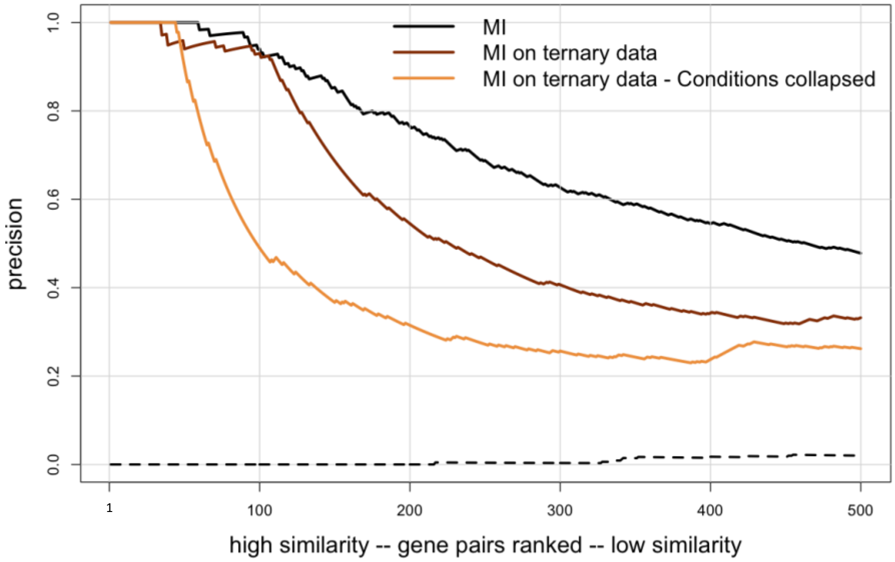

(a)
***
***
***
***
0
-2
-4
-6
log10(Mutual Information)
All gene pairs
Same pathways
Same protein complexes
Same pathways and protein complexes
Intersection of 5 annotation sets
Mean
0.10
0.08
0.06
0.04
0.02
0
(b)
***
***
***
***
0
-2
-4
-6
log10(Mutual Information)
All gene pairs
Same pathways
Same protein complexes
Same pathways and protein complexes
Intersection of 5 annotation sets
Mean
0.10
0.08
0.06
0.04
0.02
0

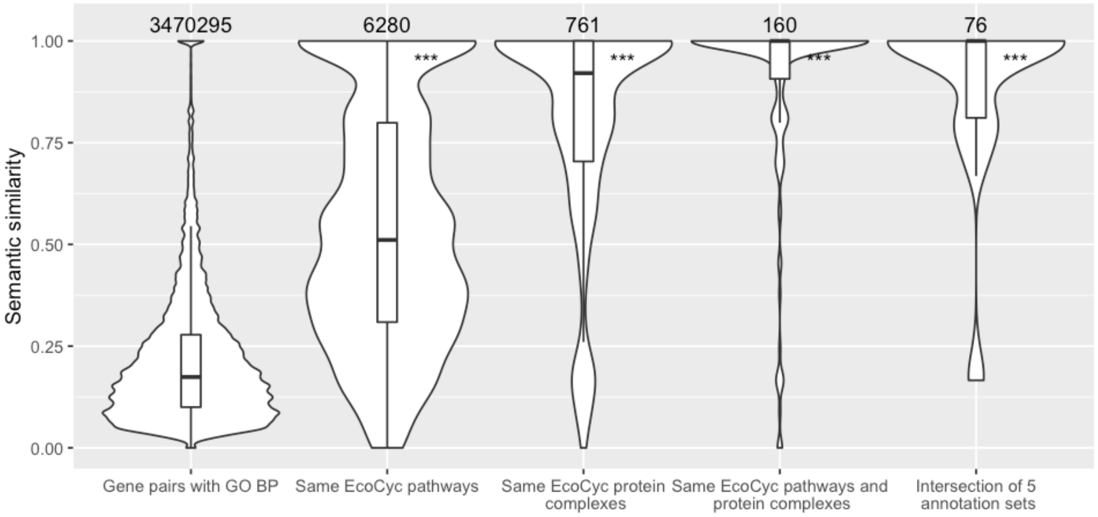
FIGURE 2.5. Phenotypic profile similarity after converting fitness scores from quantitative to qualitative, ternary values. Violin plots of the distributions of phenotypic profile similarity based on Mutual Information for, from left to right, all gene pairs, gene pairs annotated to the same EcoCyc pathway, gene pairs annotated to the same heteromeric protein complex, gene pairs annotated to the same EcoCyc pathway and protein complex, and gene pairs that are co-annotated in annotation sets 1 through 5. The MI values were log transformed after addition of a constant (1x10^{-6}) to eliminate zero values. The middle line within the box plots represents the median. Panel (A) shows the results when profile similarity was determined using all 324 growth conditions. The mean values of the distributions in (A) are 0.0006, 0.014, 0.014, 0.039, and 0.057. Panel (B) shows the results when profile similarity was determined after collapsing the growth conditions to 114 unique stresses. The mean values of the distributions in (B) are 0.0021, 0.026, 0.025, 0.073, and 0.1. ***: p-value <0.001 determined by 1-sided Mann-Whitney U test, compared to all gene pairs.



Ranking / Similarity	1st	100th	200th	300th	400th	500th
MI	1.20	0.60	0.47	0.42	0.39	0.37
MI ternary	0.72	0.20	0.20	0.20	0.20	0.18
MI ternary – collapsed	0.87	0.43	0.43	0.43	0.42	0.39

FIGURE 2.6. Precision versus ranking for quantitative versus discretized, ternary fitness scores. Gene pairs were ranked from high to low similarity based on Mutual Information and plotted versus precision using the union of annotation sets 1 through 5 to identify co-annotated gene pairs. Only the first 500 gene pairs are shown. Phenotypic profile similarity was determined with either the original quantitative fitness scores (black line), the discretized ternary scores for all growth conditions (brown line), or the discretized, ternary scores for growth conditions collapsed to 114 unique stresses (orange line). The cutoffs used to convert the quantitative scores to discretized, ternary scores were based on the 5% FDR for each condition. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control). The correspondence between ranking and similarity scores is shown below each graph.

(a)


3470295
6280
761
160
76
***
***
***
***
Semantic similarity
1.00
0.75
0.50
0.25
0.00
Gene pairs with GO BP
Same EcoCyc pathways
Same EcoCyc protein complexes
Same EcoCyc pathways and protein complexes
Intersection of 5 annotation sets

(b)

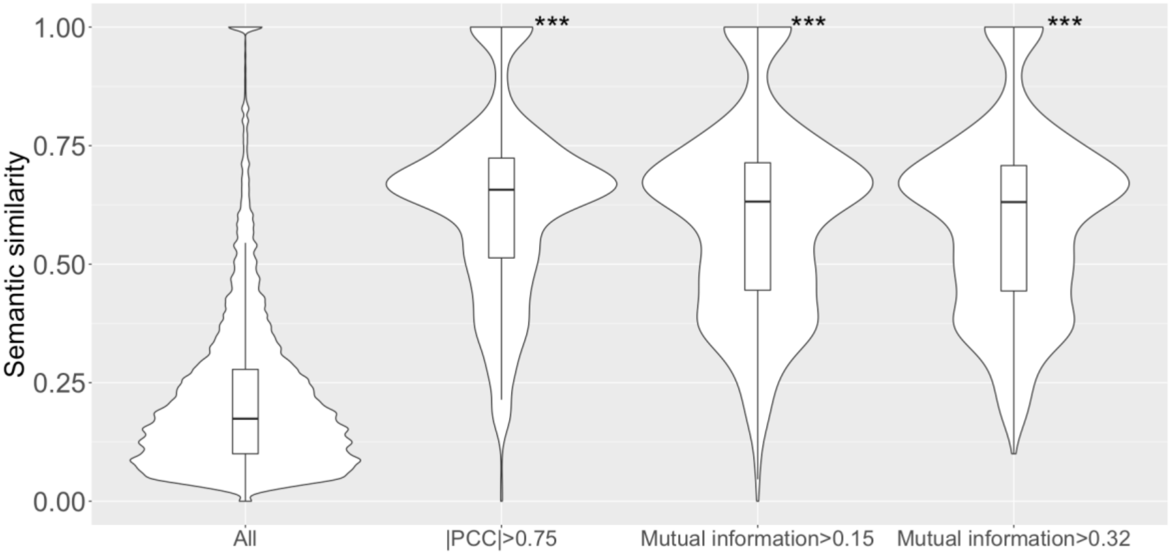

***
***
***
Semantic similarity
1.00
0.75
0.50
0.25
0.00
All
|PCC|>0.75
Mutual information>0.15
Mutual information>0.32

FIGURE 2.7. Higher semantic similarity and phenotypic profile similarity were found for co-annotated gene pairs. (A) Violin plots of the distributions of semantic similarity for, from left to right, all gene pairs annotated with GO biological process term(s), gene pairs annotated to the same EcoCyc pathway, gene pairs annotated to the same heteromeric protein complex, gene pairs annotated to both the same EcoCyc pathway and the same protein complex, and gene pairs co-annotated in annotation sets 1 through 5. Numbers above each violin plot indicate the number of gene pairs in each plot. (B) Violin plots of semantic similarity for, from left to right: all gene pairs annotated with GO biological process term(s); the subset of gene pairs with |PCC| >0.75; the subset of gene pairs with MI >0.15 (calculated based on qualitative fitness scores for all growth conditions); and MI >0.32 (calculated based on qualitative fitness scores for the collapsed set of growth conditions). The cutoffs of MI >0.15 for the third violin plot and MI >0.32 for the fourth violin plot were chosen so that all three subsets of gene pairs would contain the same number (~1,200) of top-ranked gene pairs. ***: p-value <0.001 was determined by 1-sided Mann-Whitney U test, compared to all gene pairs.

Supplemental material

TABLE 2.S1. The 366 EcoCyc pathways used in this study. This table provides the numeric labels used to identify the pathways shown in Figure S1, the pathway IDs used in the EcoCyc database, and the common names for the pathways.

Label No.	EcoCyc Pathway ID	Pathway name
1	HOMOSER-THRESYN-PWY	L-threonine biosynthesis
2	PWY0-1505	ArcAB two-component signal transduction system, quinone dependent
3	XYLCAT-PWY	xylose degradation I
4	PYRUVDEHYD-PWY	pyruvate decarboxylation to acetyl CoA
5	PWY0-1458	PhoQP two-component signal transduction system, magnesium-dependent
6	PWY0-1487	CreCB two-component signal transduction system
7	GLUTATHIONESYN-PWY	glutathione biosynthesis
8	PWY0-1509	NtrBC two-component signal transduction system, nitrogen-dependent
9	PWY0-1474	AtoSC two-component signal transduction system
10	PWY-6890	4-amino-2-methyl-5-diphosphomethylpyrimidine biosynthesis
11	PWY0-1554	5-(carboxymethoxy)uridine biosynthesis
12	PWY-66	GDP-L-fucose biosynthesis I (from GDP-D-mannose)
13	GLUTDEG-PWY	L-glutamate degradation II
14	PWY-7335	UDP-N-acetyl-alpha-D-mannosaminouronate biosynthesis
15	PWY0-1500	EnvZ two-component signal transduction system, osmotic responsive
16	PWY0-1470	QseBC two-component signal transduction system, quorum sensing related
17	PWY0-1468	DcuSR two-component signal transduction system, dicarboxylate-dependent
18	PWY-6153	autoinducer AI-2 biosynthesis I
19	PWY0-1490	EvgSA two-component signal transduction system
20	BETSYN-PWY	glycine betaine biosynthesis I (Gram-negative bacteria)
21	PWY0-1499	DpiBA two-component signal transduction system
22	PWY-7343	UDP-alpha-D-glucose biosynthesis I
23	2PHENDEG-PWY	phenylethylamine degradation I

24	PWY0-1264	biotin-carboxyl carrier protein assembly
25	PWY-7761	NAD salvage pathway II
26	PWY0-1559	BtsSR two-component signal transduction system
27	PWY0-1550	YpdAB two-component signal transduction system
28	GLUAMCAT-PWY	N-acetylglucosamine degradation I
29	GLUTSYN-PWY	L-glutamate biosynthesis I
30	GLUCONSUPER-PWY	D-gluconate degradation
31	RIBOKIN-PWY	ribose phosphorylation
32	PWY-6910	hydroxymethylpyrimidine salvage
33	ALKANEMONOX-PWY	Two-component alkanesulfonate monooxygenase
34	PWY-6147	6-hydroxymethyl-dihydropterin diphosphate biosynthesis I
35	PWY-40	putrescine biosynthesis I
36	PWY0-1182	trehalose degradation II (trehalase)
37	PWY0-461	L-lysine degradation I
38	TREDEGLOW-PWY	Trehalose degradation I (low osmolarity)
39	PWY0-1492	UhpBA two-component signal transduction system
40	PWY0-1483	PhoRB two-component signal transduction system, phosphate-dependent
41	PWY0-1485	CpxAR two-component signal transduction system
42	PWY-901	methylglyoxal degradation II (no longer recognized as a pathway in ecocyc)
43	PWY0-1587	N6-L-threonylcarbamoyladenosine37-modified tRNA biosynthesis
44	PWY0-1498	ZraSR two-component signal transduction system
45	PWY0-1482	BasSR two-component signal transduction system
46	CYANCAT-PWY	cyanate degradation
47	PWY-7247	beta-D-glucuronide and D-glucuronate degradation
48	PWY0-1021	L-alanine biosynthesis III
49	PWY-2161	folate polyglutamylation
50	PWY0-1503	GlrKR two-component signal transduction system
51	PWY-6019	pseudouridine degradation
52	ENTNER-DOUDOROFF-PWY	Entner-Doudoroff pathway I
53	BSUBPOLYAMSYN-PWY	spermidine biosynthesis I
54	TRESYN-PWY	trehalose biosynthesis I
55	PWY0-1477	ethanolamine utilization
56	PWY-7194	pyrimidine nucleobases salvage II
57	PWY0-1433	tetrahydromonapterin biosynthesis
58	PWY-6605	adenine and adenosine salvage II

59	PWY0-1588	HprSR two-component signal transduction system
60	PWY0-1280	ethylene glycol degradation
61	PWY0-1317	L-lactaldehyde degradation (aerobic)
62	PWY-5459	methylglyoxal degradation IV
63	ALANINE-SYN2-PWY	L-alanine biosynthesis II
64	PWY-7179	purine deoxyribonucleosides degradation I
65	PWY-7176	UTP and CTP de novo biosynthesis
66	PWY0-1519	aerotactic two-component signal transduction system
67	PWY0-1481	BaeSR two-component signal transduction system
68	PWY0-1501	BarA UvrY two-component signal transduction system
69	PWY0-1512	CusSR two-component signal transduction system
70	PWY0-1506	TorSR two-component signal transduction system, TMAO dependent
71	PWY-6703	preQ$_0$ biosynthesis
72	PWY-7197	pyrimidine deoxyribonucleotide phosphorylation
73	PWY-7205	CMP phosphorylation
74	PWY0-1534	hydrogen sulfide biosynthesis I
75	ASPARAGINESYN-PWY	L-asparagine biosynthesis II
76	PWY0-1325	superpathway of L-asparagine biosynthesis
77	PWY-7193	pyrimidine ribonucleosides salvage I
78	PWY-6537	4-aminobutanoate degradation II
79	PWY0-1495	KdpDE two-component signal transduction system, potassium-dependent
80	PWY0-1517	sedoheptulose bisphosphate bypass
81	PWY0-1309	chitobiose degradation
82	PWY0-1497	RstBA two-component signal transduction system
83	PWY-5123	trans, trans-farnesyl diphosphate biosynthesis
84	PWY0-661	PRPP biosynthesis II
85	PROSYN-PWY	L-proline biosynthesis I
86	GLYCLEAV-PWY	glycine cleavage
87	SERSYN-PWY	L-serine biosynthesis
88	PWY-5340	sulfate activation for sulfonation
89	PWY-5901	2,3-dihydroxybenzoate biosynthesis
90	CYSTSYN-PWY	L-cysteine biosynthesis I
91	PWY0-1515	NarX two-component signal transduction system, nitrate dependent
92	KDOSYN-PWY	kdo transfer to lipid IVA I
93	PWY0-1514	NarQ two-component signal transduction system, nitrate dependent

94	PWY0-1275	lipoate biosynthesis and incorporation II
95	PWY0-901	L-selenocysteine biosynthesis I (bacteria)
96	PWY0-521	fructoselysine and psicoselysine degradation
97	PANTO-PWY	phosphopantothenate biosynthesis I
98	PWY-7221	guanosine ribonucleotides de novo biosynthesis
99	AMMASSIM-PWY	ammonia assimilation cycle III
100	PWY-5965	fatty acid biosynthesis initiation III
101	IDNCAT-PWY	L-idonate degradation
102	LYXMET-PWY	L-lyxose degradation
103	PUTDEG-PWY	putrescine degradation I
104	GALACTCAT-PWY	D-galactonate degradation
105	HOMOSERSYN-PWY	L-homoserine biosynthesis
106	PWY-1801	formaldehyde oxidation II (glutathione-dependent)
107	THREONINE-DEG2-PWY	L-threonine degradation II
108	PWY0-1303	aminopropylcadaverine biosynthesis
109	PWY0-1312	acetate formation from acetyl-CoA I
110	SALVPURINE2-PWY	xanthine and xanthosine salvage
111	ASPARAGINE-DEG1-PWY	L-asparagine degradation I
112	PWY0-44	D-allose degradation
113	ALADEG-PWY	L-alanine degradation I
114	NADPHOS-DEPHOS-PWY	NAD phosphorylation and dephosphorylation
115	PWY0-1493	RcsCDB two-component signal transduction system
116	PPGPPMET-PWY	ppGpp biosynthesis
117	PWY-6543	4-aminobenzoate biosynthesis
118	PLPSAL-PWY	pyridoxal 5'-phosphate salvage I
119	PWY0-1415	superpathway of heme b biosynthesis from uroporphyrinogen-III
120	PWY0-1518	chemotactic two-component signal transduction
121	OXIDATIVEPENT-PWY	pentose phosphate pathway (oxidative branch) I
122	PWY-6038	citrate degradation
123	PWY0-823	L-arginine degradation III (arginine decarboxylase/agmatinase pathway)
124	PWY-7181	pyrimidine deoxyribonucleosides degradation
125	THIOREDOX-PWY	thioredoxin pathway
126	PWY0-1337	oleate beta-oxidation
127	PWY-6614	tetrahydrofolate biosynthesis
128	PWY-6535	4-aminobutanoate degradation I
129	PWY0-1300	2-O-alpha-mannosyl-D-glycerate degradation

130	PWY-7208	superpathway of pyrimidine nucleobases salvage
131	PWY-5698	allantoin degradation to ureidoglycolate II (ammonia producing)
132	PYRIDNUCSAL-PWY	NAD salvage pathway I
133	ETOH-ACETYLCOA-ANA-PWY	ethanol degradation I
134	PWY-5162	2-oxopentenoate degradation
135	THRDLCTCAT-PWY	L-threonine degradation III (to methylglyoxal)
136	UDPNAGSYN-PWY	UDP-N-acetyl-D-glucosamine biosynthesis I
137	PWY0-1319	CDP-diacylglycerol biosynthesis II
138	PWY0-1569	autoinducer AI-2 degradation
139	PWY-5436	L-threonine degradation IV
140	PWY0-1324	N-acetylneuraminate and N-acetylmannosamine degradation I
141	PWY0-43	conversion of succinate to propanoate
142	SER-GLYSYN-PWY	superpathway of L-serine and glycine biosynthesis I
143	PWY0-1241	ADP-L-glycero-beta-D-manno-heptose biosynthesis
144	PWY-6708	ubiquinol-8 biosynthesis (prokaryotic)
145	PWY-7545	pyruvate to cytochrome bd oxidase electron transfer
146	PYRIDNUCSYN-PWY	NAD biosynthesis I (from aspartate)
147	PWY0-1568	NADH to cytochrome bd oxidase electron transfer II
148	PANTOSYN-PWY	superpathway of coenzyme A biosynthesis I (bacteria)
149	PWY-7242	D-fructuronate degradation
150	PWY-6897	thiamine salvage II
151	GLYCEROLMETAB-PWY	glycerol degradation V
152	FUCCAT-PWY	fucose degradation
153	PWY-6556	pyrimidine ribonucleosides salvage II
154	PWY0-1338	polymyxin resistance
155	PWY-5966	fatty acid biosynthesis initiation II
156	PWY-7195	pyrimidine ribonucleosides salvage III
157	PWY-7446	sulfoquinovose degradation I
158	ACETOACETATE-DEG-PWY	acetoacetate degradation (to acetyl CoA)
159	PWY0-301	L-ascorbate degradation I (bacterial, anaerobic)
160	KDO-LIPASYN-PWY	(Kdo)2-lipid A biosynthesis I
161	GLYCOGENSYNTH-PWY	glycogen biosynthesis I (from ADP-D-Glucose)
162	PWY-6700	queuosine biosynthesis
163	AST-PWY	L-arginine degradation II (AST pathway)
164	ALANINE-VALINESYN-PWY	L-alanine biosynthesis I
165	PWY-4381	fatty acid biosynthesis initiation I

166	PWY0-1507	biotin biosynthesis from 8-amino-7-oxononanoate I
167	PWY-6611	adenine and adenosine salvage V
168	PWY0-1573	nitrate reduction VIIIb (dissimilatory)
169	PWY-7180	2'-deoxy-alpha-D-ribose 1-phosphate degradation
170	SERDEG-PWY	L-serine degradation
171	DARABCATK12-PWY	D-arabinose degradation I
172	PWY-5785	di-trans,poly-cis-undecaprenyl phosphate biosynthesis
173	PWY0-1221	putrescine degradation II
174	TYRSYN	L-tyrosine biosynthesis I
175	PWY0-1545	cardiolipin biosynthesis III
176	PWY0-181	salvage pathways of pyrimidine deoxyribonucleotides
177	PWY-1269	CMP-3-deoxy-D-manno-octulosonate biosynthesis
178	PWY-7206	pyrimidine deoxyribonucleotides dephosphorylation
179	PWY-5705	allantoin degradation to glyoxylate III
180	PWY0-1295	pyrimidine ribonucleosides degradation
181	GLYOXDEG-PWY	glycolate and glyoxylate degradation II
182	PWY-6164	3-dehydroquinate biosynthesis I
183	CARNMET-PWY	L-carnitine degradation I
184	PWY-5350	thiosulfate disproportionation IV (rhodanese)
185	PWY-5659	GDP-mannose biosynthesis
186	PWY-6122	5-aminoimidazole ribonucleotide biosynthesis II
187	PWY-6121	5-aminoimidazole ribonucleotide biosynthesis I
188	PWY0-1565	D-lactate to cytochrome bo oxidase electron transfer
189	PWY0-1567	NADH to cytochrome bo oxidase electron transfer II
190	PWY0-1544	proline to cytochrome bo oxidase electron transfer
191	PWY-7544	pyruvate to cytochrome bo oxidase electron transfer
192	PWY0-1561	glycerol-3-phosphate to cytochrome bo oxidase electron transfer
193	PWY-6123	inosine-5'-phosphate biosynthesis I
194	UBISYN-PWY	superpathway of ubiquinol-8 biosynthesis (prokaryotic)
195	TRPSYN-PWY	L-tryptophan biosynthesis
196	PWY0-501	lipoate biosynthesis and incorporation I
197	DAPLYSINESYN-PWY	L-lysine biosynthesis I
198	GALACTUROCAT-PWY	D-galacturonate degradation I
199	GALACTMETAB-PWY	galactose degradation I (Leloir pathway)
200	LCYSDEG-PWY	L-cysteine degradation II
201	ACETATEUTIL-PWY	superpathway of acetate utilization and formation

202	PWY0-41	allantoin degradation IV (anaerobic)
203	PWY-6961	L-ascorbate degradation II (bacterial, aerobic)
204	COBALSYN-PWY	adenosylcobalamin salvage from cobinamide I
205	PWY-6012	acyl carrier protein metabolism
206	FASYN-INITIAL-PWY	superpathway of fatty acid biosynthesis initiation (E. coli)
207	PWY-4621	arsenate detoxification II (glutaredoxin)
208	DTDPRHAMSYN-PWY	dTDP-L-rhamnose biosynthesis I
209	GALACTARDEG-PWY	D-galactarate degradation I
210	PWY-6620	guanine and guanosine salvage
211	PHESYN	L-phenylalanine biosynthesis I
212	PWY-4261	glycerol degradation I
213	PWY-5386	methylglyoxal degradation I
214	PWY-5668	cardiolipin biosynthesis I
215	GLUCARDEG-PWY	D-glucarate degradation I
216	PWY0-1296	purine ribonucleosides degradation
217	PWY-6151	S-adenosyl-L-methionine cycle I
218	PWY0-1546	muropeptide degradation
219	GLUT-REDOX-PWY	glutathione-glutaredoxin redox reactions
220	GLCMANNANAUT-PWY	superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation
221	PWY0-1471	uracil degradation III
222	PWY-5971	palmitate biosynthesis II (bacteria and plants)
223	PWY0-862	(5Z)-dodec-5-enoate biosynthesis I
224	4AMINOBUTMETAB-PWY	superpathway of 4-aminobutanoate degradation
225	PWY-6277	superpathway of 5-aminoimidazole ribonucleotide biosynthesis
226	GLUTORN-PWY	L-ornithine biosynthesis I
227	PYRIDOXSYN-PWY	pyridoxal 5'-phosphate biosynthesis I
228	THRESYN-PWY	superpathway of L-threonine biosynthesis
229	P2-PWY	citrate lyase activation
230	DETOX1-PWY	superoxide radicals degradation
231	RIBOSYN2-PWY	flavin biosynthesis I (bacteria and plants)
232	PWY0-1584	nitrate reduction X (dissimilatory, periplasmic)
233	GLUCUROCAT-PWY	superpathway of beta-D-glucuronosides degradation
234	PWY-6579	superpathway of guanine and guanosine salvage
235	PWY-7315	dTDP-N-acetylthomosamine biosynthesis
236	HOMOSER-METSYN-PWY	L-methionine biosynthesis I

237	NRI-PWY	Nitrogen regulation two-component system
238	PWY-6952	glycerophosphodiester degradation
239	PWY-5437	L-threonine degradation I
240	GLUCARGALACTSUPER-PWY	superpathway of D-glucarate and D-galactarate degradation
241	PWY-6609	adenine and adenosine salvage III
242	PWY-5453	methylglyoxal degradation III
243	PWY0-42	2-methylcitrate cycle I
244	PWY-6163	chorismate biosynthesis from 3-dehydroquinate
245	PWY0-1297	superpathway of purine deoxyribonucleosides degradation
246	GLYOXYLATE-BYPASS	glyoxylate cycle
247	POLYISOPRENSYN-PWY	polyisoprenoid biosynthesis (E. coli)
248	PWY-6282	palmitoleate biosynthesis I (from (5Z)-dodec-5-enoate)
249	FASYN-ELONG-PWY	fatty acid elongation - saturated
250	LEUSYN-PWY	L-leucine biosynthesis
251	ILEUSYN-PWY	L-isoleucine biosynthesis I (from threonine)
252	METSYN-PWY	L-homoserine and L-methionine biosynthesis
253	PWY0-1353	succinate to cytochrome bd oxidase electron transfer
254	ASPASN-PWY	superpathway of L-aspartate and L-asparagine biosynthesis
255	PWY0-1533	methylphosphonate degradation I
256	PWY-7220	adenosine deoxyribonucleotides de novo biosynthesis II
257	PWY-7222	guanosine deoxyribonucleotides de novo biosynthesis II
258	PWY0-1582	glycerol-3-phosphate to fumarate electron transfer
259	NONOXIPENT-PWY	pentose phosphate pathway (non-oxidative branch)
260	FAO-PWY	fatty acid beta-oxidation I
261	ORNDEG-PWY	superpathway of ornithine degradation
262	KETOGLUCONMET-PWY	ketogluconate metabolism
263	PWY0-381	glycerol and glycerophosphodiester degradation
264	PWY-5837	1,4-dihydroxy-2-naphthoate biosynthesis
265	GLYCOCAT-PWY	glycogen degradation I
266	PWY-7187	pyrimidine deoxyribonucleotides de novo biosynthesis II
267	PWY-7184	pyrimidine deoxyribonucleotides de novo biosynthesis I
268	PWY0-1298	superpathway of pyrimidine deoxyribonucleosides degradation
269	GLYCOLATEMET-PWY	glycolate and glyoxylate degradation I
270	PWY-6284	superpathway of unsaturated fatty acids biosynthesis (E. coli)

271	PWY-5973	cis-vaccenate biosynthesis
272	GLUCOSE1PMETAB-PWY	glucose and glucose-1-phosphate degradation
273	SO4ASSIM-PWY	sulfate reduction I (assimilatory)
274	PWY-5686	UMP biosynthesis I
275	PWY0-1329	succinate to cytochrome bo oxidase electron transfer
276	VALSYN-PWY	L-valine biosynthesis
277	ENTBACSYN-PWY	enterobactin biosynthesis
278	PWY-6892	thiazole biosynthesis I (facultative anaerobic bacteria)
279	PWY0-845	superpathway of pyridoxal 5'-phosphate biosynthesis and salvage
280	GALACT-GLUCUROCAT-PWY	superpathway of hexuronide and hexuronate degradation
281	NAGLIPASYN-PWY	lipid IVA biosynthesis
282	PWY-6690	cinnamate and 3-hydroxycinnamate degradation to 2-oxopent-4-enoate
283	HCAMHPDEG-PWY	3-phenylpropanoate and 3-(3-hydroxyphenyl)propanoate degradation to 2-oxopent-4-enoate
284	GALACTITOLCAT-PWY	galactitol degradation
285	PWY-6612	superpathway of tetrahydrofolate biosynthesis
286	PWY0-1355	formate to trimethylamine N-oxide electron transfer
287	PWY0-1576	hydrogen to fumarate electron transfer
288	FUC-RHAMCAT-PWY	superpathway of fucose and rhamnose degradation
289	PWY0-1061	superpathway of L-alanine biosynthesis
290	PWY0-1479	tRNA processing
291	PWY-6519	8-amino-7-oxononanoate biosynthesis I
292	PWY0-163	salvage pathways of pyrimidine ribonucleotides
293	NONMEVIPP-PWY	methylerythritol phosphate pathway I
294	PWY0-881	superpathway of fatty acid biosynthesis I (E. coli)
295	HISTSYN-PWY	L-histidine biosynthesis
296	LIPA-CORESYN-PWY	Lipid A-core biosynthesis
297	PWY-6823	molybdenum cofactor biosynthesis
298	PWY-6125	superpathway of guanosine nucleotides de novo biosynthesis II
299	PWY0-1581	nitrate reduction IX (dissimilatory)
300	PWY0-1356	formate to dimethyl sulfoxide electron transfer
301	PWY0-1578	hydrogen to trimethylamine N-oxide electron transfer
302	POLYAMSYN-PWY	superpathway of polyamine biosynthesis I
303	OANTIGEN-PWY	O-antigen building blocks biosynthesis (E. coli)
304	PHOSLIPSYN-PWY	superpathway of phospholipid biosynthesis I (bacteria)

305	PWY-7196	superpathway of pyrimidine ribonucleosides salvage
306	ECASYN-PWY	enterobacterial common antigen biosynthesis
307	PWY0-162	superpathway of pyrimidine ribonucleotides de novo biosynthesis
308	PWY-7219	adenosine ribonucleotides de novo biosynthesis
309	GLUTAMINDEG-PWY	L-glutamine degradation I
310	MET-SAM-PWY	superpathway of S-adenosyl-L-methionine biosynthesis
311	1CMET2-PWY	N10-formyl-tetrahydrofolate biosynthesis
312	PWY0-1577	hydrogen to dimethyl sulfoxide electron transfer
313	PENTOSE-P-PWY	pentose phosphate pathway
314	ARO-PWY	chorismate biosynthesis I
315	COLANSYN-PWY	colanic acid building blocks biosynthesis
316	PWY0-1261	anhydromuropeptides recycling I
317	PWY0-1585	formate to nitrite electron transfer
318	PWY0-321	phenylacetate degradation I (aerobic)
319	PWY-5838	superpathway of menaquinol-8 biosynthesis I
320	THISYN-PWY	superpathway of thiamine diphosphate biosynthesis I
321	PWY-6387	UDP-N-acetylmuramoyl-pentapeptide biosynthesis I (meso-diaminopimelate containing)
322	PWY-7805	aminomethylphosphonate degradation
323	PWY-6608	guanosine nucleotides degradation III
324	GLYCOL-GLYOXDEG-PWY	superpathway of glycol metabolism and degradation
325	ARGSYN-PWY	L-arginine biosynthesis I (via L-ornithine)
326	PEPTIDOGLYCANSYN-PWY	peptidoglycan biosynthesis I (meso-diaminopimelate containing)
327	PWY0-1277	3-phenylpropanoate and 3-(3-hydroxyphenyl)propanoate degradation
328	PWY0-1321	nitrate reduction III (dissimilatory)
329	ARGDEG-PWY	superpathway of L-arginine, putrescine, and 4-aminobutanoate degradation
330	BIOTIN-BIOSYNTHESIS-PWY	biotin biosynthesis I
331	TRNA-CHARGING-PWY	tRNA charging
332	PWY-6071	superpathway of phenylethylamine degradation
333	PWY0-166	superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis (E. coli)
334	SALVADEHYPOX-PWY	adenosine nucleotides degradation II
335	METHGLYUT-PWY	superpathway of methylglyoxal degradation
336	PWY0-1347	NADH to trimethylamine N-oxide electron transfer
337	ORNARGDEG-PWY	superpathway of L-arginine and L-ornithine degradation
338	PWY0-1334	NADH to cytochrome bd oxidase electron transfer I

339	PWY0-1348	NADH to dimethyl sulfoxide electron transfer
340	SULFATE-CYS-PWY	superpathway of sulfate assimilation and cysteine biosynthesis
341	PWY0-1335	NADH to cytochrome bo oxidase electron transfer I
342	PWY0-1336	NADH to fumarate electron transfer
343	P4-PWY	superpathway of L-lysine, L-threonine and L-methionine biosynthesis I
344	PWY-7211	superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis
345	BRANCHED-CHAIN-AA-SYN-PWY	superpathway of branched chain amino acid biosynthesis
346	PWY0-1586	peptidoglycan maturation (meso-diaminopimelate containing)
347	TCA	TCA cycle I (prokaryotic)
348	PWY-6126	superpathway of adenosine nucleotides de novo biosynthesis II
349	GLUCONEO-PWY	gluconeogenesis I
350	PWY0-1352	nitrate reduction VIII (dissimilatory)
351	KDO-NAGLIPASYN-PWY	superpathway of (Kdo)2-lipid A biosynthesis
352	GLYCOLYSIS	glycolysis I (from glucose 6-phosphate)
353	PWY-5484	glycolysis II (from fructose 6-phosphate)
354	COMPLETE-ARO-PWY	superpathway of aromatic amino acid biosynthesis
355	PWY0-781	aspartate superpathway
356	TCA-GLYOX-BYPASS	superpathway of glyoxylate bypass and TCA
357	GLYCOLYSIS-E-D	superpathway of glycolysis and the Entner-Doudoroff pathway
358	THREOCAT-PWY	superpathway of L-threonine metabolism
359	ARG+POLYAMINE-SYN	superpathway of arginine and polyamine biosynthesis
360	LPSSYN-PWY	superpathway of lipopolysaccharide biosynthesis
361	HEXITOLDEGSUPER-PWY	superpathway of hexitol degradation (bacteria)
362	DENOVOPURINE2-PWY	superpathway of purine nucleotides de novo biosynthesis II
363	FERMENTATION-PWY	mixed acid fermentation
364	GLYCOLYSIS-TCA-GLYOX-BYPASS	superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass
365	PRPP-PWY	superpathway of histidine, purine, and pyrimidine biosynthesis
366	ALL-CHORISMATE-PWY	superpathway of chorismate metabolism

TABLE 2.S2. The 271 EcoCyc heteromeric protein complexes used in this study. This table provides the numeric labels used to identify the protein complexes in Figure S2, the protein complex IDs used in the EcoCyc database, and the common names for the protein complexes.

Label No.	EcoCyc protein complex ID	Name of complex
1	3-ISOPROPYLMALISOM-CPLX	3-isopropylmalate dehydratase
2	CPLX0-8178	peptidoglycan glycosyltransferase / peptidoglycan DD-transpeptidase - MrcB-LpoB complex
3	SULFITE-REDUCT-CPLX	assimilatory sulfite reductase (NADPH)
4	TRYPSYN	tryptophan synthase
5	PC00027	DNA-binding transcriptional dual regulator IHF
6	GLUTAMIDOTRANS-CPLX	imidazole glycerol phosphate synthase
7	SULFATE-ADENYLYLTRANS-CPLX	sulfate adenylyltransferase
8	CPLX0-7609	5-carboxymethylaminomethyluridine-tRNA synthase [multifunctional]
9	CPLX0-3107	ClpXP
10	CARBPSYN-CPLX	carbamoyl phosphate synthetase
11	SUCCCOASYN	succinyl-CoA synthetase
12	PYRUVATEDEH-CPLX	pyruvate dehydrogenase
13	ABC-63-CPLX	Zn^{2+} ABC transporter
14	CYSSYNMULTI-CPLX	cysteine synthase complex
15	RNAP70-CPLX	RNA polymerase sigma 70
16	CPLX0-2021	DNA-binding transcriptional dual regulator HU
17	CPLX-3946	exodeoxyribonuclease VII
18	CPLX0-7910	DNA polymerase III, Psi-Chi subunit
19	CPLX0-3949	thiazole synthase
20	CPLX0-1321	HflK-HflC complex; regulator of FtsH protease
21	ANTHRANSYN-CPLX	anthranilate synthase
22	CPLX0-7994	poly-N-acetyl-D-glucosamine synthase
23	CPLX0-7529	polysaccharide export complex
24	CPLX0-2502	molybdopterin synthase
25	CPLX0-3104	ClpAP
26	CPLX0-3959	Xer site-specific recombination system
27	CPLX0-231	galactitol-specific PTS enzyme II
28	CPLX-156	mannitol-specific PTS enzyme II CmtBA
29	NAP-CPLX	periplasmic nitrate reductase

30	TMAOREDUCTI-CPLX	trimethylamine N-oxide reductase 1
31	CPLX0-7720	undecaprenyl-phosphate-alpha-L-Ara4N flippase
32	CPLX0-1163	HslVU protease
33	ABC-6-CPLX	glutathione / L-cysteine ABC exporter CydDC
34	CPLX0-8239	Grx4-IbaG complex
35	ACETOACETYL-COA-TRANSFER-CPLX	acetoacetyl-CoA transferase
36	CPLX0-7852	GadE-RcsB DNA-binding transcriptional activator
37	CPLX0-3925	DNA polymerase V
38	CPLX-63	trimethylamine N-oxide reductase 2
39	ACETOLACTSYNIII-CPLX	acetolactate synthase / acetohydroxybutanoate synthase
40	CPLX0-4	aromatic carboxylic acid efflux pump
41	GLUTAMATESYN-DIMER	glutamate synthase
42	GLUTAMATESYN-CPLX	glutamate synthase
43	CPLX0-3821	HypA-HypB heterodimer
44	PHES-CPLX	phenylalanine-tRNA ligase
45	CPLX0-2661	McrBC restriction endonuclease
46	CPLX0-5	enterobactin export complex EntS-TolC
47	NRDACTMULTI-CPLX	anaerobic nucleoside-triphosphate reductase activating system
48	CPLX0-7976	translocation and assembly module
49	ABC-54-CPLX	divisome protein complex FtsEX
50	CPLX-3945	curli secretion and assembly complex
51	CPLX0-241	tagatose-1,6-bisphosphate aldolase 2
52	CPLX0-7	N-acetylmuramic acid-specific PTS enzyme II
53	ABC-21-CPLX	putative transport complex, ABC superfamily
54	FAO-CPLX	aerobic fatty acid oxidation complex
55	CPLX0-7704	ATP-dependent Lipid A-core flippase
56	RIBONUCLEOSIDE-DIP-REDUCTII-CPLX	ribonucleoside-diphosphate reductase 2
57	DTDPRHAMSYNTHMULTI-CPLX	dTDP-L-rhamnose synthetase complex
58	APP-UBIOX-CPLX	cytochrome bd-II ubiquinol oxidase
59	CPLX0-2221	Colicin E9 translocon
60	CPLX0-8238	putative menaquinol-cytochrome c reductase NrfCD
61	CPLX0-8182	N6-L-threonylcarbamoyladenine synthase
62	CPLX0-3976	Enterobacterial Common Antigen Biosynthesis Protein Complex
63	CPLX0-8179	peptidoglycan glycosyltransferase / peptidoglycan DD-transpeptidase - MrcA-LpoA complex

64	ASPCARBTRANS-CPLX	aspartate carbamoyltransferase
65	CPLX0-8230	HigB-HigA toxin/antitoxin complex and DNA-binding transcriptional repressor
66	PABASYN-CPLX	4-amino-4-deoxychorismate synthase
67	CPLX0-7684	L-valine exporter
68	PC00084	RcsAB DNA-binding transcriptional dual regulator
69	CPLX0-8232	carnitine monooxygenase
70	CPLX0-1668	anaerobic fatty acid beta-oxidation complex
71	RNAP54-CPLX	RNA polymerase sigma54
72	PYRNUTRANSHYDROGEN-CPLX	pyridine nucleotide transhydrogenase
73	ETHAMLY-CPLX	ethanolamine ammonia-lyase
74	YDGEF-CPLX	multidrug/spermidine efflux pump
75	CPLX-159	putative PTS enzyme II FrvAB
76	CPLX0-8213	periplasmic protein-L-methionine sulfoxide reducing system
77	RNAPS-CPLX	RNA polymerase sigma S
78	CPLX-158	fructose-specific PTS enzyme II
79	CPLX0-3922	primosome
80	CPLX0-7909	RnlA-RnlB toxin-antitoxin complex
81	CPLX0-7624	YhaV-PrlF toxin-antitoxin complex
82	CPLX0-7791	RelB-RelE antitoxin/toxin complex / DNA-binding transcriptional repressor
83	CPLX0-7610	N-acetyl-D-galactosamine specific PTS (cryptic)
84	CPLX0-7823	DosC-DosP complex
85	ABC-61-CPLX	putative transport complex, ABC superfamily
86	CPLX0-7787	DinJ-YafQ antitoxin/toxin complex / DNA-binding transcriptional repressor
87	CPLX0-7988	PaaF-PaaG hydratase-isomerase complex
88	CPLX0-3930	FlhDC DNA-binding transcriptional dual regulator
89	CPLX0-8174	Cas1-Cas2 complex
90	CPLX0-245	alkyl hydroperoxide reductase
91	CPLX0-7916	RcsB-BglJ DNA-binding transcriptional activator
92	CPLX0-7788	NAD-dependent dihydropyrimidine dehydrogenase
93	CPLX-157	glucose-specific PTS enzyme II
94	CPLX0-3241	ubiquinol-[NapC cytochrome c] reductase NapGH
95	CPLX0-8227	FicT-FicA complex
96	CPLX0-3937	evolved beta-D-galactosidase
97	CPLX0-1841	predicted xanthine dehydrogenase
98	CPLX0-7942	Grx4-BolA complex

99	SECD-SECF-YAJC-YIDC-CPLX	Sec translocon accessory complex
100	FABZ-CPLX	3-hydroxy-acyl-[acyl-carrier-protein] dehydratase
101	NITRITREDUCT-CPLX	nitrite reductase - NADH dependent
102	MONOMER0-2461	MtlR-HPr
103	LTARTDEHYDRA-CPLX	L(+)-tartrate dehydratase
104	CPLX0-7986	HypCD complex involved in hydrogenase maturation
105	CPLX0-3781	YefM-YoeB antitoxin/toxin complex / DNA-binding transcriptional repressor
106	CPLX0-7425	HipAB toxin/antitoxin complex / DNA-binding transcriptional repressor
107	NRFMULTI-CPLX	periplasmic nitrite reductase NrfAB
108	CPLX0-7822	MqsA-MqsR antitoxin/toxin complex
109	ACETOLACTSYNI-CPLX	acetohydroxybutanoate synthase / acetolactate synthase
110	CPLX0-2561	bacterial condensin MukBEF
111	RNAP32-CPLX	RNA polymerase sigma 32
112	CPLX0-240	tagatose-1,6-bisphosphate aldolase 1
113	CPLX0-3957	ATP dependent structure specific DNA nuclease
114	CPLX-168	trehalose-specific PTS enzyme II
115	CPLX-3942	sulfurtransferase complex TusBCD
116	TRANS-CPLX-201	multidrug efflux pump AcrAB-TolC
117	GCVMULTI-CPLX	glycine cleavage system
118	F-O-CPLX	ATP synthase Fo complex
119	ABC-45-CPLX	intermembrane phospholipid transport system
120	RECFOR-CPLX	RecFOR complex
121	UVRABC-CPLX	excision nuclease UvrABC
122	ENTMULTI-CPLX	enterobactin synthase
123	CYT-D-UBIOX-CPLX	cytochrome bd-I ubiquinol oxidase
124	RUVABC-CPLX	resolvasome
125	CPLX0-7450	flagellar motor switch complex
126	ABC-18-CPLX	D-galactose / methyl-beta-D-galactoside ABC transporter
127	CPLX0-1923	energy transducing Ton complex
128	CPLX0-1924	vitamin B12 outer membrane transport complex
129	MUTHLS-CPLX	MutHLS complex, methyl-directed mismatch repair
130	CPLX0-3108	ClpAXP
131	ABC-19-CPLX	molybdate ABC transporter
132	ANGLYC3PDEHYDROG-CPLX	anaerobic glycerol-3-phosphate dehydrogenase
133	ABC-33-CPLX	xylose ABC transporter
134	ABC-11-CPLX	iron(III) hydroxamate ABC transporter

135	CPLX0-8167	hydrogenase 1, oxygen tolerant hydrogenase
136	FORMHYDROGI-CPLX	hydrogenase 1
137	TRANS-200-CPLX	macrolide ABC exporter
138	CPLX0-1341	SufBC2D Fe-S cluster scaffold complex
139	ABC-12-CPLX	L-glutamine ABC transporter
140	NITRATREDUCTZ-CPLX	nitrate reductase Z
141	CPLX-155	N,N'-diacetylchitobiose-specific PTS enzyme II
142	CPLX0-3958	EcoKI restriction-modification system
143	NITRATREDUCTA-CPLX	nitrate reductase A
144	EIISGA	L-ascorbate specific PTS enzyme II
145	ABC-56-CPLX	aliphatic sulfonate ABC transporter
146	ABC-32-CPLX	thiamin(e) ABC transporter
147	FORMATEDEHYDROGO-CPLX	formate dehydrogenase O
148	RECBCD	exodeoxyribonuclease V
149	DIMESULFREDUCT-CPLX	dimethyl sulfoxide reductase
150	TSR-CPLX	chemotaxis signaling complex - serine sensing
151	TSR-GLUME	chemotaxis signaling complex - serine sensing containing Tsr$^{\text{Glu-methyl}}$
152	TSR-GLN	chemotaxis signaling complex - serine sensing Tsr$^{\text{Gln}}$
153	TSR-GLU	chemotaxis signaling complex - serine sensing Tsr$^{\text{Glu}}$
154	ABC-64-CPLX	taurine ABC transporter
155	CPLX0-8152	cystine / cysteine ABC transporter
156	ABC-2-CPLX	arabinose ABC transporter
157	CPLX0-7807	putative multidrug efflux pump MdtNOP
158	ABC-57-CPLX	multidrug ABC exporter
159	PABSYNMULTI-CPLX	para-aminobenzoate synthase multi-enzyme complex
160	CPLX0-3932	multidrug efflux pump AcrAD-TolC
161	TAP-GLU	chemotaxis signaling complex - dipeptide sensing containing Tap$^{\text{Glu}}$
162	TAP-CPLX	chemotaxis signaling complex - dipeptide sensing
163	TAP-GLUME	chemotaxis signaling complex - dipeptide sensing containing Tap$^{\text{Glu-methyl}}$
164	TAP-GLN	chemotaxis signaling complex - dipeptide sensing containing Tap$^{\text{Gln}}$
165	CPLX0-3801	DNA polymerase III, preinitiation complex
166	CPLX0-761	putative xanthine dehydrogenase
167	CPLX0-2081	dihydroxyacetone kinase
168	CPLX0-2982	FtsH/HflKC protease complex
169	CITLY-CPLX	citrate lyase, inactive

170	ACECITLY-CPLX	citrate lyase
171	CPLX0-2141	multidrug efflux pump AcrEF-TolC
172	CPLX-170	galactosamine-specific PTS enzyme II (cryptic)
173	ABC-49-CPLX	glutathione ABC transporter
174	TRG-CPLX	chemotaxis signaling complex - ribose/galactose/glucose sensing
175	TRG-GLUME	chemotaxis signaling complex - ribose/galactose/glucose sensing containingTrg$^{\text{Glu-Methyl}}$
176	TRG-GLN	chemotaxis signaling complex - ribose/galactose/glucose sensing containing Trg$^{\text{Gln}}$
177	TRG-GLU	chemotaxis signaling complex - ribose/galactose/glucose sensing containing Trg$^{\text{Glu}}$
178	TRANS-CPLX-203	2,3-diketo-L-gulonate:Na^{+} symporter
179	CPLX-169	sorbitol-specific PTS enzyme II
180	SEC-SECRETION-CPLX	Sec Holo-Translocon
181	CPLX0-2121	multidrug efflux pump EmrAB-TolC
182	ABC-5-CPLX	vitamin B12 ABC transporter
183	CPLX0-2361	DNA polymerase III, core enzyme
184	ABC-42-CPLX	D-allose ABC transporter
185	TRANS-CPLX-204	multidrug efflux pump MdtEF-TolC
186	CPLX-165	mannose-specific PTS enzyme II
187	METNIQ-METHIONINE-ABC-CPLX	L-methionine/D-methionine ABC transporter
188	CPLX0-7458	glycolate dehydrogenase
189	ABC-28-CPLX	ribose ABC transporter
190	ALPHA-SUBUNIT-CPLX	formate dehydrogenase N, subcomplex
191	FORMATEDEHYDROGN-CPLX	formate dehydrogenase N
192	CPLX0-2161	multidrug efflux pump EmrKY-TolC
193	EIISGC	putative PTS enzyme II SgcBCA
194	ABC-60-CPLX	putative transport complex, ABC superfamily
195	CPLX0-7805	aldehyde dehydrogenase
196	TAR-CPLX	chemotaxis signaling complex - aspartate sensing
197	TAR-GLUME	chemotaxis signaling complex - aspartate sensing containing Tar$^{\text{Glu-methyl}}$
198	TAR-GLN	chemotaxis signaling complex - aspartate sensing containing Tar$^{\text{Gln}}$
199	TAR-GLU	chemotaxis signaling complex - aspartate sensing containing Tar$^{\text{Glu}}$
200	ABC-48-CPLX	putative transport complex, ABC superfamily
201	ABC-26-CPLX	glycine betaine ABC transporter
202	CPLX0-8119	putative PTS enzyme II FryBCA

203	CYT-O-UBIOX-CPLX	cytochrome bo3 ubiquinol oxidase
204	ABC-10-CPLX	ferric enterobactin ABC transporter
205	ABC-16-CPLX	maltose ABC transporter
206	ABC-7-CPLX	thiosulfate/sulfate ABC transporter
207	F-1-CPLX	ATP synthase F1 complex
208	SUCC-DEHASE	succinate:quinone oxidoreductase subcomplex
209	CPLX0-8160	succinate:quinone oxidoreductase
210	ABC-27-CPLX	phosphate ABC transporter
211	TATABCE-CPLX	twin arginine protein translocation system
212	CPLX0-8120	putative ABC transporter ArtPQMI
213	CPLX0-1941	ferric enterobactin outer membrane transport complex
214	CPLX0-3323	holocytochrome c synthetase
215	ABC-24-CPLX	spermidine preferential ABC transporter
216	ABC-70-CPLX	sulfate/thiosulfate ABC transporter
217	CPLX0-1721	copper/silver export system
218	CPLX0-3401	fimbrial complex
219	CPLX-160	putative PTS enzyme II FrwCBDPtsA
220	ABC-35-CPLX	heme trafficking system CcmABCDE
221	CPLX0-1601	selenate reductase
222	CPLX0-7952	ferric coprogen outer membrane transport complex
223	ABC-4-CPLX	L-arginine ABC transporter
224	CPLX0-1943	ferric citrate outer membrane transport complex
225	CPLX0-1942	ferrichrome outer membrane transport complex
226	ABC-34-CPLX	sn-glycerol 3-phosphate / glycerophosphodiester ABC transporter
227	CPLX0-1762	phenylacetyl-CoA 1,2-epoxidase
228	ABC-29-CPLX	putrescine ABC exporter
229	ABC-55-CPLX	putative transport complex, ABC superfamily
230	CPLX0-7958	methylphosphonate degradation complex
231	HCAMULTI-CPLX	putative 3-phenylpropionate/cinnamate dioxygenase
232	CPLX0-7935	carbon-phosphorus lyase core complex
233	ABC-25-CPLX	putrescine ABC transporter
234	ABC-14-CPLX	histidine ABC transporter
235	CPLX0-7628	lipopolysaccharide transport system - outer membrane assembly complex
236	ABC-41-CPLX	putative oligopeptide ABC transporter
237	FUMARATE-REDUCTASE	fumarate reductase
238	ABC-3-CPLX	lysine / arginine / ornithine ABC transporter

239	ABC-52-CPLX	putative transport complex, ABC superfamily
240	ABC-51-CPLX	putative transport complex, ABC superfamily
241	FORMHYDROG2-CPLX	hydrogenase 2
242	ABC-13-CPLX	glutamate / aspartate ABC transporter
243	ABC-46-CPLX	galactofuranose ABC transporter
244	ATPASE-1-CPLX	K+ transporting P-type ATPase
245	ABC-58-CPLX	Autoinducer-2 ABC transporter
246	ABC-40-CPLX	glycine betaine ABC transporter, non-osmoregulatory
247	ABC-9-CPLX	ferric citrate ABC transporter
248	TRANS-CPLX-202	multidrug efflux pump MdtABC-TolC
249	CPLX0-2201	The Tol-Pal Cell Envelope Complex
250	CPLX0-3361	NADH:quinone oxidoreductase I, peripheral arm
251	ABC-22-CPLX	oligopeptide ABC transporter
252	CPLX0-3970	murein tripeptide ABC transporter
253	CPLX0-7725	CRISPR-associated complex for antiviral defense
254	ABC-59-CPLX	putative D,D-dipeptide ABC transporter
255	CPLX0-7992	lipopolysaccharide transport system
256	CPLX0-2381	degradosome
257	ABC-20-CPLX	Ni(2+) ABC transporter
258	ABC-15-CPLX	branched chain amino acid / phenylalanine ABC transporter
259	ABC-304-CPLX	leucine / L-phenylalanine ABC transporter
260	ABC-8-CPLX	dipeptide ABC transporter
261	HYDROG3-CPLX	hydrogenase 3
262	ATPSYN-CPLX	ATP synthase / thiamin triphosphate synthase
263	FHLMULTI-CPLX	formate hydrogenlyase complex
264	CPLX0-3803	DNA polymerase III, holoenzyme
265	CPLX0-7451	flagellar export apparatus
266	CPLX0-250	hydrogenase 4
267	CPLX0-3933	Outer Membrane Protein Assembly Complex
268	NADH-DHI-CPLX	NADH:quinone oxidoreductase I
269	CPLX0-3382	Type II secretion system
270	FLAGELLAR-MOTOR-COMPLEX	flagellar motor complex
271	CPLX0-7452	flagellum

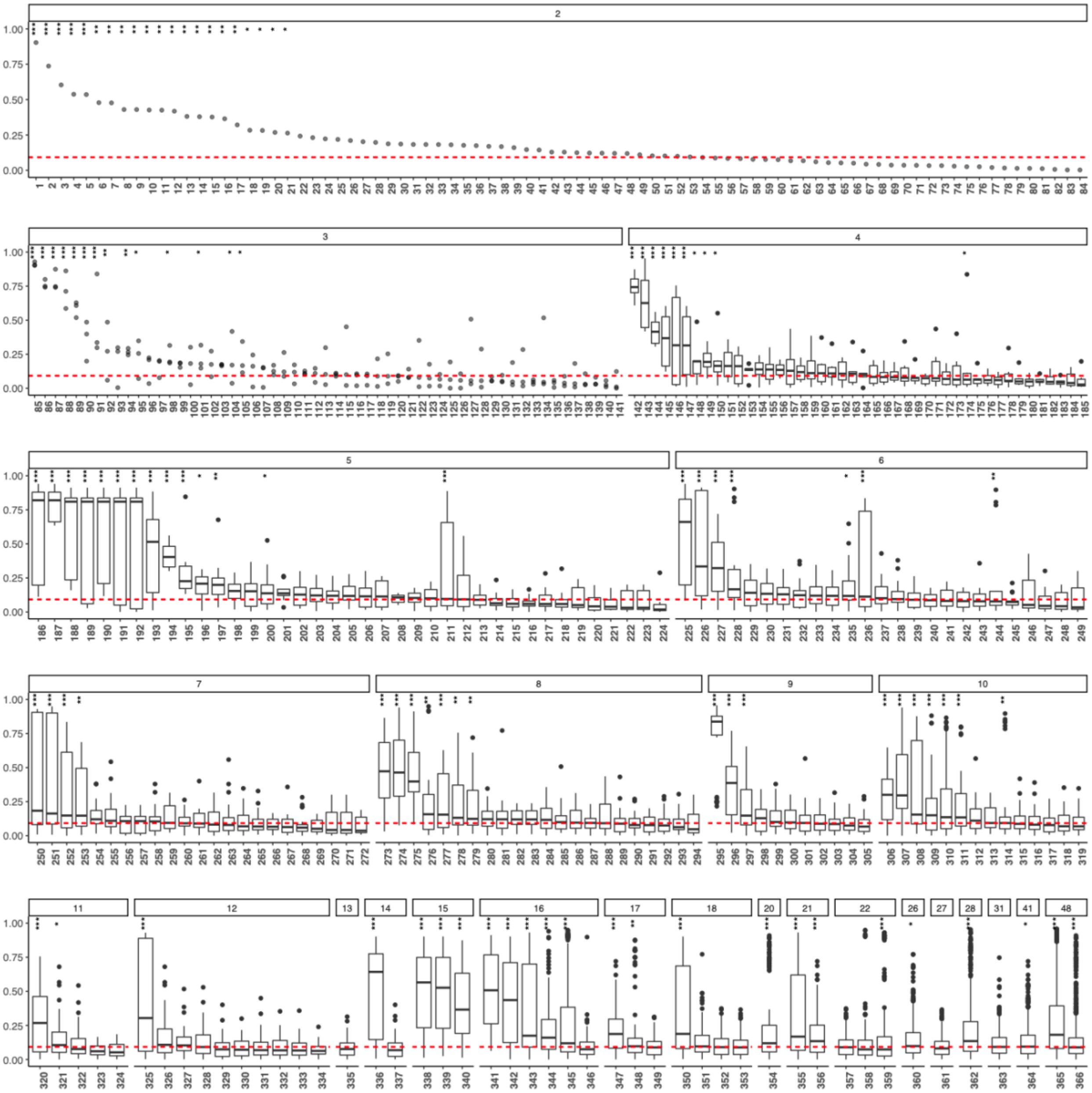

2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
21
22
26
27
28
31
41
48
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84
85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141
142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185
186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224
225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249
250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272
273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294
295 296 297 298 299 300 301 302 303 304 305
306 307 308 309 310 311 312 313 314 315 316 317 318 319
320 321 322 323 324
325 326 327 328 329 330 331 332 333 334
335
336 337
338 339 340
341 342 343 344 345 346
347 348 349
350 351 352 353
354
355 356
357 358 359
360
361
362
363
364
365 366

FIGURE 2.S1. Pairwise phenotypic profile similarity for genes in the same Ecocyc pathway. The distribution of phenotypic profile similarity values determined by |PCC| for all pairwise combinations of genes assigned to each of 366 EcoCyc pathways is shown. Profile similarity is plotted on the y-axis and the individual pathways are arrayed along the x-axis. The identity of each pathway is indicated by a numeric label that is defined in Table S1. The pathways are sorted by the number of genes in the pathway and then by the median |PCC| value. For pathways that have two or three members, the results are shown as scatter plots. For pathways with more than three genes, the results are shown as box plots with the outliers shown as black dots. The red dashed line shows the mean |PCC| value for all possible gene pairs. Asterisks indicate the permutation-based FDR-corrected p-values: * for $p<0.05$, ** for $p<0.01$, and *** for $p<0.001$.

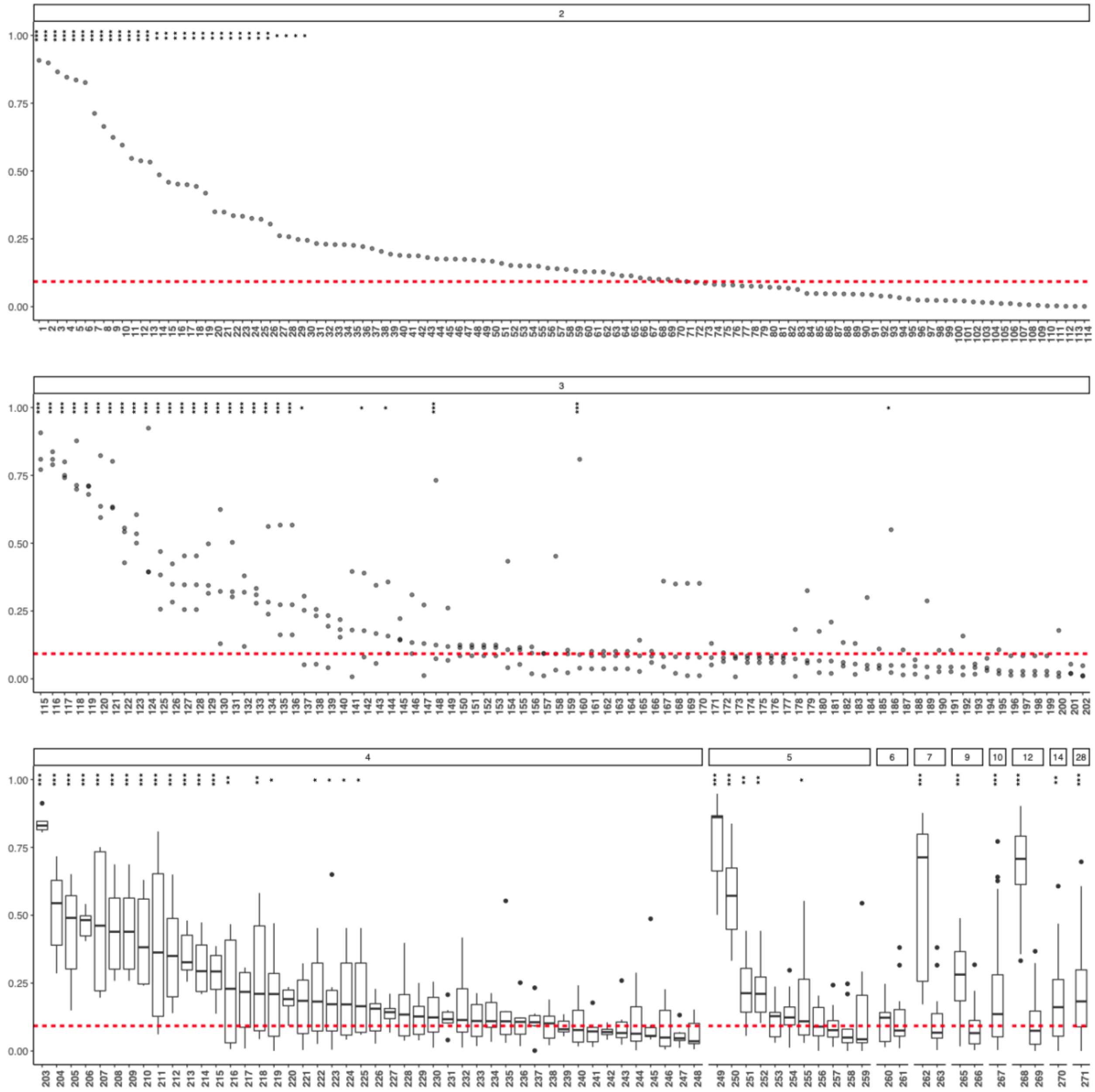

2
3
4
5
6
7
9
10
12
14
28
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114
115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202
203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248
249 250 251 252 253 254 255 256 257 258 259
260 261
262 263
265 266
267
268 269
270
271

FIGURE 2.S2. Phenotypic profile similarity for genes in the same EcoCyc heteromeric protein complex. The distribution of phenotypic profile similarity values determined by |PCC| for all pairwise combinations of genes assigned to each of 271 EcoCyc heteromeric protein complexes is shown. Profile similarity is plotted on the y-axis and the individual protein complexes are arrayed along the x-axis. The name of each protein complex is indicated by a numeric label that is defined in Table S2. The complexes are sorted by the number of genes that encode the complex and then by the median |PCC| value. For protein complexes that have two or three members, the results are shown as scatter plots. For protein complexes with more than three genes, the results are shown as box plots with the outliers shown as black dots. The red dashed line shows the mean |PCC| value for all possible gene pairs. Asterisks indicate the permutation-based FDR-corrected p-values: * for p<0.05, ** for p<0.01, and *** for p<0.001.

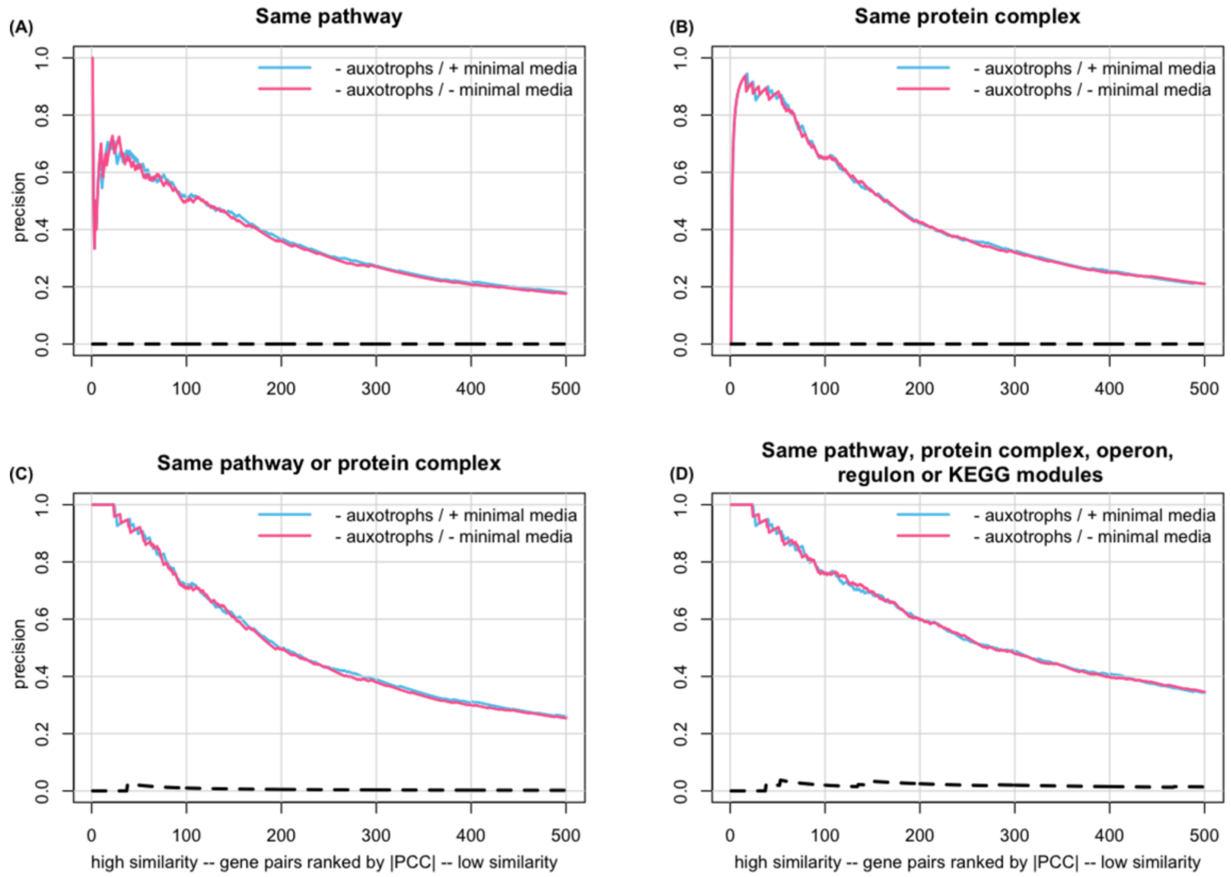


FIGURE 2.S3. Precision for co-annotated gene pairs when auxotrophic mutants were excluded. Gene pairs were ranked from high to low similarity based on |PCC| after strains with an auxotrophic phenotype were excluded (only the first 500 gene pairs are shown). Precision was calculated using: (A) gene pairs co-annotated to the same EcoCyc pathway, (B) gene pairs co-annotated to the same heteromeric protein complex, (C) gene pairs co-annotated to either the same pathway or the same protein complex, and (D) gene pairs co-annotated to the union of annotation sets 1 to 5 (EcoCyc pathways, heteromeric protein complexes, operon, regulon, or KEGG module). In each panel, the blue line shows precision for all growth conditions, and the red line shows precision when growth conditions involving minimal media were excluded. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control).

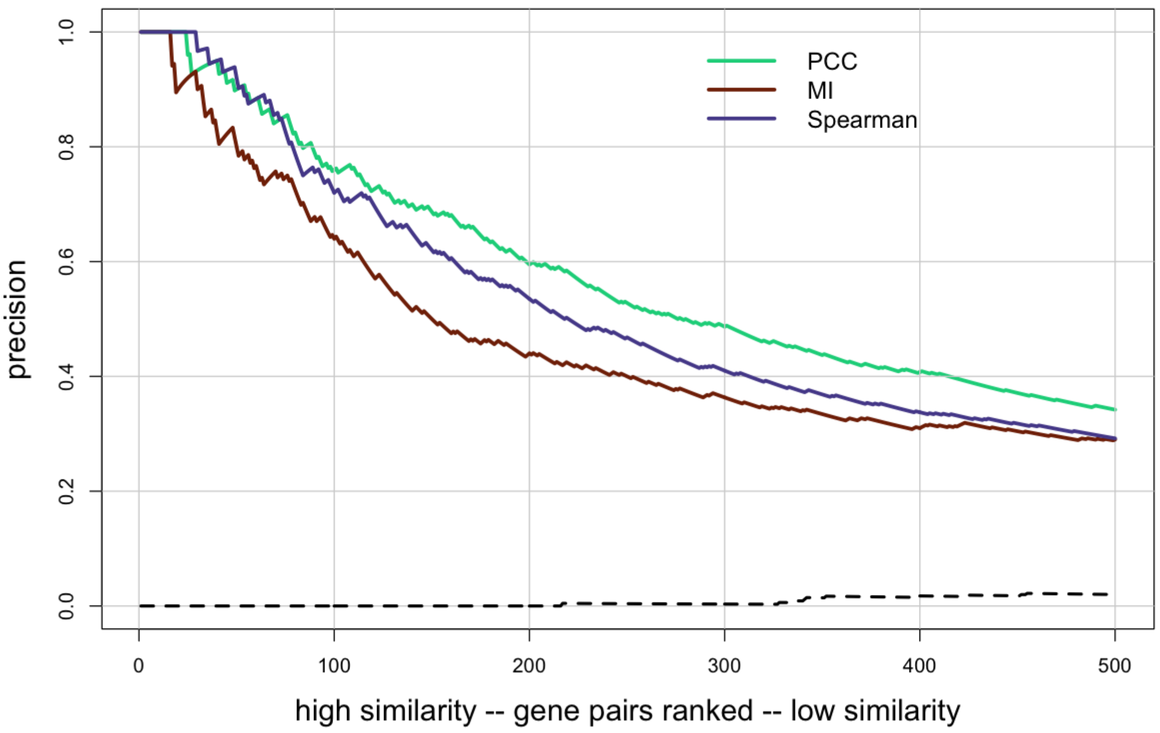


FIGURE 2.S4. Precision values determined by the three different similarity metrics were similar when auxotrophic mutants were excluded. Gene pairs were ranked from high to low similarity based on either |PCC| (green line), MI (brown line), or |SRCC| (blue line) determined after strains with a known auxotrophic phenotype were excluded and plotted versus precision, using the union of annotation sets 1 through 5 to identify co-annotated gene pairs. Only the first 500 gene pairs are shown. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control).

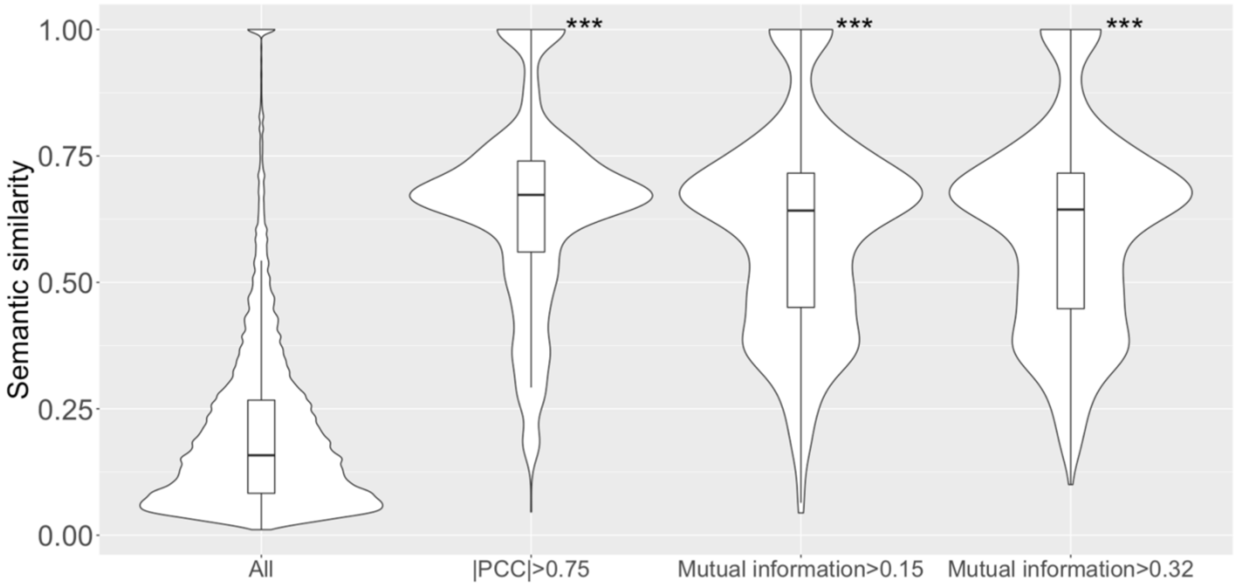


FIGURE 2.S5. Higher semantic similarity and phenotypic profile similarity were still found when GO biological process annotations inferred from electronic annotation (IEA) were excluded. Violin plots of the distribution of semantic similarity for, from left to right: all gene pairs annotated with GO biological process term(s); the subset of gene pairs with |PCC| >0.75; the subset of gene pairs with MI >0.15 (calculated based on qualitative fitness scores for all growth conditions); and MI >0.32 (calculated based on qualitative fitness scores for the collapsed set of growth conditions). The cutoffs of MI >0.15 for the third violin plot and MI >0.32 for the fourth violin plot were chosen so that all three subsets of gene pairs would contain the same number (~1,000) of top-ranked gene pairs. ***: p-value <0.001 was determined by 1-sided Mann-Whitney U test, compared to all gene pairs.

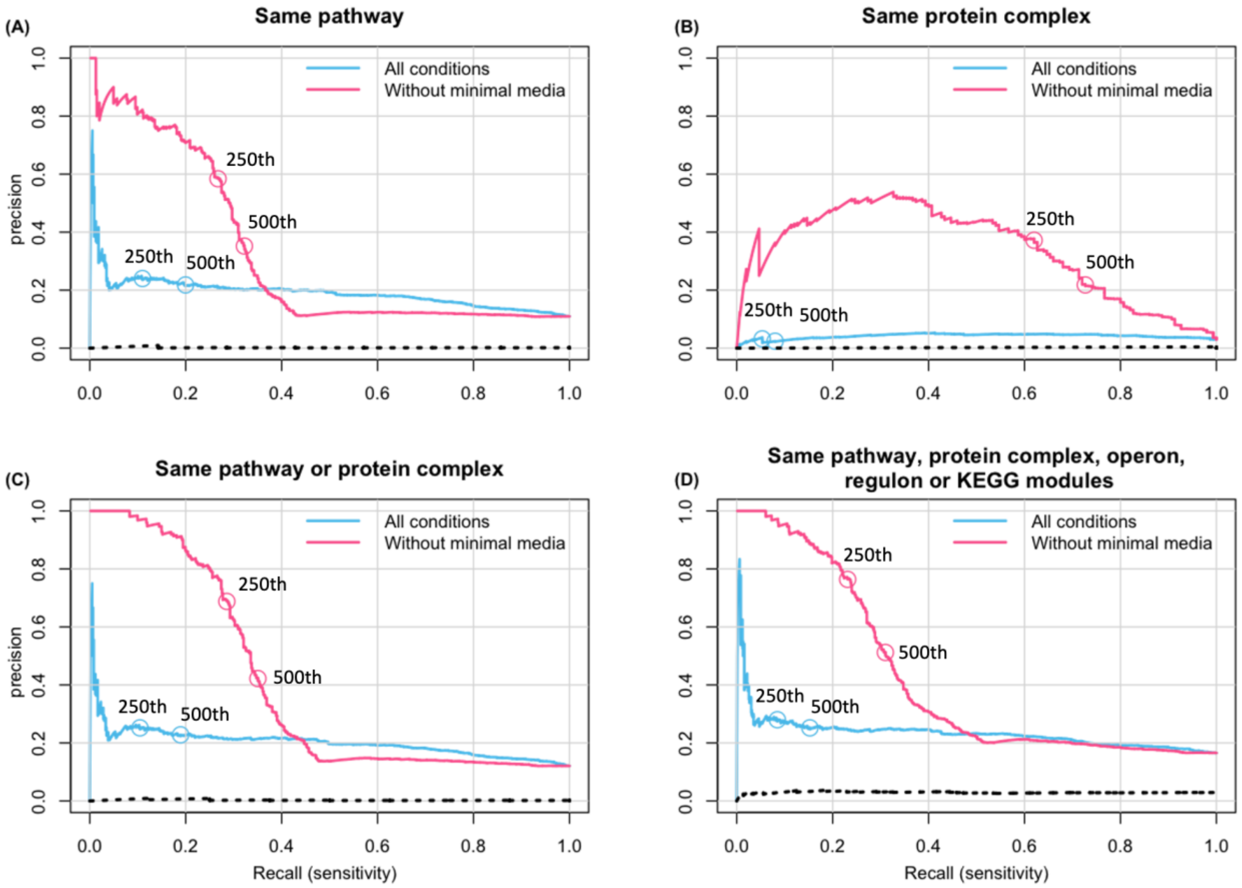


FIGURE 2.S6. Precision-recall curves when phenotypic profile similarity is determined by |PCC|.

Gene pairs were ranked from high to low similarity based on |PCC| using either all growth conditions (blue line) or after minimal media conditions were excluded (red line). Precision and recall were then calculated for the 5,000 top-ranked gene pairs. The panels show precision-recall curves for gene pairs annotated to either (A) the same EcoCyc pathway, (B) the same heteromeric protein complex, (C) the same pathway or complex, or (D) gene pairs co-annotated in any of annotation sets 1 through 5. Circles indicate the positions of the 250th and 500th top-ranked gene pairs. The dashed lines show precision for randomly ordered gene pairs generated as described in the Methods (negative control). The areas under the curve for randomly ordered gene pairs, for all growth conditions, and when minimal media conditions are excluded are: (A) 0.11, 0.19, and 0.50, respectively; (B) 0.03, 0.04, and 0.31, respectively; (C), 0.12, 0.20, and 0.57, respectively; and (D) 0.16, 0.24, and 0.47, respectively.

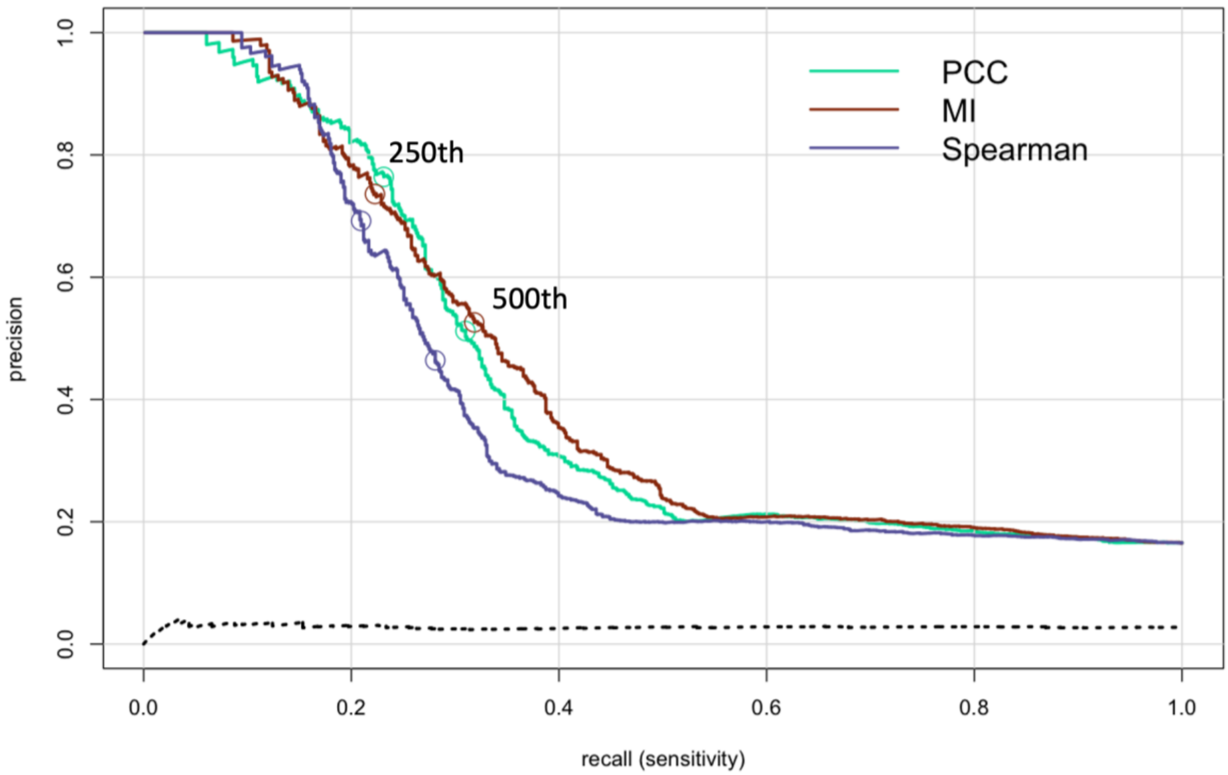


FIGURE 2.S7. Precision-recall curves when phenotypic profile similarity is determined using different metrics. Gene pairs were ranked from high to low similarity based on |PCC| (green line), MI (brown line), or |SRCC| (blue line). Precision and recall were then calculated for the 5,000 top-ranked gene pairs using the union of annotation sets 1 through 5 to identify co-annotated gene pairs. Circles indicate the positions of the 250th and 500th top-ranked gene pairs. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control). The areas under the curve (AUC) when pairwise profile similarity is determined using |PCC|, MI, |SRCC| or for randomly ordered gene pairs are: 0.42, 0.43, 0.39, and 0.17, respectively.

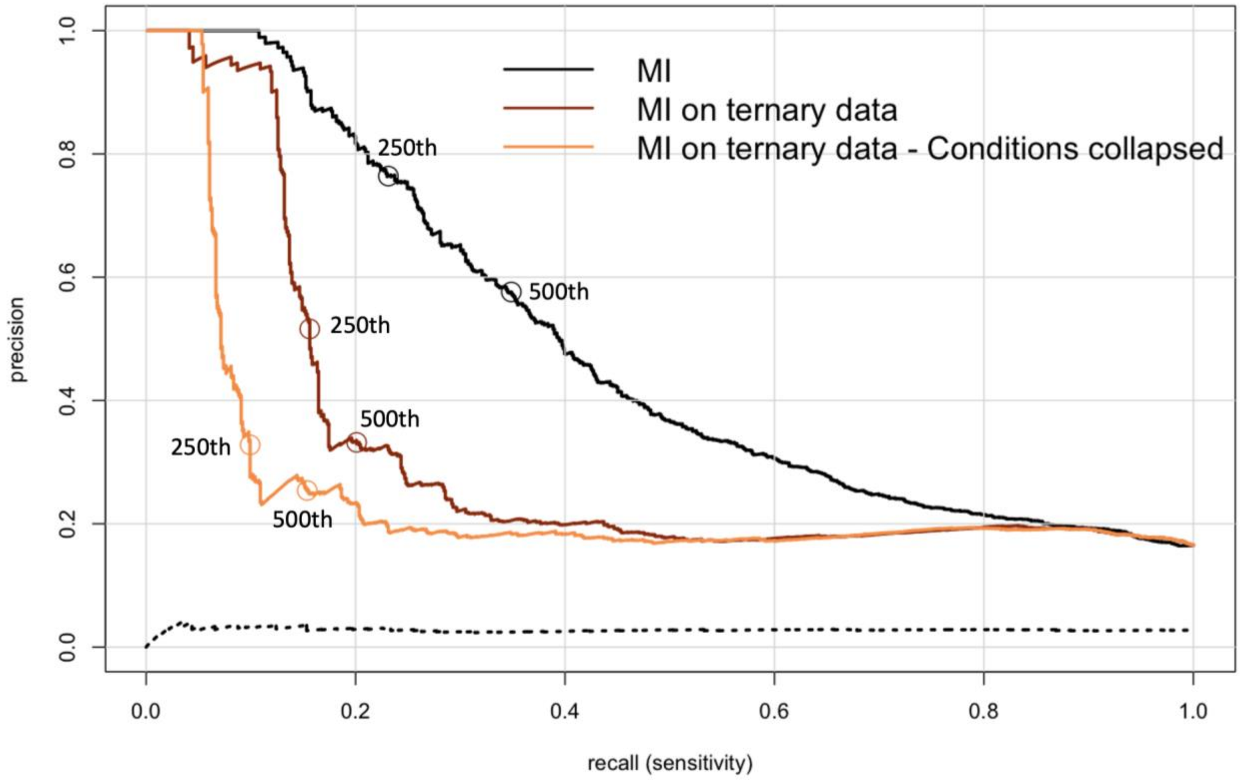


FIGURE 2.S8. Precision-recall curves for quantitative versus discretized, ternary fitness scores. Gene pairs were ranked from high to low similarity based on MI. Precision and recall were then calculated for the 5,000 top-ranked gene pairs using the union of annotation sets 1 through 5 to identify co-annotated gene pairs. Circles indicate the positions of the 250th and 500th top-ranked gene pairs. The phenotypic profiles contained either the original quantitative fitness scores (black line), the discretized, ternary scores for all growth conditions (brown line), or the discretized, ternary scores for growth conditions collapsed to 114 unique stresses (orange line). The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control). Areas under the curve (AUC) when MI was determined using either quantitative fitness scores; discretized, ternary scores for all growth conditions; or discretized, ternary scores for the collapsed set of growth conditions were 0.48, 0.32, 0.25, and 0.17, respectively.

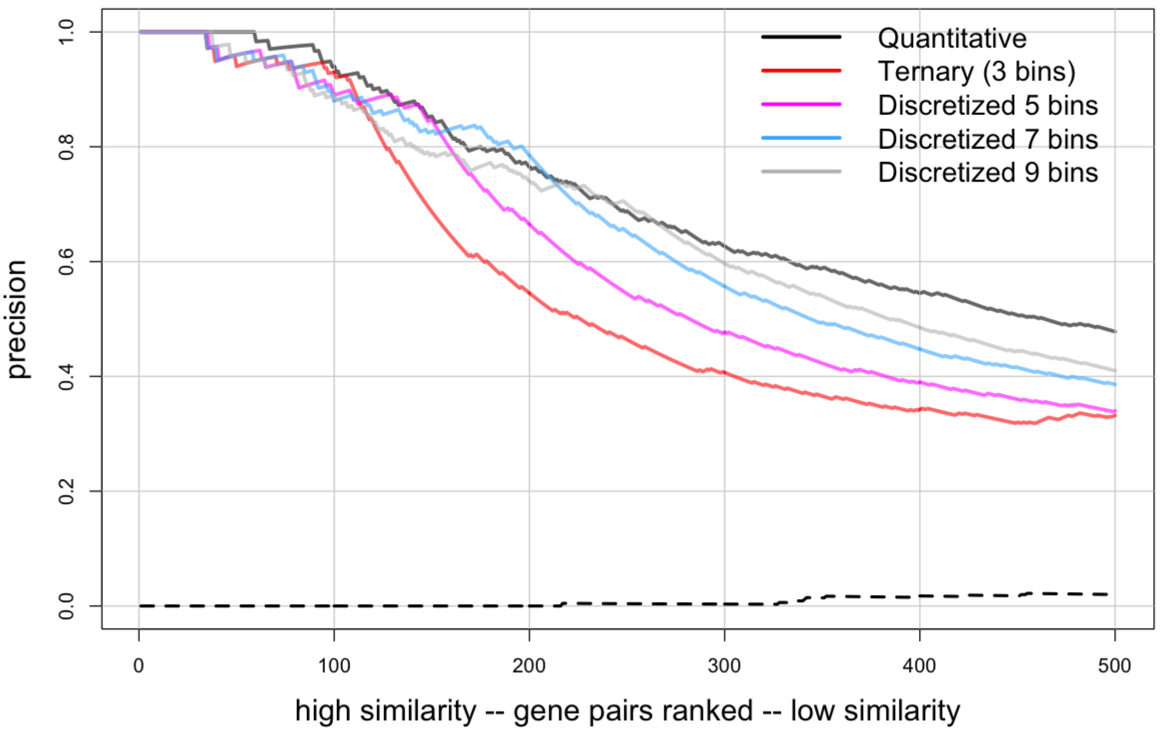


FIGURE 2.S9. Precision increased when quantitative fitness scores were partitioned into larger numbers of bins. Gene pairs were ranked from high to low similarity based on MI. Precision was then calculated using the union of annotation sets 1 through 5 to identify co-annotated gene pairs. Only the first 500 gene pairs are shown. The conversion of the quantitative fitness scores into discretized scores were based on the false discovery rates: 5% FDR for 3 bins; 5% and 10% FDR for 5 bins; 5%, 10%, and 15% FDR for 7 bins; and 5%, 10%, 15%, and 20% FDR for 9 bins. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control).

CHAPTER 3. SYSTEMATIC REANALYSIS OF MULTIPLE HIGH-THROUGHPUT *E. COLI* PHENOTYPIC DATASETS REVEALS FUNCTIONAL CONNECTIONS BETWEEN GENES

ABSTRACT

Phenotypes play a crucial role in understanding the functions of genes. The ability to infer new functions significantly increases when large numbers of phenotypes can be experimentally captured and systematically analyzed. Here, we report the results of systematically reanalyzing three published, high-throughput phenotypic datasets with the help of several functional annotation sets. We found that using a guilt-by-association approach we have published previously on one of the datasets leads to the same conclusion: phenotypic profile similarity strongly associates with functional similarity. When the phenotypes from two of the three studies were combined, associations between phenotypes and gene function were still observed but not improved compared to using single datasets. In addition, we have annotated the phenotypes from the three datasets using the Ontology of Microbial Phenotypes and done a preliminary analysis for pairwise semantic similarity that shows that in the long run, OMP can be used to make microbial phenotype data interoperable.

INTRODUCTION

The combination of forward genetics, biochemical and molecular biological approaches have led to identification of the function of many genes in bacterial genomes. But even for a well-studied bacterium such as Escherichia coli K-12, the function of 35% of its genes remains unknown (Ghatak et al.,2019). Reverse genetics with high-throughput phenotypic screens are an alternate approach for finding clues to the function of these orphan genes, which are also known as y-genes. There are many such studies that aim to generate phenotypes in large quantities in order to relate phenotypes to gene functions. These pioneering research projects discovered some functions of genes using the "guilt by association" approach. We have performed a systematic and unbiased reanalysis of a high-throughput E. coli phenotypic dataset (P. I.-F. Wu et al., 2021), and shown that high phenotypic similarity strongly correlates with functional similarity. With the expectation that the ability of phenotype data to predict gene function will increase with the amount of phenotype information available, we decided to perform a similar systematic analysis of other high-throughput datasets. The existing high-throughput phenotypic studies for *E. coli* (Campos et al., 2018; Fuhrer et al., 2017; Mutalik et al., 2020; Nichols et al., 2011; Price et al., 2018; Rishi et al., 2020; Shiver et al., 2020; Tong et al., 2020) can potentially provide insights into the functions of these genes.

In this research, we have conducted systematic reanalysis of a high-throughput phenotypic profile dataset (Price et al., 2018), the reanalysis after combining two datasets (Nichols et al., 2011; Price et al., 2018) and three datasets (Campos et al., 2018; Nichols et al., 2011; Price et al., 2018). We found that many existing functional annotations correlate well with these phenotype data, suggesting that testing hundreds of thousands of phenotypes in parallel can serve as a strong indicator for gene functions. In addition, we have highlighted the importance of using an ontology (Chibucos et al., 2014) to systematically curate microbial phenotypes (Siegele et al., 2019).

MATERIALS AND METHODS

Functional annotations used

E. coli functional annotations were downloaded from various sources: pathway, protein complex and operon annotations were downloaded from EcoCyc (Keseler et al., 2017). KEGG module annotations were downloaded from Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2016). Regulon annotations were from RegulonDB (Gama-Castro et al., 2016). Protein-protein interaction annotations were from STRING (Szklarczyk et al., 2015).

Data preprocessing

We preprocessed the data from Price *et al.* (Price et al., 2018) before calculating the phenotypic similarity scores: fitness scores were averaged for growth conditions where results from multiple experiments were included in the dataset. For combining phenotypic profile datasets of Nichols *et al.* (Nichols et al., 2011) and Price *et al.* (Price et al., 2018), only genes present in both studies were used.

Software and statistics

Statistical software R with Rstudio IDE and Jupyter Notebook loaded with IRkernnel were used to perform all analyses. All the source code is deposited in https://github.com/peterwu19881230/analyze_multiple_ecoli_phenotypic_profiles. Phenotypic profile similarities for OMP-based phenotype annotations were calculated using the Lin method with the best-match-average (BMA) strategy, as implemented by Greene *et al.* (Greene et al., 2017).

RESULTS

Reanalysis of a high-throughput *E. coli* phenotypic dataset reveals functional associations

We previously described the reanalysis of a high-throughput phenotypic profile dataset where fitness scores were based on imaging colony sizes on agar plates (Nichols et al., 2011). We showed in a systematic way across the entire dataset that phenotypic profile similarity significantly correlates with shared functional annotations. We wondered whether comparable results would be found if other phenotypic datasets were analyzed in the same way. We picked the phenotypic profile dataset from Price *et al*. (Price et al., 2018) to analyze because different experimental methods were used. Price *et al*. performed competitive fitness assays in liquid culture (Wetmore et al., 2015) for pools of mini-Tn5 insertion mutants under a variety of growth conditions for four to eight population doublings. In contrast, Nichols et *al*. (Nichols et al., 2011) determined fitness scores for each mutant strain individually on solid medium. Growth of each strain was assayed after a time equal to >15 doublings. The comparison of their methods is in Table 3.1.

To assess the association between phenotypes and functions, we calculated the pairwise phenotypic similarity for all mutants using the Pearson Correlation Coefficient (PCC) and then compared the functional annotations associated with each gene in a pair. We used high-quality functional annotations, the majority of which were manually curated and based on experimental results. The annotation sets included annotations to metabolic pathways, heteromeric protein complexes, operons, regulons, KEGG modules and protein-protein interactions (Gama-Castro et al., 2016; Kanehisa et al., 2016; Keseler et al., 2017; Keseler et al., 2014; Szklarczyk et al., 2015; The Gene Ontology Consortium, 2017). Figure 3.1a shows the distribution of phenotypic similarity values for all possible gene pairs and for sets of gene pairs that share the same annotations (co-annotated gene pairs). Not surprisingly, for each of the functional annotation sets, except for regulons (data not shown), co-annotated gene pairs had, on average, significantly higher phenotypic similarity. When we examined the profile similarity of gene pairs that are co-annotated in both pathways and heteromeric protein complexes or co-annotated in all six annotation sets, even greater enrichment for phenotypic similarity was seen (Figure 3.1b).

To assess whether gene pairs that are more phenotypically similar are more likely to share functions, we first ranked gene pairs based on phenotypic similarity and then define determined the fraction of co-annotated gene pairs among all gene pairs with a phenotypic similarity score above a specific cutoff. This fraction represents precision: the fraction of results that a test identifies as positive that represent true positives [TP/(TP+FP)]. For example, if there are 10 gene pairs whose phenotypic similarity is >0.90, and 6 of the gene pairs are co-annotated, the precision is 6/10=0.6. Expecting phenotypes to be associated with more than one set of functions, we calculated precision for ranked gene pairs using the six functional annotation sets either singly or in combination. The graph in Figure 3.2 shows the relationship between precision and phenotypic similarity for gene pairs that are co-annotated to the same pathway, to the same heteromeric protein complex, to either the same pathway or same protein complex (the union of the two annotation sets), or are co-annotated in any of the annotation sets (the union of all six annotation sets used). As expected, enriched precision was seen for all sets of co-annotated gene pairs, except for gene pairs co-annotated to the same protein complex. The highest precision was seen for the union of all the annotation sets indicating that phenotypic profiles associate with multiple categories of functions.

Combining phenotypic datasets did not increase the association between phenotypes and functions

For mutant genes in the two high-throughput datasets we analyzed (Nichols et al., 2011; Price et al., 2018), we combined the results from the two studies into single phenotypic profile dataset of shared 3,527 *E. coli* genes, which is expected to be more informative in terms of associating phenotypes with functions. When pairwise phenotypic similarity was calculated using PCC, co-annotated gene pairs were enriched for higher phenotypic similarity compared to all gene pairs (Figure 3.3a). Gene pairs that are co-annotated in both pathways and complexes or co-annotated in all six annotation sets, had even greater enrichment for phenotypic profile similarity (Figure 3.3b). However, when these distributions were compared to those shown in Figure 3.1 for the single data set, there was no consistent increase in the mean phenotypic similarity of co-annotated gene pairs.

To assess the association between phenotypic profile similarity and function after combining the two datasets, we used the precision versus ranking approach described above. Gene pairs that share co-annotations in each single annotation set, except for protein complex annotations, or that share co-annotations in more than one annotation set show enriched precision compared to random expectation (Figure 3.4a). The highest precision was seen for gene pairs that share co-annotations from any of the annotation sets (the union of all the annotation sets). The precision curve for gene pairs that share pathway annotations isn't visible in the figure because it overlaps with the curve for gene pairs that share both pathway and protein complex annotations. The dotted line indicates random expectation.

When we compared the precision curves from the combined dataset with the precision curves from each single dataset, we unexpectedly found that combining the phenotypes did not significantly increase the precision of co-annotated gene pairs. Figure 3.4b shows the comparison for gene pairs co-annotated to the union of the six annotation sets. No significant increase in precision was seen when phenotypic profile similarity was determined using either Mutual Information (MI) or Spearman’s Rank Correlation Coefficient (SRCC) (data not shown).

Prediction for functions is more accurate when more functional annotations are available

It is worth noting that although the functional annotations used in this study are of high quality, they do not yet capture all available published information. Moreover, our understanding of gene functions is still incomplete, although it is expected to increase with time. Nevertheless, we hypothesize that the power of phenotypes to predict functions will improve as the functions of more genes are experimentally determined and thus, more functional annotations are made. To test this idea, we removed some annotations from the six functional annotation sets and then reassessed the association between phenotypes and functions for the combined dataset of Nichols *et al*. (Nichols et al., 2011) and Price *et al*. (Price et al., 2018) using PCC to assess profile similarity. There was no significant change in the average phenotypic similarity for co-annotated gene pairs when annotations were removed (Figure 3.5a). This result is reasonable because removing annotations shouldn't decrease the phenotypic similarity of gene pairs. However, when assessing the enrichment for phenotypic similarity between co-annotated gene pairs, the precision (fraction of co-annotated pairs above a similarity cutoff) significantly dropped when increasing number of annotations were taken out, either when a single annotation set, such as pathways, was used (Figure 3.5b) or when all six annotation sets were used (Figure 3.5c). This indicates that as additional functional annotations are made, precision should significantly increase. The increased precision should strengthen the hypothesis that systematically collecting phenotypes under many conditions can very well explain functional connections between gene pairs.

Preliminary results indicate that Ontology of Microbial Phenotypes can help probe the functions of genes

The Ontology of Microbial Phenotypes (OMP) (Chibucos et al., 2014) was developed to capture phenotype information from different microorganisms using a common vocabulary. As a formal ontology, the terms in the ontology are connected by logical relationships generating a hierarchical structure that forms a directed-acyclic graph (DAG). We wondered if the association between phenotypic similarity and functional similarity would still occur, if phenotypic similarity was based on the semantic similarity of phenotype annotations instead of the phenotypes observed in the original high-throughput experiments. To test this hypothesis, we used OMP to make annotations for the statistically significant phenotypes from three high-throughput studies: Nichols *et al*. (Nichols et al., 2011), Price *et al*. (Price et al., 2018) and Campos *et al*. (Campos et al., 2018). We then calculated the semantic similarity of the OMP annotations for pairwise combinations of genes present in all three studies. Gene pairs that shared the same functional annotation(s) had enhanced phenotypic similarity (Figure 3.6a and 3.6b), compared to all gene pairs. However, when ranked gene pairs were used to calculate precision no increase in precision relative to random expectation was seen regardless of which annotation set was used to identify co-annotated gene pairs (data not shown). This is probably due to loss of phenotype information that would contribute to an accurate estimate of phenotypic profile similarity that occurred when the annotations were made using OMP. The fitness scores that were not statistically significant were ignored. Presumably, when more OMP-based phenotype annotations become available, stronger association between phenotypes and functions will be seen.

In addition to the online data browser that was made for the data from Nichols *et al.* (P. I.-F. Wu et al., 2021), we have made data browsers for the results from Price *et al.* and Campos et al. All of them allow browsing information for strains, conditions and pairwise phenotypic similarities. The data browsers for Price *et al.* and Campos *et al.* are available at: https://microbialphenotypes.org/wiki/index.php?title=Special:Ecolispecialpage_price, and the results from Campos *et al*. at: https://microbialphenotypes.org/wiki/index.php?title=Special:CamposSpecialpage, respectively.

Since PCC is not the only useful metric for determining phenotypic profile similarity, for the data from Price *et al*. we also calculated similarity based on Spearman's rank correlation coefficient and Mutual Information. Mutual Information was used to determine profile similarity not only for the original quantitative fitness scores, but also after the quantitative scores were discretized to ternary values (-1, 0, +1) using all conditions, and for discretized ternary scores using only the unique growth conditions (referred to as collapsed conditions). The ability to sort by different phenotypic similarity metrics will be useful since additional similarity metrics might identify highly associated gene pairs that PCC-based similarity is not able to detect.

DISCUSSION

In this study, we re-analyzed a second high-throughput phenotypic profile dataset (Price et al., 2018) and found the co-occurrence of high phenotypic profile similarity and high functional similarity we had previously observed for a different phenotypic dataset (P. I. F. Wu et al., 2021). We next combined the phenotypic data from the two datasets (Nichols et al., 2011; Price et al., 2018) and repeated the analysis. In contrast to our original expectation, combining the datasets did not result in increased association between phenotypes and functions. It remains to be seen whether the same result will be seen for other combinations of datasets. The outcome may also depend on how the data are processed prior to being combined. Combining quantitative phenotypes may require rescaling or renormalizing the data. Combining qualitative phenotypes from different studies or combining qualitative or quantitative phenotype data presents additional challenges. Alternatively, it is possible that replacing the specific observed phenotypes with phenotype annotations may make it easier to combine phenotype data from different studies. To test this hypothesis, we calculated phenotypic profile similarity using annotation of genes made using OMP, the Ontology of Microbial Phenotypes, to allow the integration of phenotype data from separate studies. Enrichment for phenotypic profile similarity was observed for co-annotated gene pairs.

To test the hypothesis that the observed association between phenotypes and functions will improve as more experiments are performed and more functional annotations are made, we repeated our analysis after removing annotations from the current annotation sets to simulate how having increasing numbers of annotations available affects precision. The results demonstrated that the ability to predict function from phenotypes was improved by having more annotations.

One aim of phenotypic profiling is to predict functions for genes of unknown function based on finding similarity to the phenotypic profiles of genes whose function is known. To identify this category of gene pairs, we used the list of genes of unknown function annotated by Ghatak *et* al. (Ghatak et al., 2019) and screened highly correlated gene pairs, which had been identified using the data from Price *et al*, Nichols *et al*, or the combined data set, for ones where at least one gene was a gene whose function is unknown. We highlight of these because they may be of potential interest for future experiments. One example is the gene pair *yajR* and *cyoD*, which had a high PCC = 0.92. The gene product of *cyoD* is a subunit of the cytochrome bo3 ubiquinol oxidase, the terminal component of the aerobic electron transport chain, which reduces O_2 to H_2O (Nakamura et al., 1997). The function of the *yajR* gene product is unknown, but it is predicted to encode a putative inner membrane transport protein (Jiang et al., 2013). Another example is the gene pair *maoP* and *hdfR*, which have PCC > 0.95 using the quantitative data from Price *et al*, and also have a high similarity of 0.82 when similarity was determined using OMP-based phenotype annotations (similarity based on OMP annotations ranges from 0 to 1). The *hdfR* gene encodes DNA-binding transcriptional dual regulator HdfR (Ko & Park, 2000), which is known to positively regulate transcription of *maoP*, which encodes the macrodomain Ori protein (Valens et al., 2016). In addition, the gene pair of *ybjM* and *puuP*, which were not highly correlated when quantitative data from Nichols *et al*. or Price *et al*. were used to determine similarity. However, the gene pair had an OMP similarity of 0.86. The current knowledge about *puuP* (Kurihara et al., 2005) is that it is a proton dependent putrescine transporter, while the only information about ybjM is that it encodes a putative inner membrane protein (Daley et al., 2005). To sum up, it may be worthwhile to further characterize potential functions based on these highly correlated pairs.

When the phenotypic similarity was calculated using similarity metrics like PCC from Price *et al.* (Price et al., 2018) and Nichols *et al.* (Nichols et al., 2011), simply incorporating all the phenotypes didn't improve the precision of identifying shared functions (Figure 3). One possible explanation for this is that the fitness scores in Price *et al.* (Price et al., 2018) were determined after fewer growth cycles than those in Nichols *et al.* (Nichols et al., 2011), which might result in different changes in fitness for many mutants. Indeed, when the two studies measured growth in the presence of the same chemicals, the correlation between the fitness scores (determined using PCC) is low (data not shown). It is possible that pre-selecting the conditions that associate most strongly with the functional annotations, or using some other similarity metrics might increase precision.

It is worth mentioning that analyzing the data from Campos *et al* (Campos et al., 2018) didn't show as strong a connection between phenotypes and functional annotations as those seen using the data from Price *et al.* (Price et al., 2018) or Nichols *et al.* (Nichols et al., 2011) (data not shown). However, since there are 324 conditions for phenotypes in Nichols *et al.* (Nichols et al., 2011), more than 100 conditions in Price *et al.* (Price et al., 2018) but only less than 30 phenotypes measured in Campos *et al* (Campos et al., 2018), it is possible that results from using phenotypes only from Campos *et al* (Campos et al., 2018) are simply due to not having enough variety of phenotypes as variables.

It is possible that some machine learning methods can be applied to give stronger and/or exact prediction of certain functions. Phenotype profiles from many conditions can be used as explanatory variables (predictors) for the input, while the labels of the mutually exclusive functional categories for genes can be used as the response variables (predicted result). As more high-throughput phenotype data and more functional annotations become available, along with expert level biocuration and rigorous biostatistics, we anticipate an increased rate of identifying new roles of genes.

Table 3.1. Comparison between the 3 phenotypic profile datasets used in this study

Studies	**Nichols et al., 2011**	**Price et al., 2018**	**Campos et al., 2018**
Mutant construct	Single gene deletions from Keio collection (Baba et al., 2006)	Insertions made using bar-coded transposon (Wetmore et al., 2015)	Single gene deletions from Keio collection (Baba et al., 2006)
Phenotype Assay	Colony size measurements	Growth inferred from relative abundance of barcoded-sequences	Single cell imaging supported by support vector machine classification
No. of generations in growth condition	>15	4-8	5-7
No. of phenotype observations	3,979 mutants X 324 conditions = 1,289,196	3,789 mutants X 162 conditions = 613,818	3,815 mutants X 30 phenotypic characteristics = 134,010
Method used to determine the significant phenotype observations	5% FDR	5% FDR and a t-like test statistic	s-score transformation (similar to z-score transformation)
No. of significant phenotype observations	15,833	27,225	4,415
No. of genes that have significant phenotypes	2,210	1,425	1,180

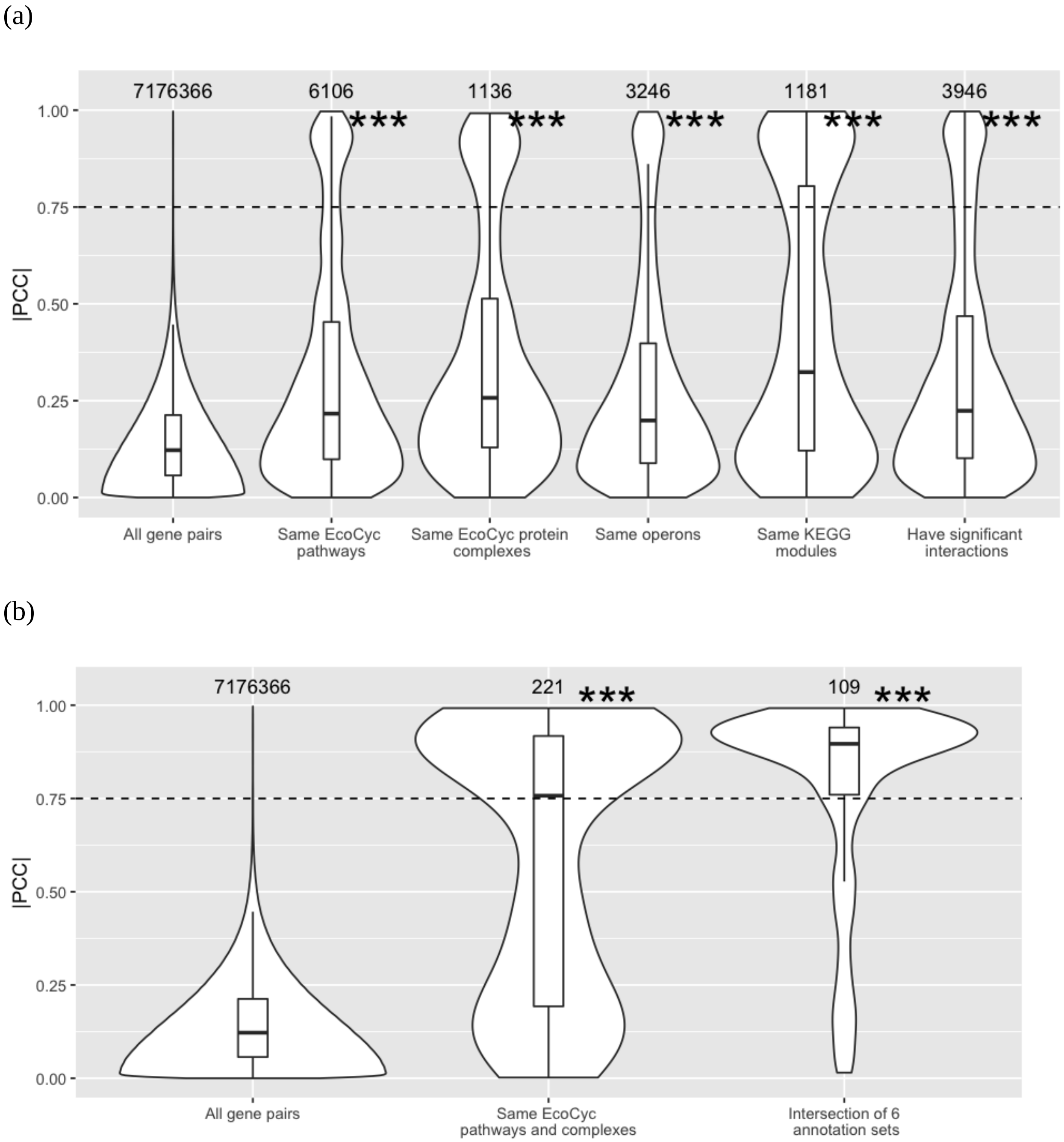


Figure 3.1. Distributions of phenotypic profile similarity for co-annotated gene pairs using fitness scores from Price *et al*.

(a) |PCC| from co-annotated pairs within each functional annotation set. The leftmost violin plot is the distribution of all pairwise |PCC| calculated from Price et al. Other violin plots represent the distribution of |PCC| from genes that are annotated to the same functional annotation(s). (b) |PCC| from co-annotated pairs within combinations of functional annotation set. ***: p value<0.001 based on one-sided Mann-Whitney U test.

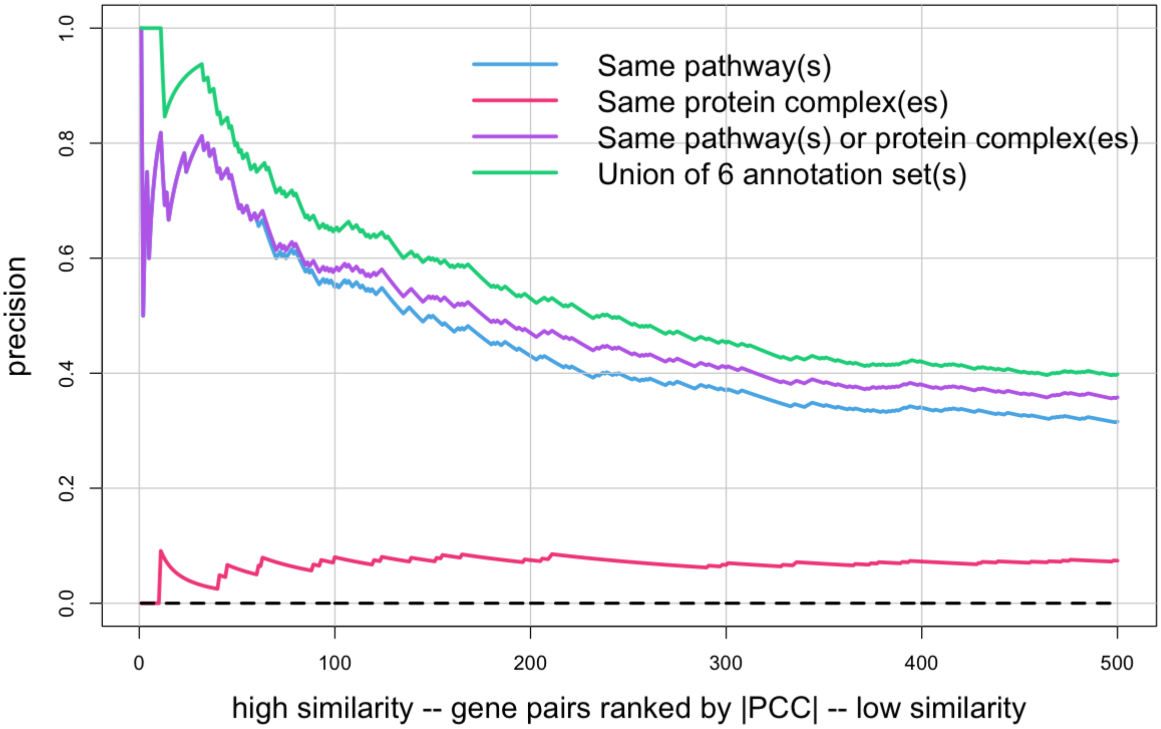


Figure 3.2. Ranking versus precision for pathways, protein complexes, pathways or protein complexes, and union of the 6 annotation sets.

Gene pairs were ranked from high to low similarity based on |PCC| values and plotted versus precision as described in the text. Only the first 500 gene pairs are shown. The different colored lines indicate either gene pairs that are annotated to the same EcoCyc pathway (blue), to the same heteromeric protein complex (pink), to either the same EcoCyc pathway or protein complex (purple), or are co-annotated in any of the following annotation sets: EcoCyc pathways, heteromeric protein complexes, operon, regulon, KEGG module or STRING interaction. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control).

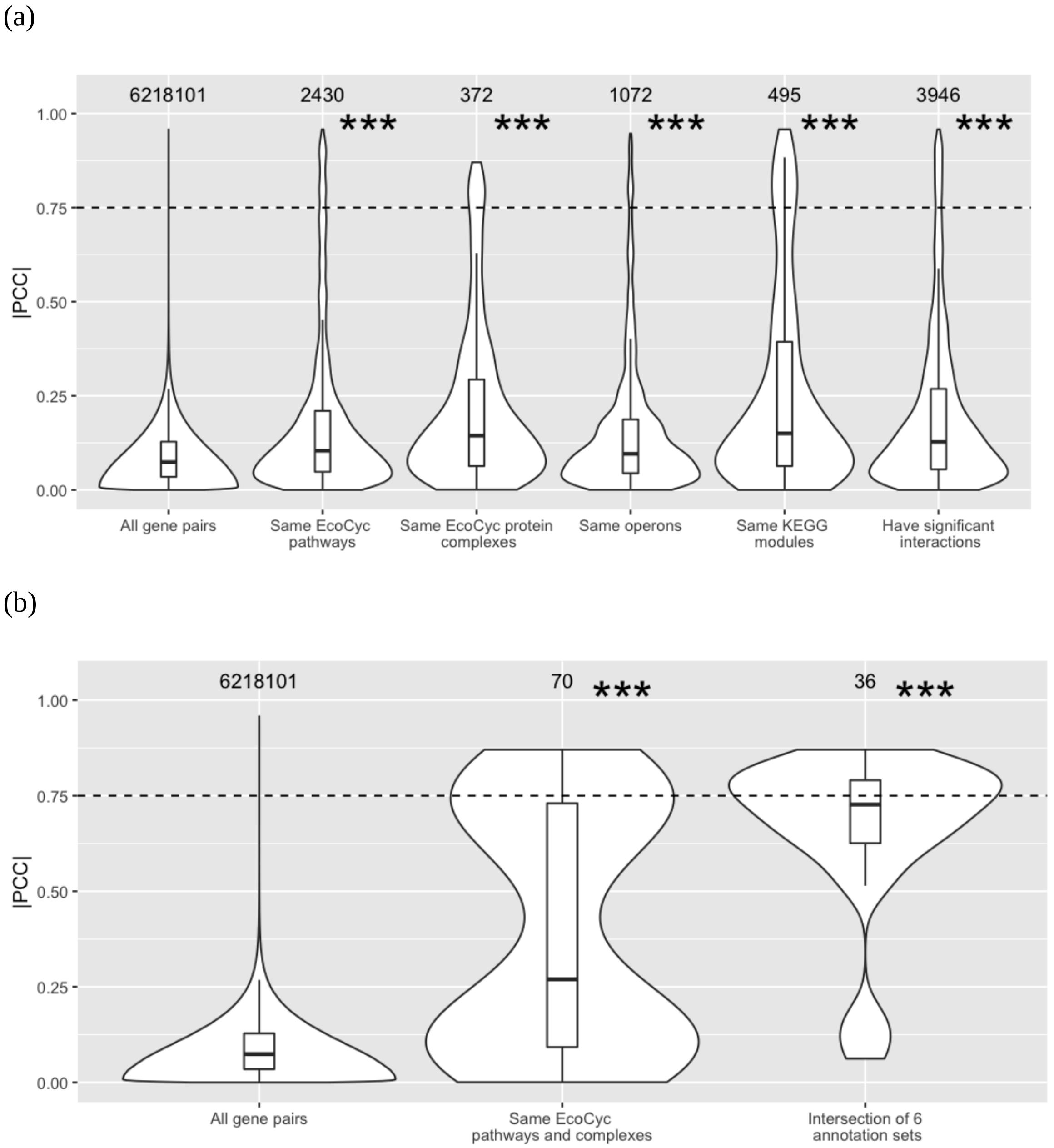
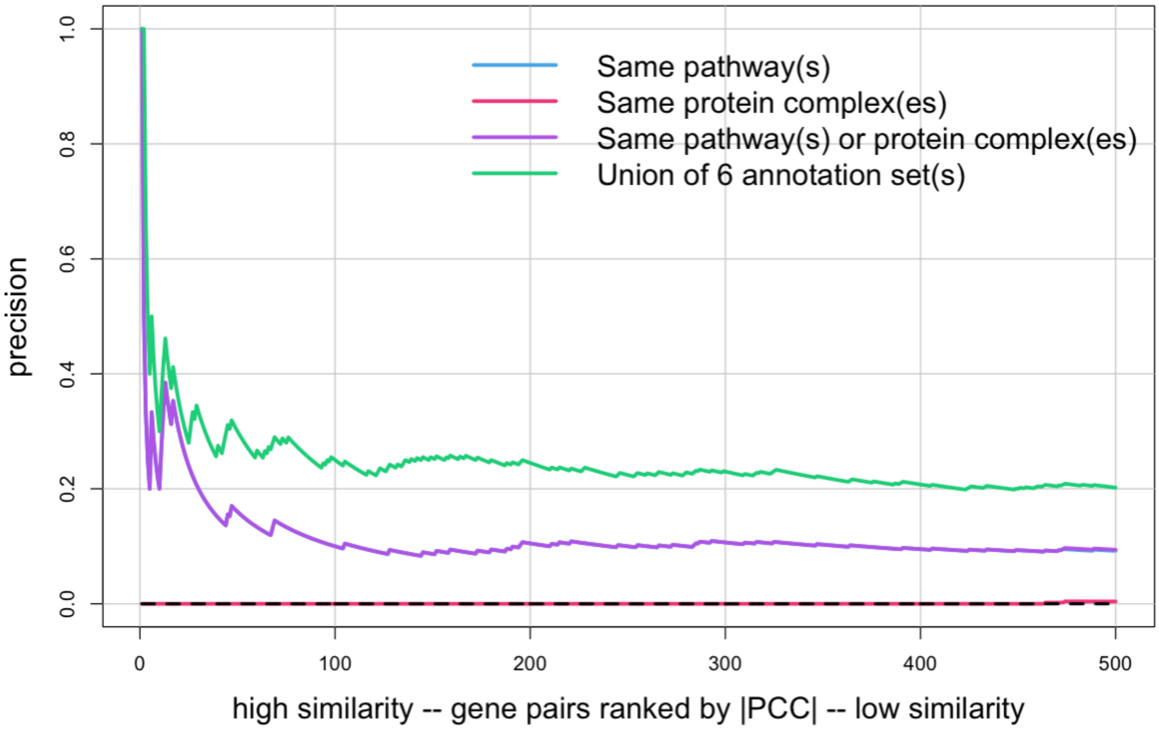


Figure 3.3 Distributions of phenotypic profile similarity for co-annotated gene pairs after combining fitness scores from two datasets (Nichols *et al*. and Price *et al*.)

Violin plots of the distributions of |PCC| values for, (a) from left to right, all possible gene pairs, gene pairs annotated to the same EcoCyc pathway, heteromeric protein complex, operon, regulon, KEGG module, or have STRING interaction. (b) from left to right, all possible gene pairs, gene pairs annotated to the same EcoCyc pathway and heteromeric protein complex, and gene pairs that to the same of EcoCyc pathways, heteromeric protein complexes, operon, regulon, KEGG module and have STRING interaction. Numbers above each violin plot indicate the number of gene pairs in each plot. ***: p-value <0.001 was determined by 1-sided Mann-Whitney U test, compared to all gene pairs. The dashed line indicates |PCC| = 0.75, which was chosen as an arbitrary cutoff.

(a)


Same pathway(s)
Same protein complex(es)
Same pathway(s) or protein complex(es)
Union of 6 annotation set(s)
precision
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
high similarity -- gene pairs ranked by |PCC| -- low similarity

(b)

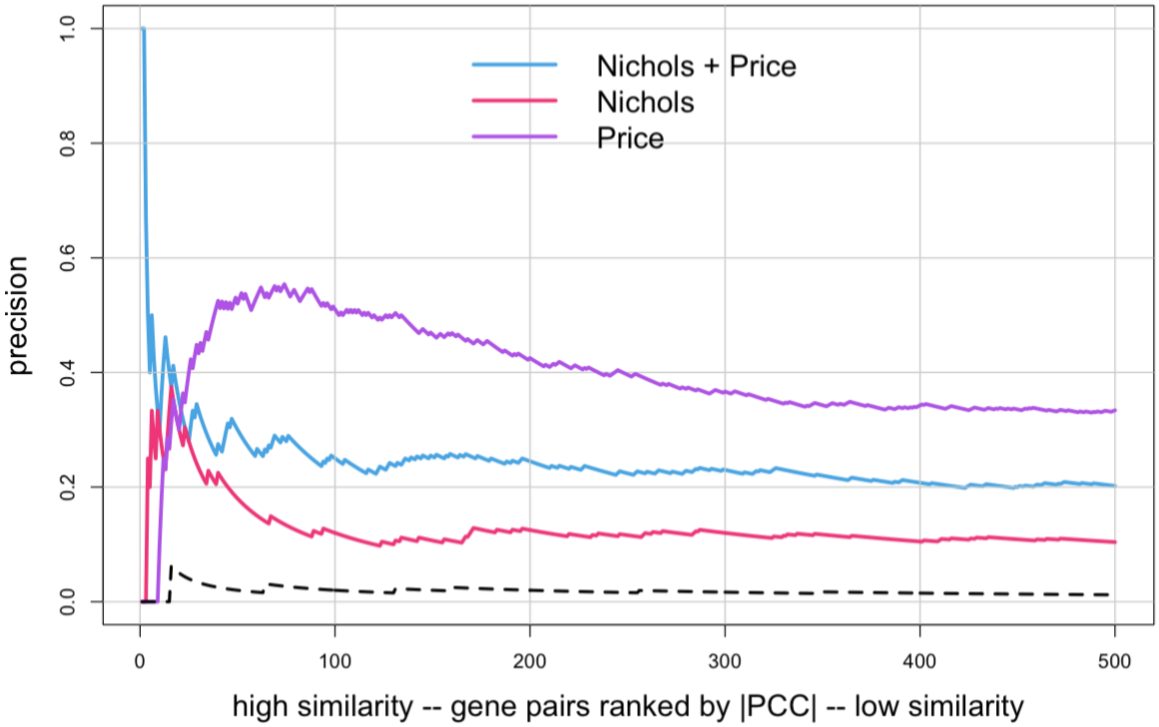
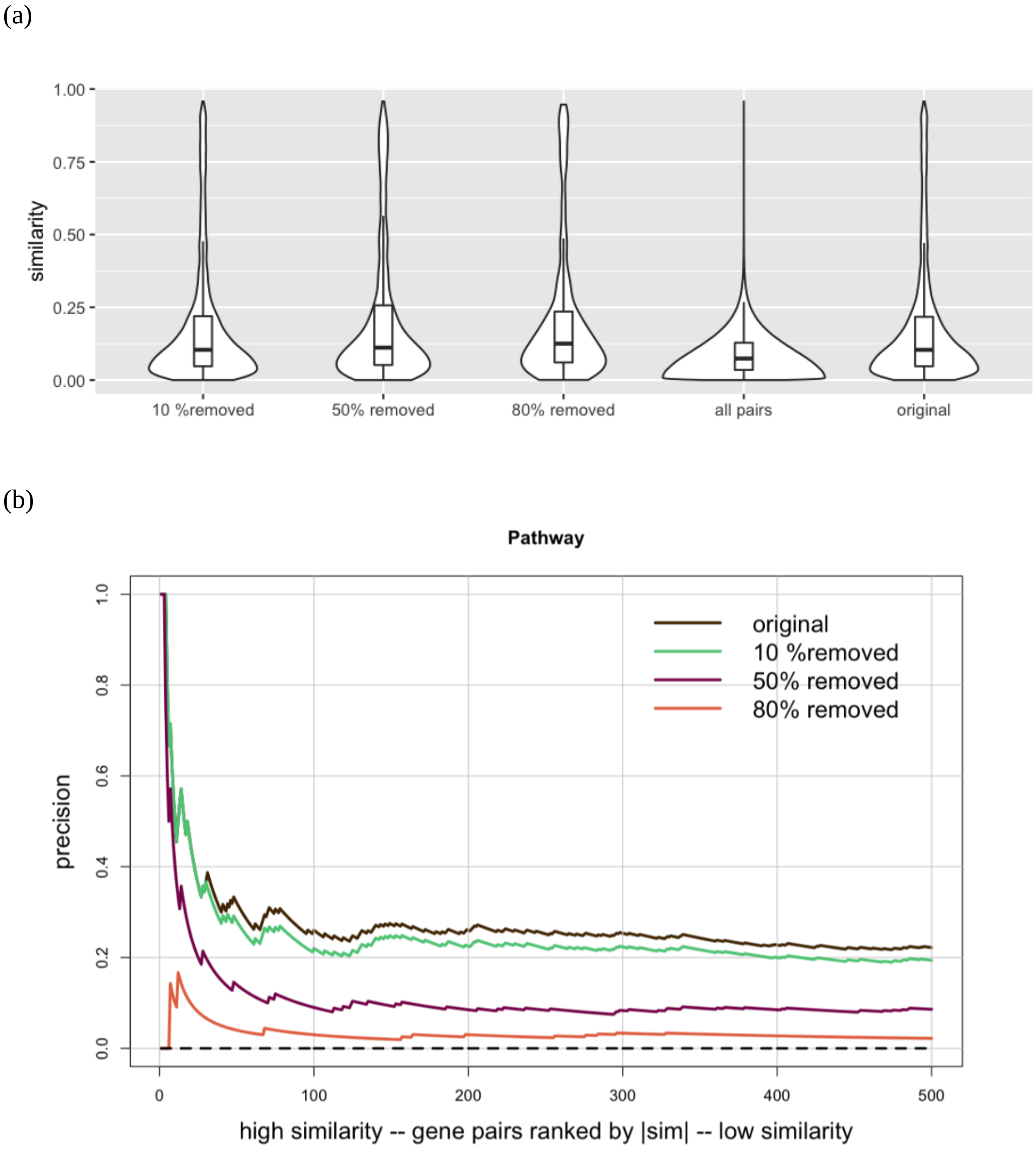

Nichols + Price
Nichols
Price
precision
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
high similarity -- gene pairs ranked by |PCC| -- low similarity

Figure 3.4. Precisions did not increase when phenotypic profiles from Price *et al*. and Nichols *et al*. were combined.

(a) Phenotypic profile similarity was calculated after combining fitness scores from Nichols *et al*. and Price *et al*. as described in the text. Gene pairs were ranked from high to low similarity based on |PCC| values and plotted versus precision as described in the text. Only the first 500 gene pairs are shown. The different colored lines indicate either gene pairs that are annotated to the same EcoCyc pathway (blue), to the same heteromeric protein complex (pink), to either the same EcoCyc pathway or protein complex (purple), or are co-annotated in any of the following annotation sets: EcoCyc pathways, heteromeric protein complexes, operon, regulon, KEGG module or STRING interaction. The dashed line shows precision for randomly ordered gene pairs generated as described in the Methods (negative control). (b) Using the same set of genes as in the previous panel, phenotypic profile similarity was calculated using the fitness scores from either Nichols *et al*., Price *et al*., or the combined fitness scores. Gene pairs were ranked from high to low similarity based on |PCC| values and plotted versus precision calculated for gene pairs co-annotated in the union of the six annotation sets. Only the first 500 gene pairs are shown. The different colored lines indicate the source of the fitness scores: Nichols *et al*. (red), Price *et al*. (purple) and the combination (blue)


(a)
1.00
0.75
0.50
0.25
0.00
similarity
10 %removed
50% removed
80% removed
all pairs
original
(b)
Pathway
1.0
0.8
0.6
0.4
0.2
0.0
precision
original
10 %removed
50% removed
80% removed
0
100
200
300
400
500
high similarity -- gene pairs ranked by |sim| -- low similarity

(c)

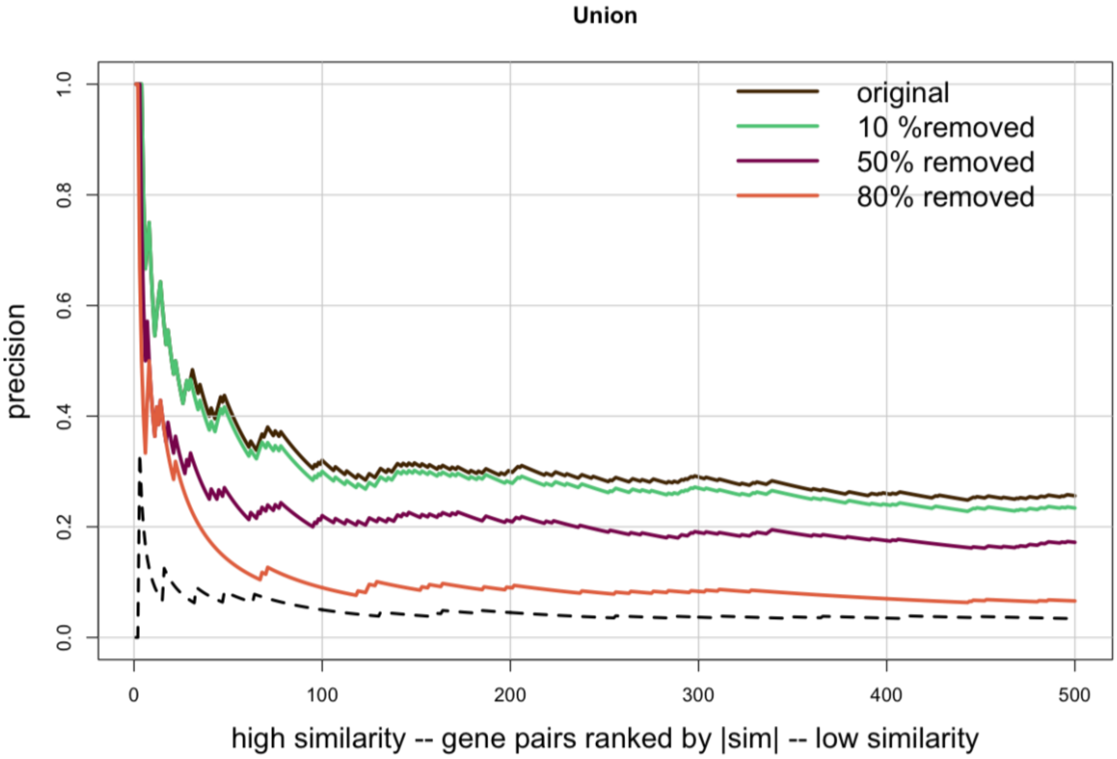


Figure 3.5. Effect on precision of removing annotations

(a) Distributions of phenotypic profile similarity for gene pairs co-annotated to the same pathway, after removal of 10%, 50%, or 80% of pathway annotations. No significant difference was observed for any group of co-annotated gene pairs when comparing to all gene pairs, based on 1-sided Mann-Whitney U test. (b) Precision versus ranking after removal of 10%, 50%, or 80% of pathway annotations for mutants. (c) Precision versus ranking after removal of 10%, 50%, or 80% of annotations for mutants from the union of the following annotation: EcoCyc pathways, heteromeric protein complexes, operon, regulon, KEGG module or STRING interaction.

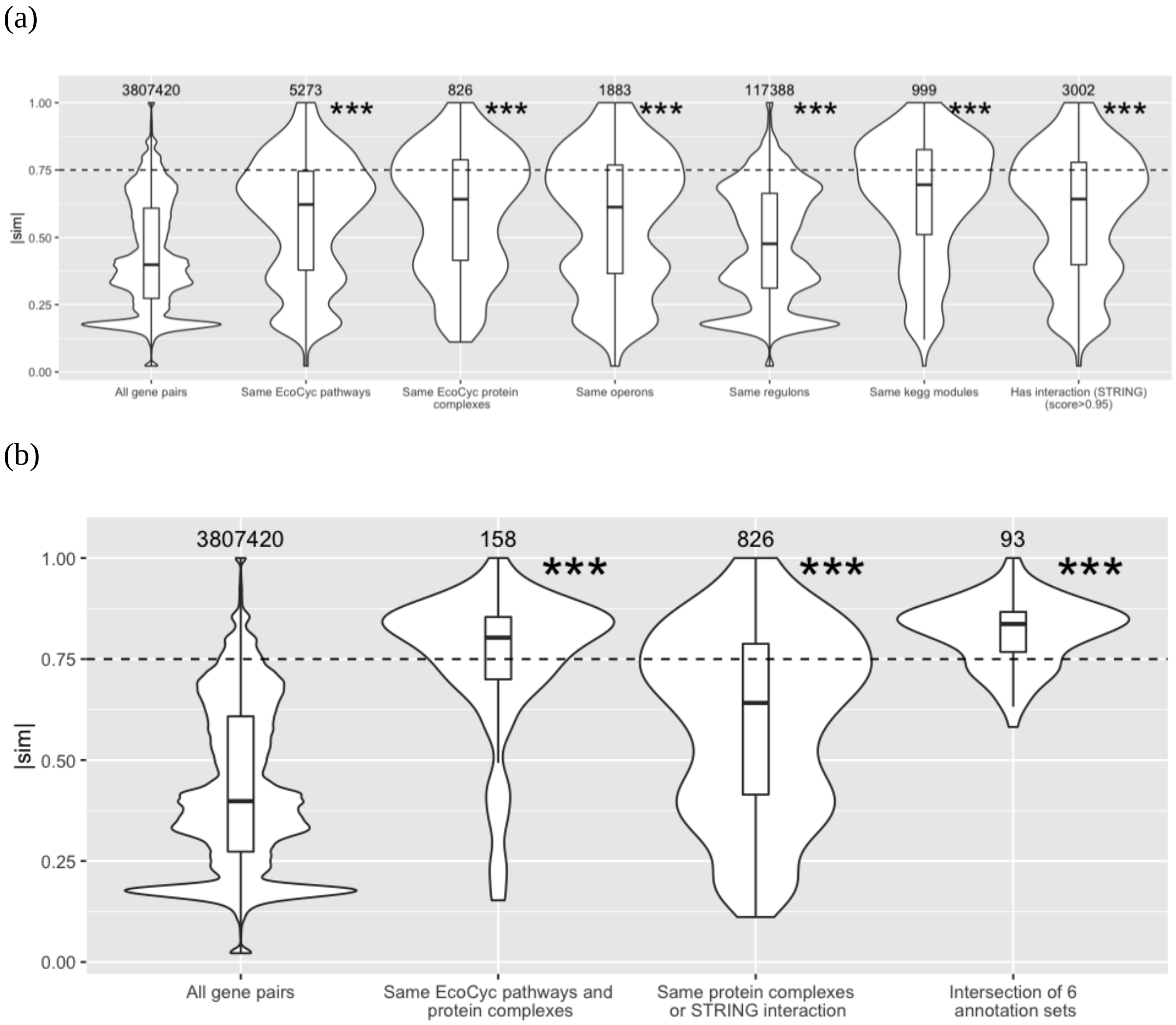


Figure 3.6. Distribution of OMP based semantic similarity for curated phenotypes from Nichols *et al.*, Price *et al.* and Campos *et al.*

Distributions of phenotypic profile similarity for gene pairs co-annotated in: (a) single annotation sets, and (b) combinations of annotation sets. ***: p value<0.001 based on one-sided Mann-Whitney U test.

CHAPTER 4. PHENOTYPIC ASSOCIATIONS AMONG CELL CYCLE GENES IN SACCHAROMYCES CEREVISIAE[2]

NOTE:

In this paper, my contribution is to the Gene-Ontology-related analysis and the critical analysis on the analytical pipelines and methods. I was listed as the co-first author with Dr. Rosa M. Bermudez.

ABSTRACT

A long-standing effort in biology is to precisely define and group phenotypes that characterize a biological process, and the genes that underpin them. In Saccharomyces cerevisiae and other organisms, functional screens have generated rich lists of phenotypes associated with individual genes. However, it is often challenging to identify sets of phenotypes and genes that are most closely associated with a given biological process. Here, we focused on the 166 phenotypes arising from loss-of-function and the 86 phenotypes from gain-of-function mutations in 571 genes currently assigned to cell cycle-related ontologies in S. cerevisiae. To reduce this complexity, we applied unbiased, computational approaches of correspondence analysis to identify a minimum set of phenotypic variables that accounts for as much of the variability in the

[2] *This is an open access article reprinted from ""Phenotypic Associations Among Cell Cycle Genes in Saccharomyces cerevisiae" by Bermudez, R. M.; Wu, P. I.; Callerame, D.; Hammer, S.; Hu, J. C. and Polymenis, M. *G3: Genes|Genomes|Genetics*, Volume 10, Issue 7, page 2345-2351 under the terms of the Creative Commons CC BY license.

data as possible. Loss-of-function phenotypes can be reduced to 20 dimensions, while gain-of-function ones to 14 dimensions. We also pinpoint the contributions of phenotypes and genes in each set. The approach we describe not only simplifies the categorization of phenotypes associated with cell cycle progression but might also potentially serve as a discovery tool for gene function.

INTRODUCTION

The generation of systematic mutant collections in a variety of model systems enables large-scale phenotypic screens, which are now standard in academic and commercial settings. The first organism for which such mutant collections became available is the budding yeast Saccharomyces cerevisiae (Giaever & Nislow, 2014). As a result, there is a wealth of phenotypes associated with most genes in that organism, displayed in easily accessible databases (Cherry et al., 2012; Engel et al., 2010). Gene Ontology (GO) techniques accurately specify the semantic relationships between terms, and they are indispensable for representing and organizing the accumulating biological knowledge (Ashburner et al., 2000). Curations of the literature and computational approaches have given rise to the systematic categorization of individual genes to biological processes.

However, given the numerous phenotypes often associated even with a single gene, the more genes involved in a biological process, the larger the number of phenotypes associated with that process. Hence, despite the plethora of phenotypic information on a per-gene basis, there is a loss in clarity and priority to the phenotypes most pertinent to the biological process in question. For example, at the time of preparing this report, based on the information on the Saccharomyces Genome Database (Cherry et al., 2012), there were at least 571 S. cerevisiae genes assigned to cell cycle related processes (see next Section). Collectively, there were 166 loss-of-function phenotypes associated with these genes, with additional qualifiers raising that number to 371 phenotypes. Among this bewildering set, identifying the phenotypic variables that cluster together in different groups and the genes that drive this classification may offer new insights into phenotype-phenotype and gene-phenotype associations within this biological process.

Network-based approaches have been used to link diseases with disease genes in humans, revealing common genetic origins of several conditions (Goh et al., 2007). Widely used multivariate statistical techniques can simplify related variables. Measuring the degree that the observed variables correlate with each other, provides the basis for the number of variables in a dataset to be reduced. If two or more phenotypic variables share some features, then based on the magnitude and direction of the relationship, the observed complexity may be simplified. Techniques implementing the above principles include factor analysis and principal component analysis (Child, 1990). For categorical data (e.g., the presence or absence of a phenotype), a related approach is that of correspondence analysis (J.-P., 1992).

Here, we identified 571 genes associated with cell division and cell cycle progression. We applied correspondence analysis to examine the numerous phenotypes associated with these genes, resulting both from loss- and gain-of-function mutations. Some phenotypic associations were generic, with mutations affecting vegetative and respiratory growth, or resistance to toxins, pH, and metals. In other cases, the clustering of some phenotypes and the gene associations was consistent with the literature. For example, loss-of-function mutations that affect shmoo formation and mating efficiency together contributed most significantly in one of the dimensions. Likewise, gain-of-function mutations affecting cellular morphology, size, and budding index together contributed significantly in another dimension. Hence, systematic phenotypic associations provide a useful dissection of biological processes and gene functions.

MATERIALS AND METHODS

DATASETS

All the individual phenotypic reports for each gene were downloaded from the Saccharomyces Genome Database (https://www.yeastgenome.org/). Loss-of-function phenotypes included not only those reported for ‘null’ alleles, but also ‘conditional’, ‘repressible’, and ‘reduction of function’ ones. Gain-of-function phenotypes included ‘activation’, and ‘overexpression’. Phenotypes that arose from ‘unspecified’ alleles were excluded from the analysis. To assemble the individual files into a single spreadsheet, we used R language packages. The files were read using the readr package. For example, for the loss-of-function files, the command was: lof_files = list.files(path = ‘…’, pattern = “*.txt”, full.names = TRUE). Then, the individual files were assembled into a list, with the command: lof_list = lapply(lof_files, read_tsv). The list components were combined into a dataframe with the following command from the dplyr package: lof_parent_child <- bind_rows(lof_list, .id = NULL). The resulting spreadsheet is in File2/sheet ‘lof_parent_child’. There were 371 loss-of-function phenotypes associated with 561 genes. However, in many cases, the phenotypic terms included qualifiers. For example, for the parent term ‘vegetative growth’ there were qualifiers, such as ‘increased’, ‘increased rate’, etc. To simplify the analysis, we removed these qualifiers and focused only on the 161 parent, loss-of-function phenotypic terms. To split the parent terms from their qualifiers, we used the following command from the tidyr package: lof_parent <- separate(data = lof_parent_child, col = phenotypes_lof, into = c(“parent_ontology”, “child_ontology”), sep = “:”, remove = TRUE, convert = FALSE, extra = “warn”, fill = “warn”). The resulting spreadsheet is in File2/sheet ‘lof_parent’. For the gain-of-function phenotypes, the analogous spreadsheets are in File3/sheet ‘gof_parent_child’ and ‘gof_parent’.

To gauge whether phenotypic profiles for genes in the loss-of-function dataset (lof_parent.txt) associate with functions, for each gene pair, we calculated the semantic similarity based on Gene Ontology annotations (Yu et al., 2010). For this analysis, the R language package infotheo was used to calculate the mutual information-based similarity metric for all pairs of genes. Then, the R language package GOSemSim was used to calculate the semantic similarity between gene pairs based on the GO annotations of either molecular function, biological process or cellular component (Yu et al., 2010). Significantly higher semantic similarity was indeed observed between genes that have more similar phenotypic profiles (Figure S1).

FACTOR ANALYSIS

Multiple correspondence analysis (MCA) was performed with the R language package FactoMiner, and the related ones factoextra, and FactoInvestigate. For the loss-of-function phenotypes, we used the lof_parent spreadsheet as input (File2/sheet ‘lof_parent’), after it was transposed, so that the phenotypic variables were columns and the genes rows. Then we used the command: lof_MCA <- MCA(lof_parent, method = “Burt”). All the Eigen values associated with the analysis are in File2/sheet ‘lof_eigen’. To identify the number of the most significant dimensions, we used the command: dimRestrict(lof_MCA), which identified 20 dimensions as the most significant. We then re-run the MCA function for 20 dimensions, as follows: lof_MCA <- MCA(lof_parent, method = “Burt”, ncp = 20). The cosine values from the correspondence analysis represent the correlation coefficients (Child 1990). The cos2 values for the phenotypic variables were obtained with the command ‘get_mca_var(lof_MCA)’ and listed in File2/sheet ‘lof_var_cos2_20dim’. The cos2 values for the individuals (genes) were obtained with the command ‘get_mca_ind(lof_MCA)’ and they are listed in File2/sheet ‘lof_ind_cos2_20dim’. Based on this analysis, each of the genes was assigned to one of the 20 most significant dimensions (shown in File2/sheet ‘lof_gene_20dim’).

To interpret the dimensions, we used the ‘dimdesc’ function of the FactoMiner R language package. For each dimension (the example is for dimension 1), we run the command: res1_dimdesc = dimdesc(lof_MCA, axes = 1:1, proba = 1). The results for each dimension, with the R2 values for each phenotype and the associated p-value, are in the sheets of File2 (e.g., ‘res1_dimdesc’ for dimension 1, and so on).

The analogous analysis was done for the gain-of-function phenotypes, and all the data are in File3.

NETWORK VISUALIZATION

For the networks shown in Figures S2-S4, we used the GeneMANIA Cytoscape plugin (Franz et al., 2018; Montojo et al., 2014; Montojo et al., 2010; Warde-Farley et al., 2010).

DATA AVAILABILITY

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables. All datasets (Files1-3) and Supplementary Figures (S1-S4) have been deposited via a public repository (figshare): https://doi.org/10.6084/m9.figshare.12234695.v1

RESULTS

GENE SET

Before analyzing any phenotypes associated with cell division and cell cycle progression, it is essential to identify the genes related to these processes. At the time of writing this report, the biological process 'cell cycle' (GO:0007049) was defined as: "The progression of biochemical and morphological phases and events that occur in a cell during successive cell replication or nuclear replication events. Canonically, the cell cycle comprises the replication and segregation of genetic material followed by the division of the cell …" (https://www.yeastgenome.org/go/7049). There were 307 genes annotated to the 'cell cycle' biological process (File1). However, we noticed that some genes that govern vital cell cycle events were not in this set. For example, SIC1, encoding a cyclin-dependent kinase inhibitor that must be destroyed for DNA replication to begin. Destruction of Sic1p is the only essential function of G1 cyclins (Schneider et al., 1996). Another gene that was not in the computationally annotated 'cell cycle' genes was MPS1, which encodes a conserved kinase that is essential for spindle pole body duplication (Liu & Winey, 2012).

Consequently, we looked at additional biological processes (File1), such as 'DNA replication' (GO:0006260), 'chromosome segregation' (GO:0007059), 'cell division' (GO:0051301). All the genes in the 'cell division' process were annotated computationally and were also in the 'cell cycle' set (Figure 4.1). However, several genes in the 'DNA replication' and 'chromosome segregation' processes, were not annotated as 'cell cycle' genes (Figure 4.1). We also noted that there was incomplete overlap between the genes that were annotated computationally or by manual curation within the 'DNA replication' and 'chromosome segregation' processes themselves (File1, sheets 0006260 and 0007059). To ensure that our list of cell cycle genes is as comprehensive as possible, we started with all the genes in the 'cell cycle' (GO:0007049), 'DNA replication' (GO:0006260), 'chromosome segregation' (GO:0007059), and 'cell division' (GO:0051301) categories, and also included all the genes in all the 'children' categories to the above gene ontology nodes. These additional categories (n = 100) are listed in File1/sheet 'categories' (see also the individual sheets numbered as the corresponding gene ontologies), and they were grouped as 'OTHER' (see File1/sheet 'sets_Figure 1'). The overlap between the 'cell cycle' (GO:0007049), 'DNA replication' (GO:0006260), 'chromosome segregation' (GO:0007059), 'cell division' (GO:0051301), and 'OTHER' sets is shown in Figure 1. A total of

185 genes were unique to the ‘OTHER’ set. Overall, there were 571 unique genes in all these, gene ontology-based, biological processes related to cell division, and cell cycle progression (File1/sheet: ‘genes’).

Before proceeding to more detailed categorization of the distinct phenotypes among cell cycle genes, we asked a more general question: Is it reasonable to expect that genes with similar function(s) will also have similar phenotypes? Indeed, we found that there is a significantly higher semantic similarity between genes that have more similar phenotypic profiles (Figure 4.S1, and Materials and Methods). In the rest of this study, we analyzed the loss- and gain-of-function phenotypes associated with each of these 571 genes.

LOSS-OF-FUNCTION PHENOTYPES

To analyze the 166 phenotypes associated with loss-of-function mutations in 561 genes, we tabulated them as we describe in the Materials and Methods. Correspondence analysis was performed with the R language package *FactoMiner*, and the related ones *factoextra* and *FactoInvestigate* (see Materials and Methods). We found that there were 20 significant dimensions, accounting for ≈2/3 of the observed variance (Figure 4.2, bottom). The percentage of the 561 genes associated with each of these 20 dimensions is shown in Figure 4.2, top. A detailed list is in File2/sheet ‘lof_gene_20dim’.

A major objective is to identify which phenotypic variables the 20 dimensions are the most linked to, in other words which phenotypes describe the best each dimension. For the loss-of-function phenotypes, this is shown graphically in Figure 4.3(detailed lists for each phenotype and dimension are in File2). The phenotypes that were most significantly associated (an arbitrary cutoff was chosen at $R^2 \geq 0.2$) with the most populous dimension (#1; 24% of all genes), were very general, and not particularly informative (Figure 4.3): chemical compound accumulation, respiratory or vegetative growth, metal resistance, etc (see File2/sheet 'res1_dimdesc'). The only other cell cycle-related phenotype in this group was 'cell size'. Cell size changes are often interpreted as perturbations in the normal coupling of cell growth with cell division (Jorgensen et al., 2002), albeit there is not a strong correlation between cell size and the length of the G1 phase of the cell cycle (Blank et al., 2018; Hoose et al., 2012). In other dimensions, interesting and expected associations were evident. For example, 'shmoo formation', 'bud neck morphology', and 'pheromone induced cell cycle arrest' were associated with Dimension 2 (Figure 4.3). Secretory processes with the phenotypes affecting 'endoplasmic reticulum distribution', 'peroxisomal morphology', 'Golgi distribution' were associated heavily with Dimension 4. Similarly, 'vesicle distribution' and 'vacuolar transport' were associated with Dimension 15. The constellation of phenotypes associated with loss-of-function mutations in *TOR2* is unique. *TOR2* is the only gene in Dimension 16, with 'metabolism and growth' and 'osmotic stress resistance' being the most prominent phenotypes. The remaining dimensions were defined by phenotypes that were only weakly ($R^2 \geq 0.2$) associated with cell cycle progression.

GAIN-OF-FUNCTION PHENOTYPES

There were 86 phenotypes associated with gain-of-function mutations in 368 genes (from a total of 571 genes). The phenotypic matrix was organized and analyzed as for the loss-of-function mutations (see Materials and Methods). Based on correspondence analysis we found that there were 14 significant dimensions (Figure 4.4, bottom), with the vast majority of genes grouped in just one dimension (#2; see Figure 4.4, top). A detailed list is in File3/sheet 'gof_gene_14dim'. We next identified the phenotypic variables for the gain-of-function mutants describe the best each dimension (Figure 4.5, detailed lists for each phenotype and dimension are in File3). Most genes (≈60%) were grouped in Dimension 2. The phenotypes that contributed most significantly ($R^2 \geq 0.2$) to Dimension 2 were: 'cellular morphology', 'budding index' (a proxy for altered cell cycle progression), 'cell size', and 'cell cycle progression in G2 phase' (Figure 4.5).

COMPARISONS WITH NETWORKS OF GENETIC AND OTHER INTERACTIONS

How does the grouping of the cell cycle genes we described above compare to other approaches? Functional interaction networks, based on genetic or physical interactions among gene products, provide the means to visualize the organization of cellular pathways. However, when we displayed the network of all the reported genetic (Figure 4.S2) or physical (Figure 4.S3) interactions among all the cell cycle genes (shown in File1/sheet: 'genes'), there were no obvious higher-order classifications. Co-localization of different proteins in the cell provides another means of gaining insight into higher-order classification of gene products. By that co-localization measure, many cell cycle genes were clearly organized in distinct clusters (Figure 4.S4). Nonetheless, there was no overlap between the gene products that were co-localized, and the genes that belonged to the groups we identified by phenotypic clustering. These results suggest that the phenotype-based approach we described provides new information and expands the efforts to reveal the higher-order organization of cell cycle gene products.

DISCUSSION

The results we presented are significant for several reasons: First, the multitude of phenotypes associated with genes involved in cell cycle progression can be grouped in a smaller number of categories, simplifying their analysis and the gene contributions to each category. Second, the phenotype-based categorization we described provides a separate, independent view of the biological process in question, which is not captured by the network of the genetic or physical interactions among the genes analyzed. Third, the approach we described ought to apply to any biological process.

When testing gene function, the old maxim "when in doubt knock-it out" took a more expansive turn with the availability of genome-wide deletion sets. For several model systems, and especially *S. cerevisiae*, these sets enable large-scale, often automated, phenotypic assays (Giaever et al., 2002; Giaever & Nislow, 2014). As the phenotypes associated with each gene increase, it becomes less clear which of the phenotypes associated with each gene are the most pertinent to the biological process in question. A key component in addressing this issue is high-quality annotation from the available databases. Gene Ontology (GO) categories standardize gene product annotations with regards to molecular function, biological process, and cellular component. *S. cerevisiae*is probably better annotated than most other experimental organisms, with computational and human-based approaches (Cherry et al., 2012). Yet, even in this organism, as we showed for the cell cycle genes (Figure 4.1), there is not a complete overlap among the different approaches, underscoring the need for continued efforts to improve systematic annotation (Siegele et al., 2019). Other approaches have also been developed that look for patterns in existing annotations, with the objective to correct or improve those annotations (Khatri et al., 2005). This is not the general objective of the approach we described. We use current annotations from curated databases to reduce the apparent complexity of the observed phenotypes to fewer, more manageable groups, revealing associations between individual phenotypes and the genes that drive these associations. The relatively simple approaches we used here to cluster the diverse phenotypes reported in the literature are scalable to other biological processes and genomes.

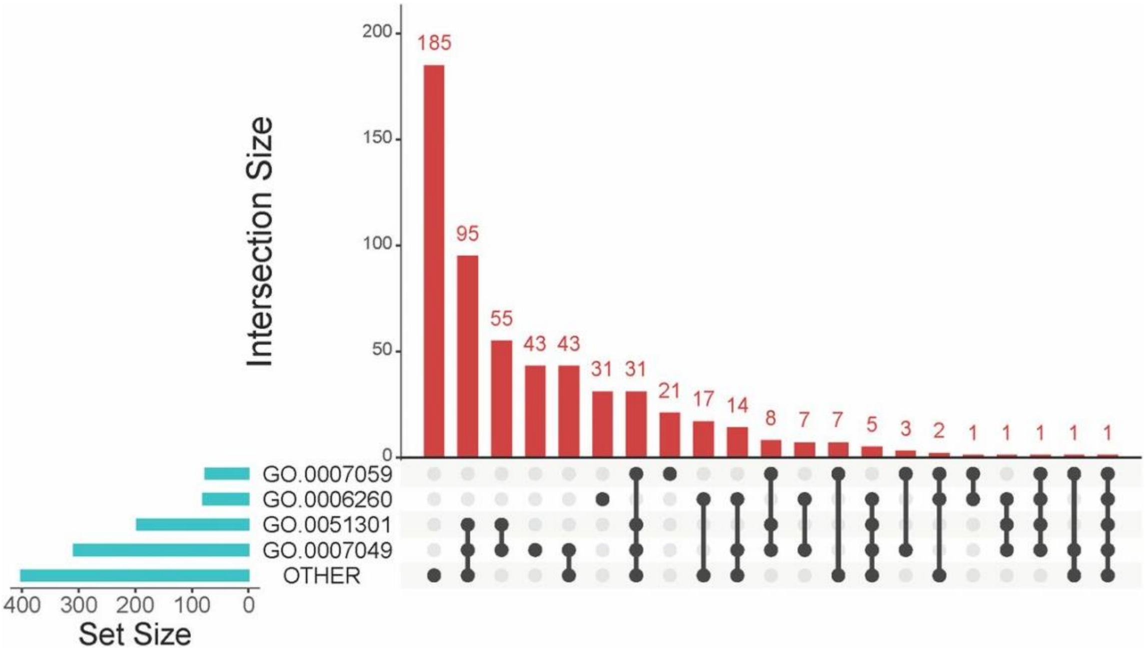


Figure 4.1. Gene ontologies related to cell cycle progression and cell division. Matrix layout for all intersections of the sets of genes we interrogated. Each red bar represents the number that are in the groups dotted black but not gray. The biological processes were 'cell cycle'

(GO:0007049), ‘DNA replication’ (GO:0006260), ‘chromosome segregation’ (GO:0007059), ‘cell division’ (GO:0051301). In ‘OTHER’ there were genes grouped together from various cell cycle-related ontologies, as described in the text and in Materials and Methods. The size of the sets is shown on the bar plot to the left. The number of genes unique to the indicated intersections is shown separately on the bar plot to the right. The names of all genes in each set are shown in File1/sheet ‘sets_Fig1’. The graph was drawn with the UpSet R language package.

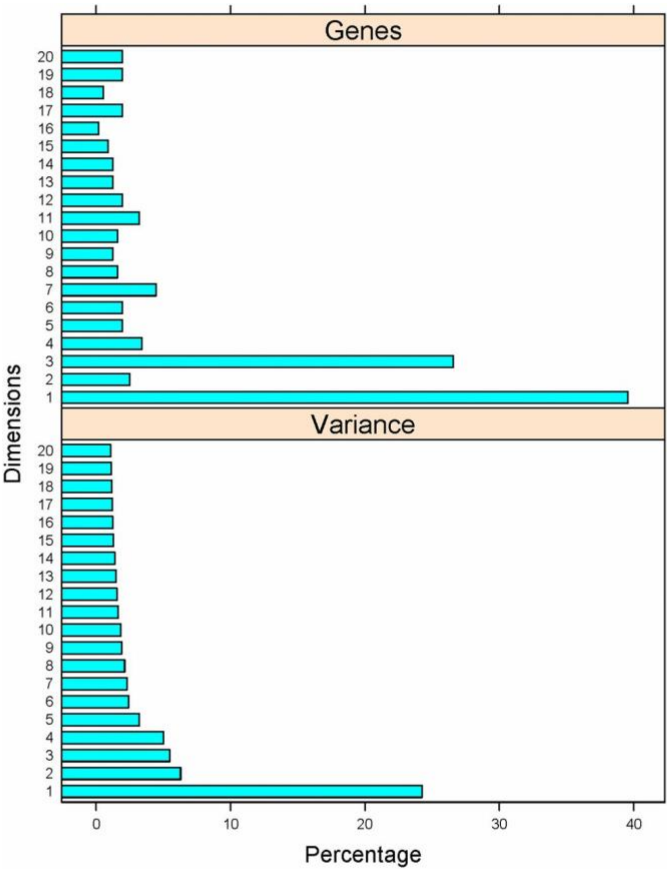


Figure 4.2. Phenotypic variance and gene associations with the 20 dimensions from the multiple correspondence analysis of the loss-of-function phenotypes of cell cycle-related genes.

Top, The percentage of genes (x-axis) most closely associated with each of the dimensions (y-axis). *Bottom*, The percentage of the variance (x-axis) explained by each of the dimensions shown (y-axis).

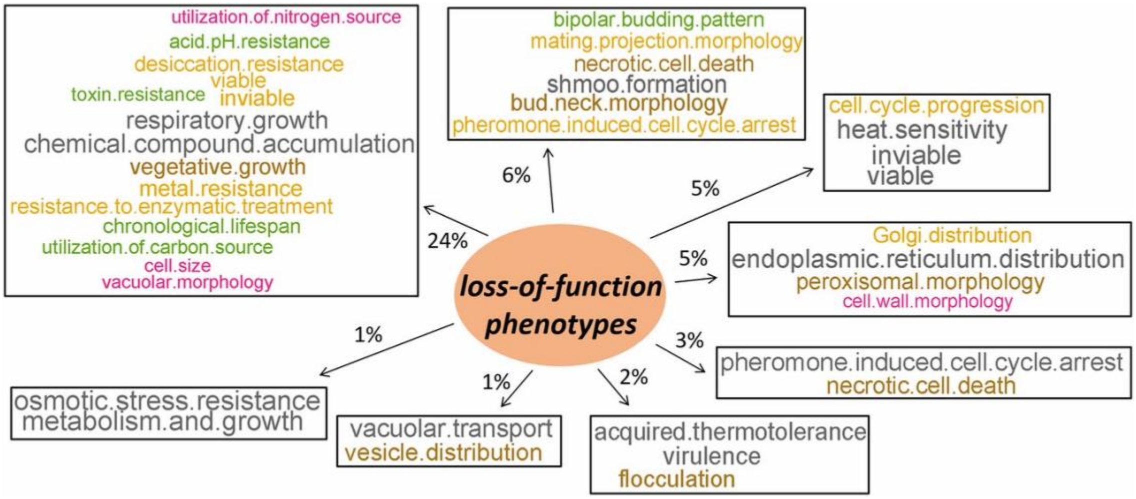


Figure 4.3. Gain-of-function phenotypes associated significantly with one of the 14 dimensions identified by MCA. The figure was generated as described for the loss-of-function phenotypes, shown in Figure 4.3.

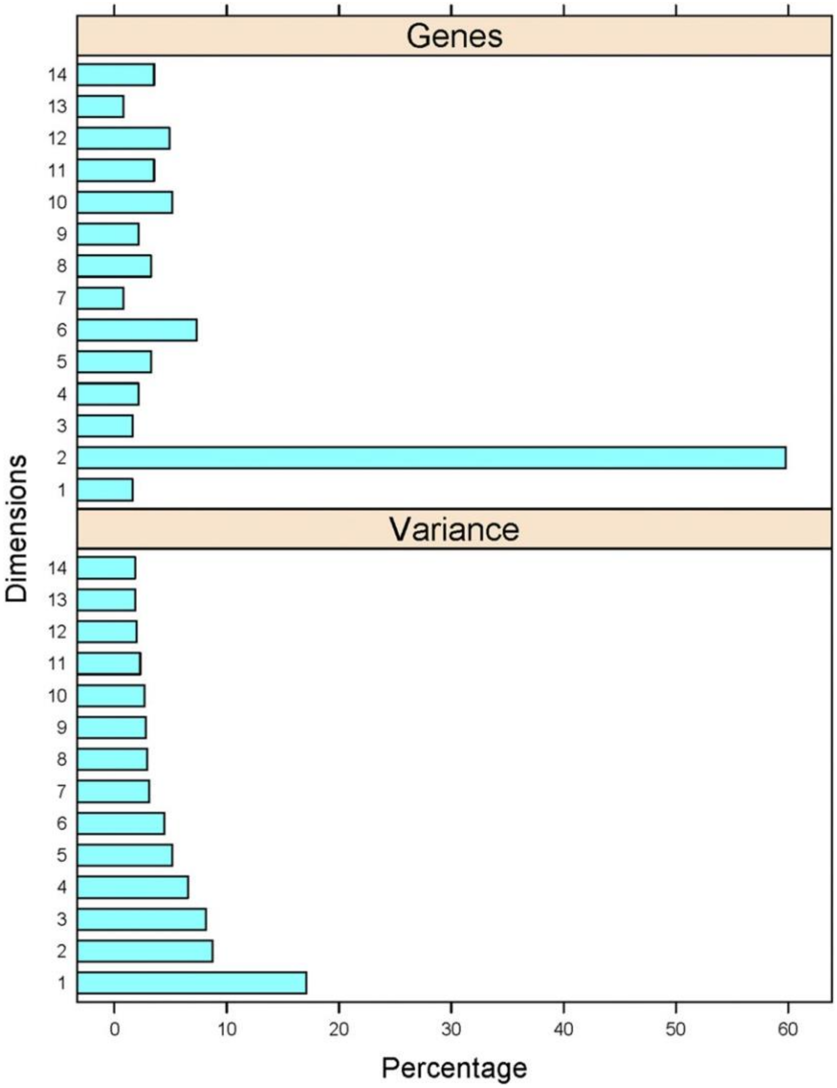


Figure 4.4. Phenotypicvarianceandgeneassociations with the 14 dimensions from the multiple correspon- dence analysis of the gain-of-function phenotypes of cell cycle-related genes. (Top) The percentage of genes (x-axis) most closely associated with each of the dimensions (y-axis). (Bottom) The percentage of the variance (x-axis) explained by each of the dimensions shown (y-axis).

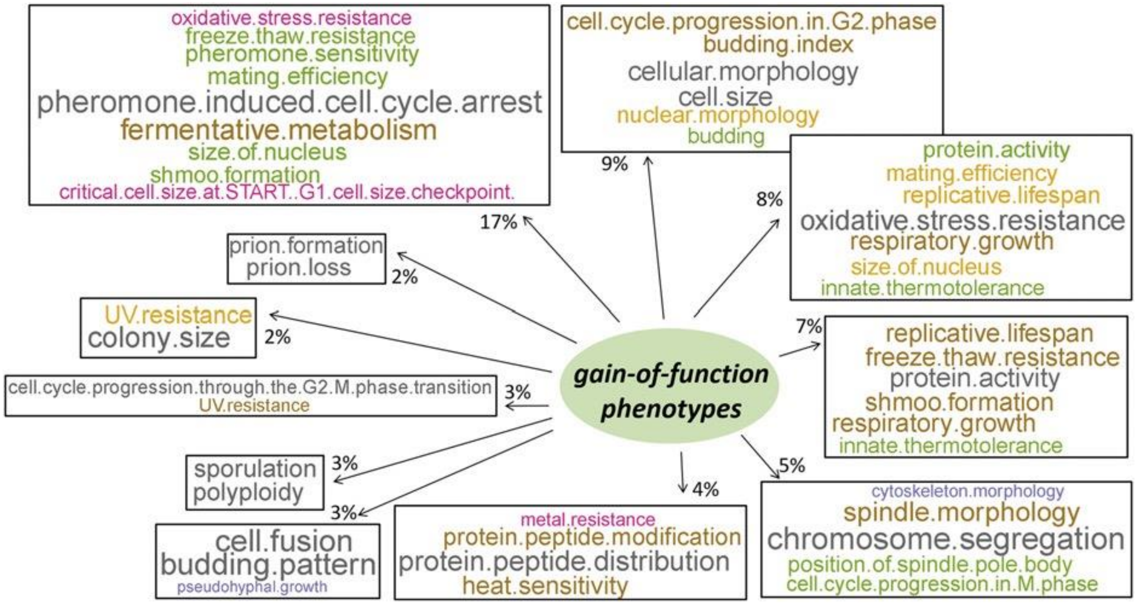


Figure 4.5. Gain-of-function phenotypes associated significantly with one of the 14 dimensions identified by MCA. The figure was generated as described for the loss-of-function phenotypes, shown in Figure 4.3.

SUPPLEMENTAL FIGURES

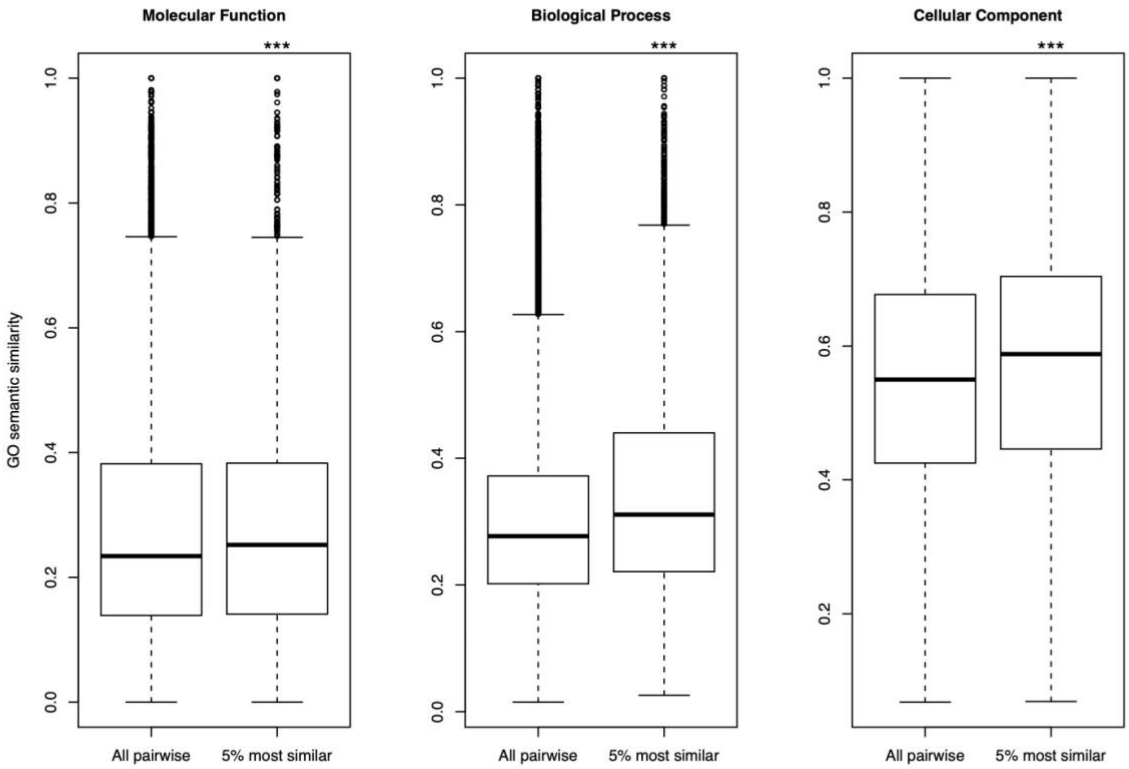


Figure 4.S1. Higher semantic similarity between genes that have similar phenotypic profiles.

Based on the SGD annotations that generated the phenotypic profiles, the top 5% phenotypically similar gene pairs have higher semantic similarity. ***: p-value<0.001 based on 1-sided Mann-Whitney U test.

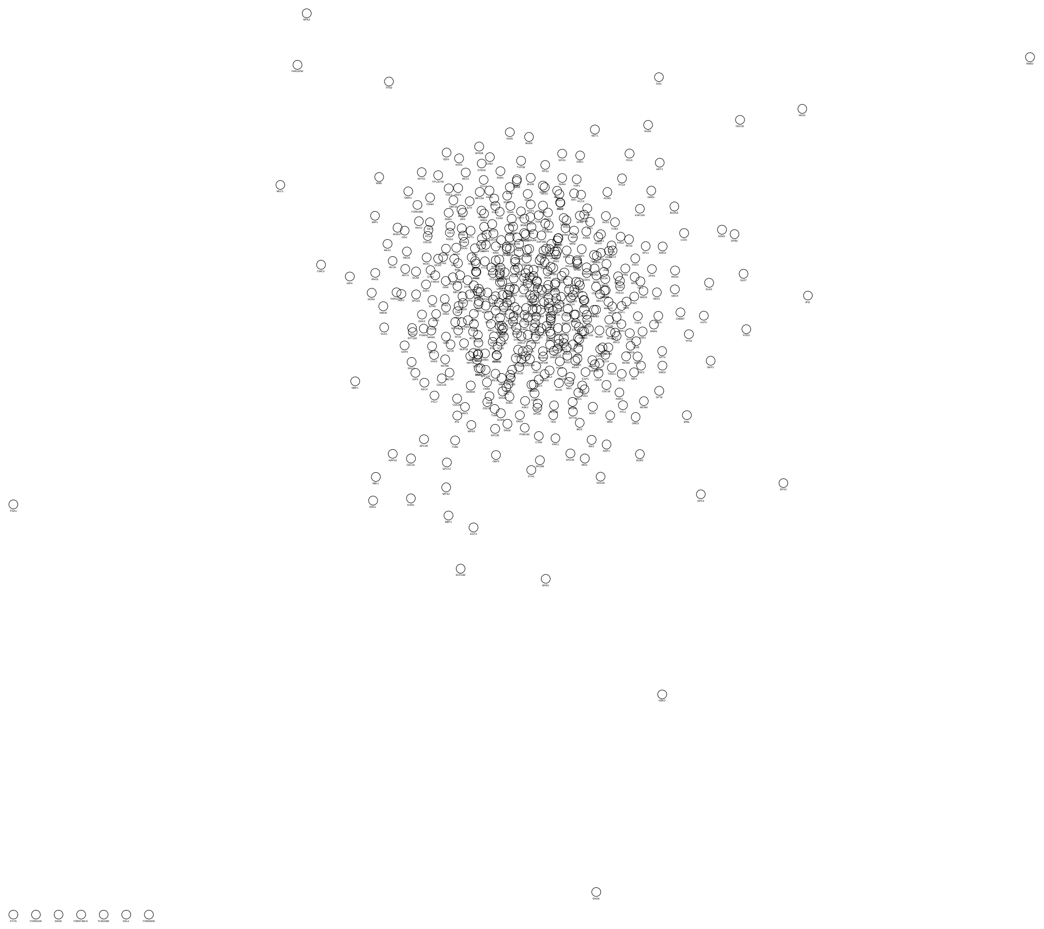


Figure 4.S2. Genetic interactions among the cell cycle genes.

Network of all the genetic interactions (n=26,522) incorporated in the GeneMANIA platform at the time of writing this manuscript among the cell cycle genes shown in File1/sheet: ‘genes’. The network was generated using the default settings of the Cytoscape software package. The gene name of each node is visible upon zooming into the provided pdf image.

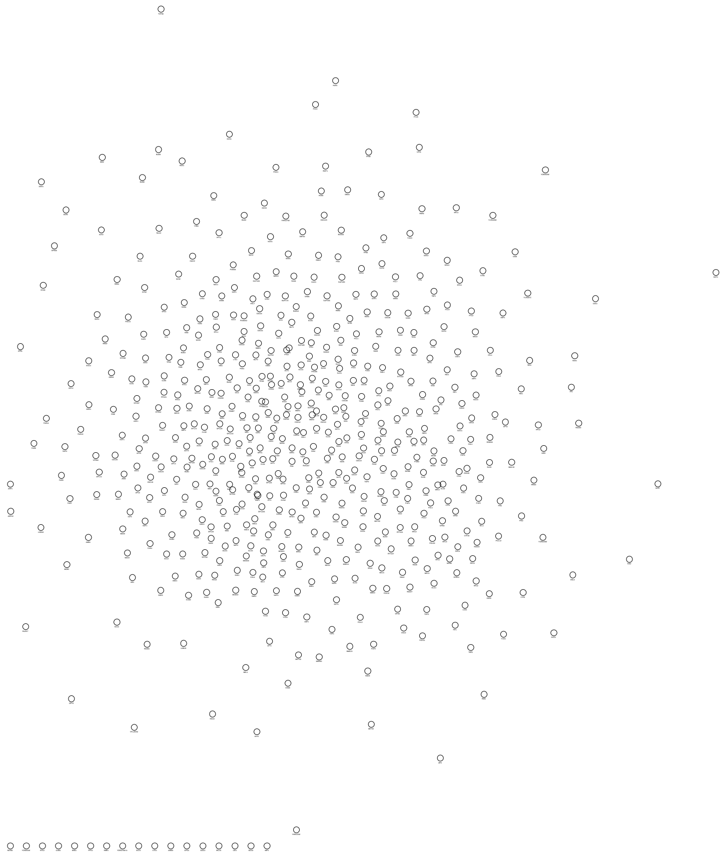


Figure 4.S3. Co-localization of cell cycle gene products.

Network of all the reported co-localizations (n=15,466) among cell cycle proteins, generated as in figure 4.S2.

ACKNOWLEDGMENTS

This work was supported by NIH grant R01GM123139 to M.P.

CHAPTER 5. MACHINE LEARNING ON MICROBIAL PHENOTYPES TO CLASSIFY GENE FUNCTIONS

ABSTRACT

High-throughput microbial phenotypes hold the potential in elucidating the functions of genes. Here we combined 2 high-throughput phenotype datasets with five categories of annotations as labels to classify genes with distinct functions. Preliminary results using complete phenotype data performed poorly, possibly due to from incomplete annotations and/or non-separable nature of functional associations of genome-wide studies. However, selecting small numbers of mutually exclusive classes significantly improves the performance, indicating that the power of high-throughput phenotyping can be coupled with machine learning to identify genes that are functionally connected.

INTRODUCTION

Phenotypes play important roles in understanding functions of genes, leading to better understanding of disease models and thus contribute to new drug discoveries. Among model organisms that can be easily manipulated to test hundreds of thousands of phenotypes in parallel, *E. coli* serves as one of the best, thanks to the scalability of its culturing. Here, we combined two different datasets (Nichols et al., 2011; Price et al., 2018) that contain phenotype data for mutants of almost every single gene in *E. coli* as the features (486 features for 3525 genes) and used five sets of gene annotations (Gama-Castro et al., 2016; Kanehisa et al., 2016; Karp et al., 2018) as the target variables to classify genes with the same function(s). The existing annotations are highly accurate, but they are incomplete and not mutually exclusive (one mutant/gene can have more than one annotation as labels). In addition, there are many annotations that do not label enough samples for most machine learning methods to train. Therefore, from each annotation set, we picked three gold-standard classes and used several well-established supervised learning techniques to demonstrate the power of high-throughput microbial phenotypes to explain gene functions.

METHODS

For each of the five annotation sets, subsets of annotated genes were selected, and their phenotypic profiles were used for analysis. Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector machine and Convoluted Neuro network were used to train supervised learning models. The code for all experiments performed in this study is can be found at: https://github.com/peterwu19881230/CSCE633_Machine_Learning

RESULTS

For each selected functional annotation set (Figure 5.1), six supervised learning techniques were applied. Maximum performance for each annotation ranges from 73% to 100% for both accuracy and precision (Figure 5.2). Table 5.1 shows the best hyperparameters. Overall, for the phenotype data used in this study, the best results were seen for gene products that are part of the same protein complex, and were worst for genes that are co-regulated. This result is reasonable since the deletion of any subunit in a protein complex is very likely to cause the same malfunction and downstream phenotypes, while co-regulated genes might perform different functions.

Comparing the different supervised learning methods (Figure 5.2), we observed that logistic regression stood out as the simplest and possibly strongest method, while the decision tree fell behind all other methods. It was surprising that although phenotypes as features shouldn't be independent of each other, more complicated methods such as Support Vector Machine and Convoluted Neuro Network didn't significantly surpass the performance of logistic regression. Further experimentation with larger sample size might help answer this question.

CONCLUSIONS

In this study, we have performed several supervised learning methods with a combined microbial phenotype dataset from 2 high-throughput studies. For every kind of annotation as labels, we get high accuracy and precision (> 70% for the best method of each annotation label). The results guarantee the utility of high-throughput, indirect phenotype measurement in explaining functions of genes.

DISCUSSION

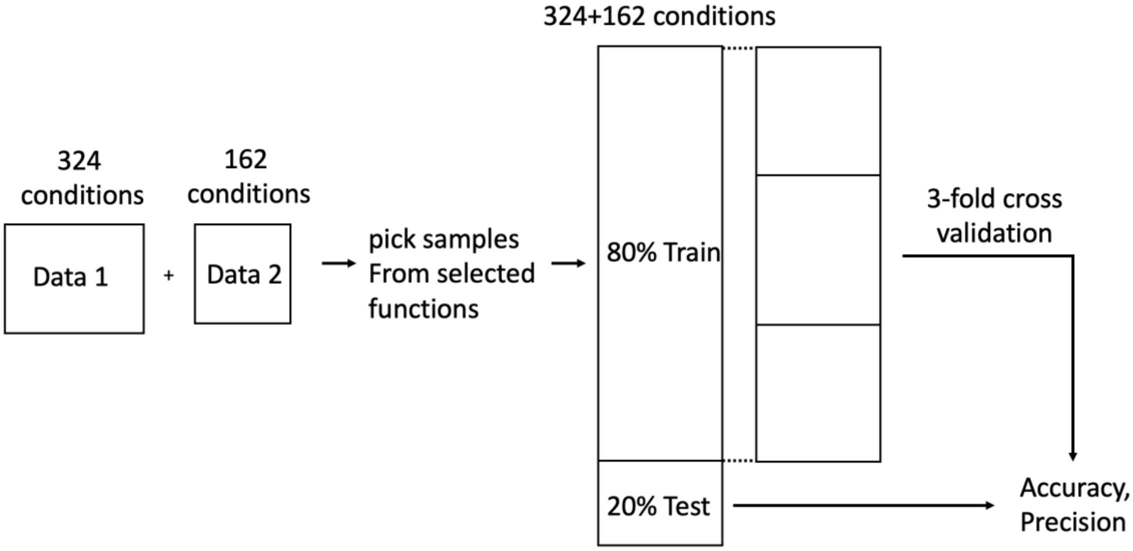
Small subsets of data picked by biochemical knowledge provide enough samples, which are mutually exclusive under distinct labels, in turn enabling high performance to learn functions from phenotypes. However, the impact of using machine learning on complete phenotype data with better labeling is yet to be done. It is worth noting that the 5 annotation sets selected here are mostly curated from experimental results from a very large number of publications. It is spectacular that different biochemical or molecular biological experiments that yield most of these annotations can be used as high-quality labels to examine phenotype data.

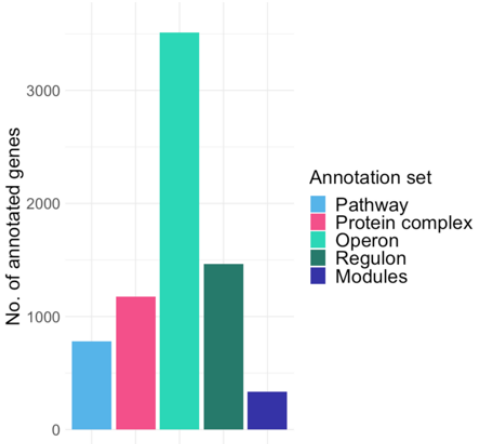
In order to see if there are obvious separation based on phenotypes, we have tried to reduce the dimensions of annotations by hierarchical clustering and separate genes into distinct categories. However, we obtained almost no separation by PCA, t-SNE or self-organizing map (Figure 5.3, Figure 5.4), and poor performance on all the supervised learning methods tested (~40% accuracy). Hopefully, with more annotations becoming available in the future, the genome-wide phenotypic data we have shown here can be much better exploited.

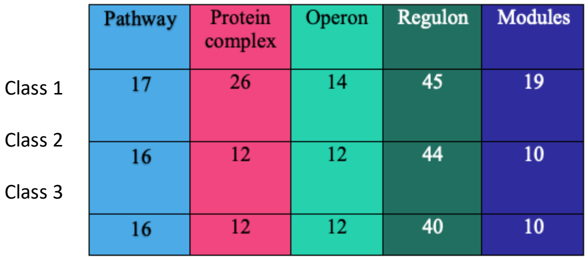
In addition to the 2 phenotype datasets described here, there are many others that measure distinct types of phenotypes (Campos et al., 2018; Fuhrer et al., 2017; Typas et al., 2008), whereas in this study, our phenotypes (features) are simply growth rates measured by the following: 1. number of pixels of colony sizes under different stress/growth conditions and 2. Sequencing results from a competitive growth assay. Incorporating other high-throughput phenotype studies and combining all of them might be an interesting future direction to decipher functions of genes in more detail as well as facilitating more generalized machine learning models to be developed.

ACKNOWLEDGEMENT

I (Peter) thank the input from Chiou-Jin Huang and Hao-Yu Miao for the code optimization and performance tuning. I also appreciate the lead from Dr. Bobak Mortazavi from the Computer Science department and the genuine support from my graduate advisors, Dr. James C. Hu and Dr. Deborah Siegele.



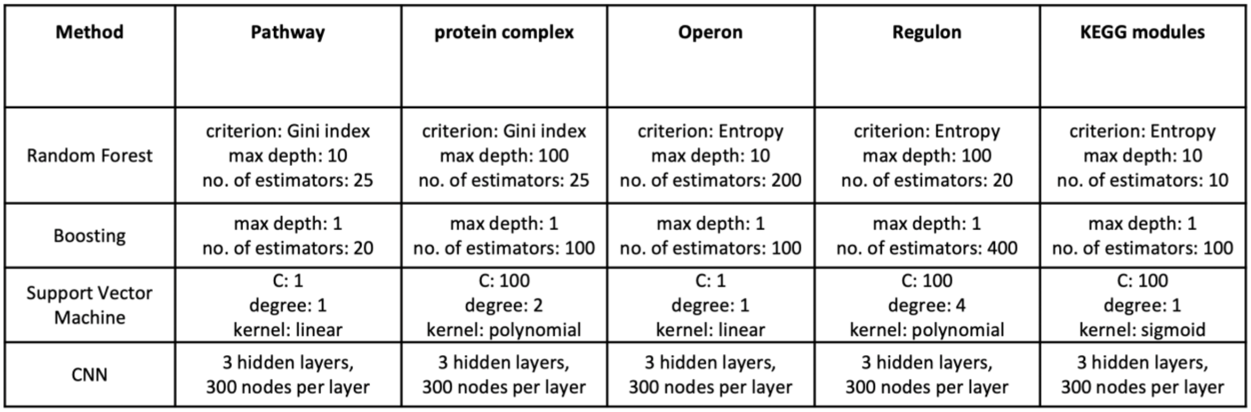




	Pathway	Protein complex	Operon	Regulon	Modules
Class 1	17	26	14	45	19
Class 2	16	12	12	44	10
Class 3	16	12	12	40	10

Figure 5.1. Supervised learning workflow and data statistics

(upper) Supervised learning workflow. (bottom left) No. of annotated samples selected for each annotation set. (bottom right) A table showing the number of samples drawn for each class within each annotation set as independent labels. 3 classes were selected for each class.



Method	Pathway	protein complex	Operon	Regulon	KEGG modules
Random Forest	criterion: Gini index max depth: 10 no. of estimators: 25	criterion: Gini index max depth: 100 no. of estimators: 25	criterion: Entropy max depth: 10 no. of estimators: 200	criterion: Entropy max depth: 100 no. of estimators: 20	criterion: Entropy max depth: 10 no. of estimators: 10
Boosting	max depth: 1 no. of estimators: 20	max depth: 1 no. of estimators: 100	max depth: 1 no. of estimators: 100	max depth: 1 no. of estimators: 400	max depth: 1 no. of estimators: 100
Support Vector Machine	C: 1 degree: 1 kernel: linear	C: 100 degree: 2 kernel: polynomial	C: 1 degree: 1 kernel: linear	C: 100 degree: 4 kernel: polynomial	C: 100 degree: 1 kernel: sigmoid
CNN	3 hidden layers, 300 nodes per layer	3 hidden layers, 300 nodes per layer	3 hidden layers, 300 nodes per layer	3 hidden layers, 300 nodes per layer	3 hidden layers, 300 nodes per layer

Table 5.1. Best hyperparameters for each supervised learning method for each annotation (labels). For Logistic regression and Decision tree there were no hyperparameter tuning.

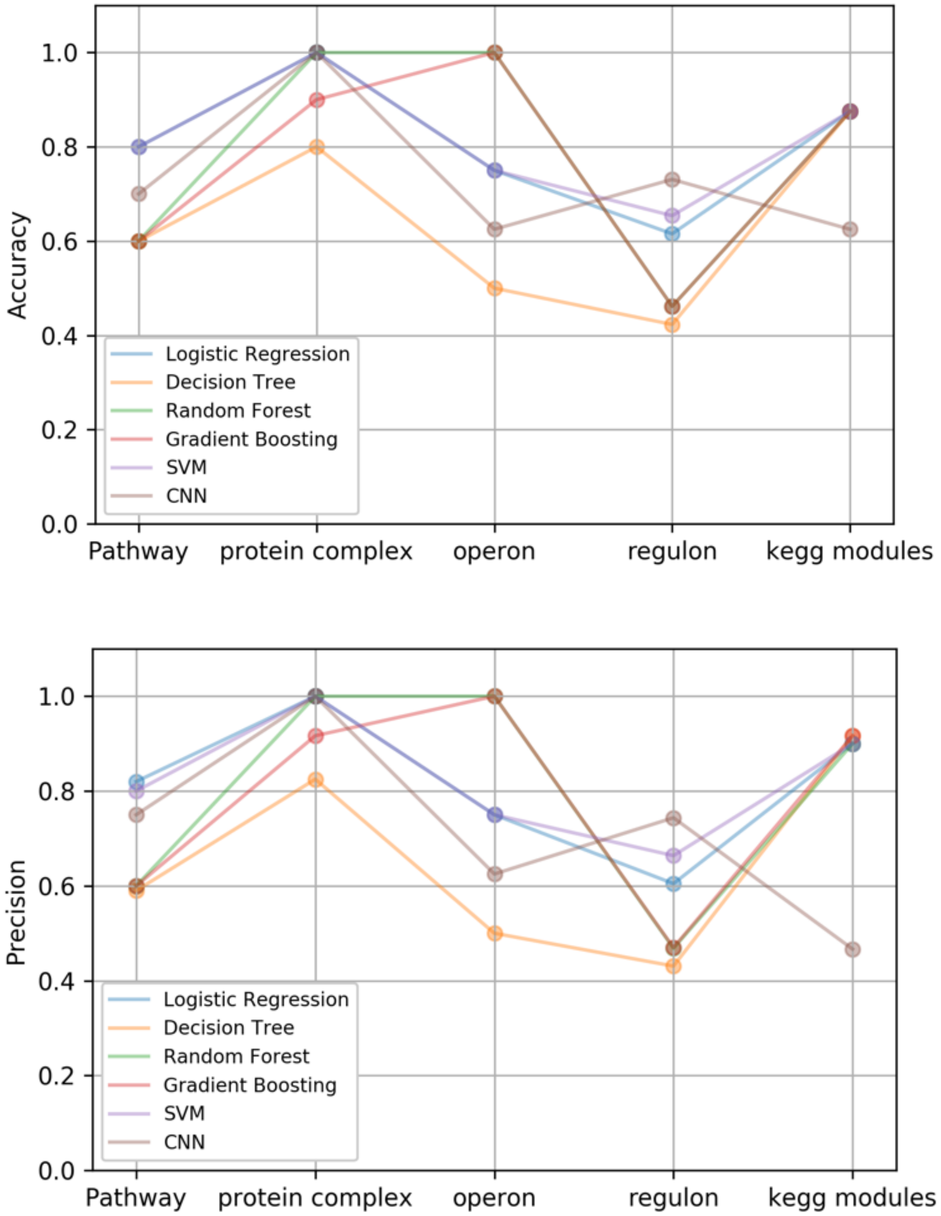


Figure 5.2. Accuracies and Precisions from each supervised machine learning methods. (Upper) Accuracy and (Bottom) precision calculated using various supervised machine learning methods when using different annotations (labels).

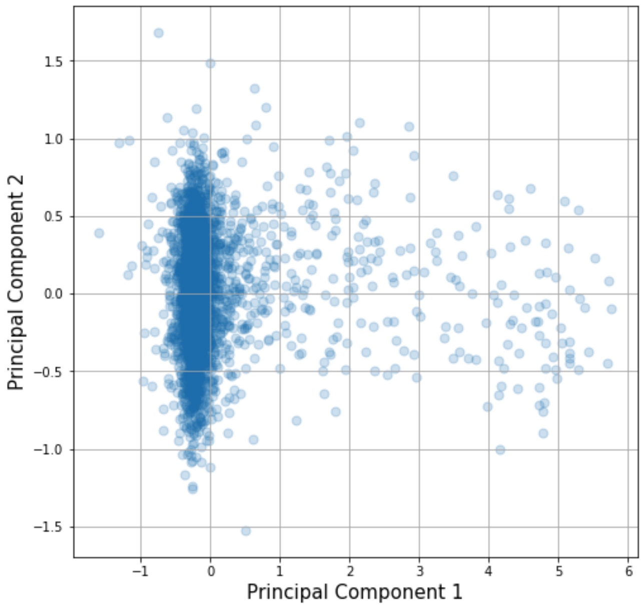


Figure. 5.3. PCA using complete phenotype dataset. There are no obvious functional clusters observed

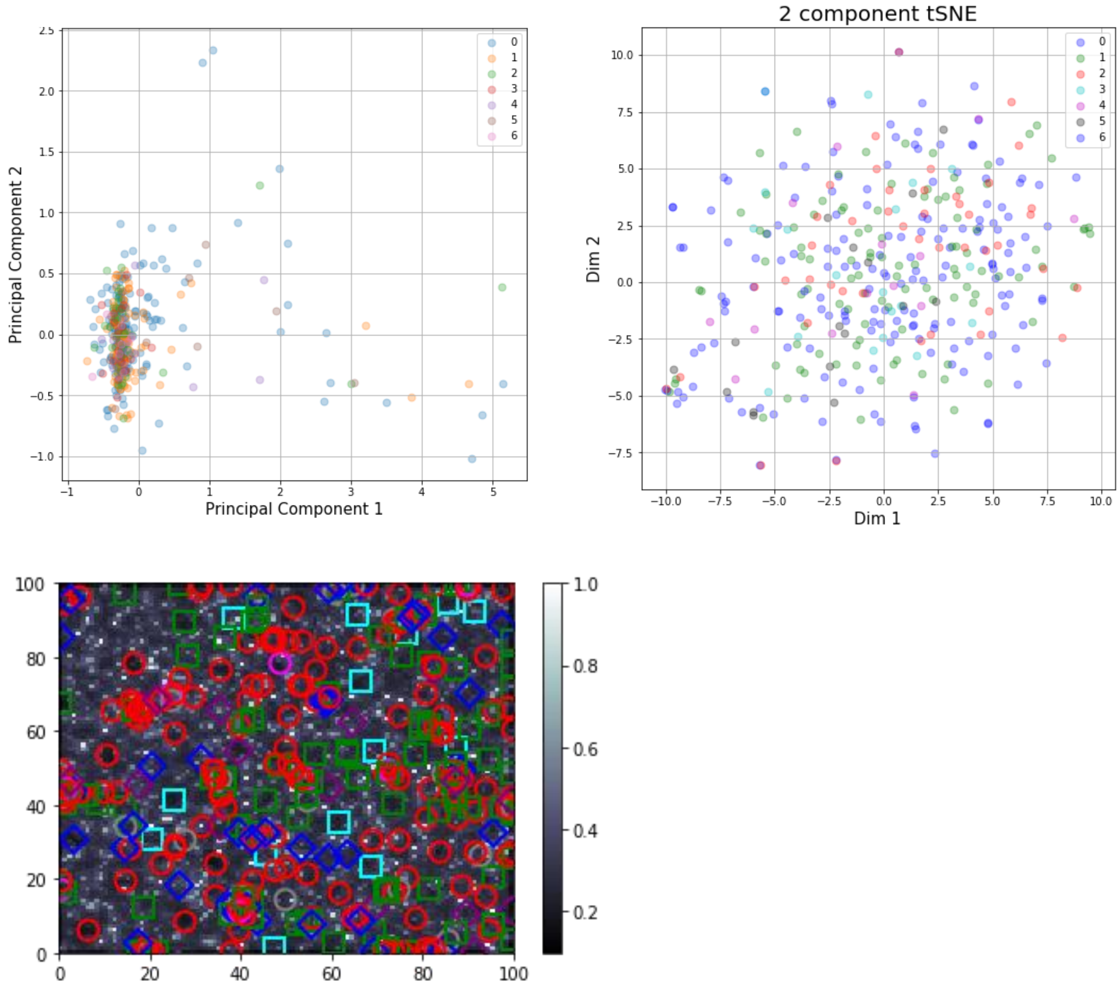


Figure. 5.4. Other preliminary unsupervised learning results.

With all described 5 annotations (labels), we tried to hierarchically cluster them and divide them into mutually exclusive groups, and then remove the number of groups that have less than 9 as the new labels. This resulted in 6 groups ready for supervised learning. However, the best accuracy obtained never goes over 50%. As for unsupervised learning methods on this subset of phenotype data, PCA (upper), t-SNE (middle) and self-organizing map (bottom) reveal not-easily separated nature, which is complementary to the supervised learning methods.

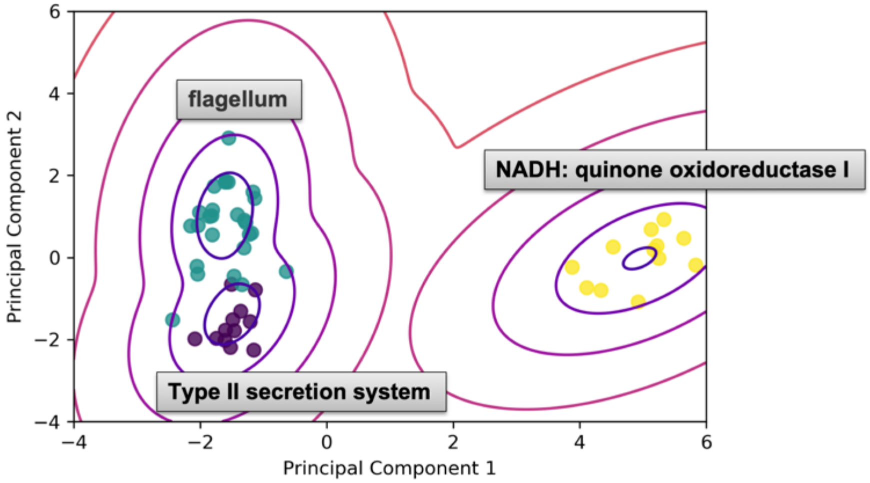


Figure 5.5. Selected protein complexes in phenotype-defined functional space, generated by Gaussian Mixture Model with Expectation Maximization.

Phenotype data naturally separate by labels, when heteromeric protein complexes of distinct functions are selected.

CHAPTER 6. MICROBIALPHENOTYPES: AN R PACKAGE THAT ANALYZES HIGH-THROUGHPUT MICROBIAL PHENOTYPE DATA

ABSTRACT

Various microbial high-throughput phenotyping techniques have been vastly conducted to infer functions of genes, generating large numbers of valuable datasets whose potential in providing insights to characterize genes hasn't been fully exploited. Therefore, computational tools that allow unbiased, systematic analysis of these data also have become vital. Here we describe a package that evaluates high-throughput microbial phenotype data by one or several sets of associated functional annotations are provided. In addition, some helper functions are provided to help clean high-throughput microbial phenotype data.

INTRODUCTION

Phenotypes play important roles in characterizing the functions of genes, particularly in microbiology, where the largest number of tests could be done much easier (Tohsato & Mori, 2008). With rising technologies (Kritikos et al., 2017; Nichols et al., 2011; Wetmore et al., 2015), high-throughput experimental approaches that measure large number of phenotypes under various conditions have flourished with complementary statistical methods in querying the behavior of gene products (Collins et al., 2006; Nichols et al., 2011; Price et al., 2018; Rishi et al., 2020). Although there are already many computational approaches written as R packages to analyze phenotype data (Deng et al., 2015; Vaas et al., 2013; Vuckovic et al., 2015) (Vehkala et al., 2015), software that quickly tests the potential of such high-throughput data in interpreting gene functions is lacking. Here we have wrapped the analytical pipeline of validating the phenotype data using annotation sets (P. I. F. Wu et al., 2021) into a package named microbialPhenotypes. In this package, we provide 8 functions that are specific in dealing with high-throughput microbial phenotype data, as well as 10 functions that are meant to be more supplementary. A high-throughput E. coli phenotype dataset (Nichols et al., 2011) is used for the examples provided below. We note that there is room for significant improvement when the analytical pipeline (P. I. F. Wu et al., 2021) is improved. Despite that the functions provided here started from a perspective of gaining biological insights for microbial phenotype data, the potential of using them as a tool for more general purposes should not be limited. If there are data from other research domains that has a similar structure to the example described here, the utility of this package can be much more extended. For example, data from animal cells (Alonezi et al., 2016, 2017) . In addition, knowing that there are resources for complicated machine learning algorithms to do classification of functions, our work here does not aim to improve those methods. Rather, it tries to quickly assess the usefulness of the newly generated high-throughput phenotype data before implementing more advanced classification schema.

INSTALLATION AND FUNCTIONS

The **MicrobialPhenotype** package can be downloaded from the github repository:

https://github.com/peterwu19881230/microbialPhenotypes

INPUT DATA

This package assumes that the input phenotypic profiles are in a text file (e.g., csv or tsv) in a matrix format where each row represents a mutant strain and each column is a growth condition where the phenotypes of the mutant strains are assayed. The values in the columns represent the phenotypes.

Users can read their data into R using either the read.csv() or read.table() function with the appropriate arguments based on the file format. See the example below:

Command:

```
> phenotype_data <-  read.csv(file="my_phenotype_profile.csv", header = TRUE)
> head(phenotype_data)
```

Sample output:

STRAIN	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5
ECK3997-purD	-1.48	-12.30	3.33	-13.46	-13.79
ECK0516-purE	0.43	-6.75	2.63	-8.27	-9.81
ECK3763-ilvD	-1.41	-11.72	0.34	-0.69	0.32
ECK3766-ilvC	-0.58	-8.53	-0.81	0.06	0.51
ECK3762-ilvE	-0.01	-11.93	0.02	-0.68	-0.46
ECK3764-ilvA	-0.15	-11.55	-0.84	-0.04	-0.19

*Cond.: condition

DISCRETIZE THE INPUT DATA

Quantitative data can be transformed into categorical data using either the BinaryConvert() or TernaryConvert() function, which produce output in a binary (0,1) or ternary (-1,0,1) form, respectively. The threshold argument is used to specify the phenotypic score cutoff used to select strains with a phenotype that is significantly different from that of the designated control, which is usually the phenotype of the wildtype. See examples, below:

Command:

```
> binary_phenotype <- binary_convert (matrix=phenotype_data, threshold =0.5)
> head(binary_phenotype)
```

Sample output:

STRAIN	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5
ECK3997-purD	1	1	1	1	1
ECK0516-purE	0	1	1	1	1
ECK3763-ilvD	1	1	0	1	0
ECK3766-ilvC	1	1	1	0	1
ECK3762-ilvE	0	1	0	1	0
ECK3764-ilvA	0	1	1	0	0

Command:

```
> ter_phenotype <- ternary_convert(matrix=phenotype_data, threshold =0.5)
> head(ter_phenotype)
```

Sample output:

STRAIN	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5
ECK3997-purD	-1	-1	1	-1	-1
ECK0516-purE	0	-1	1	-1	-1
ECK3763-ilvD	-1	-1	0	-1	0
ECK3766-ilvC	-1	-1	-1	0	1
ECK3762-ilvE	0	-1	0	-1	0
ECK3764-ilvA	0	-1	-1	0	0

CALCULATE SIMILARITIES/DISTANCES BETWEEN PHENOTYPIC PROFILES

The pairwise similarity/distance between phenotypic profiles can be calculated by a variety of functions, such as the Pearson Correlation Coefficient, Spearman Correlation Coefficient, Mutual Information. Users can also write self-defined functions. The function provided in this package is hamming distance. See the example below:

Command:

```
> hamming_dist=hamming_distance(head(ter_phenotype))
> hamming_dist
```

Sample output:

	ECK3997-purD	ECK0516-purE	ECK3763-ilvD	ECK3766-ilvC	ECK3762-ilvE	ECK3764-ilvA
ECK3997-purD	0	1	2	3	3	4
ECK0516-purE	1	0	3	4	2	3
ECK3763-ilvD	2	3	0	3	1	3
ECK3766-ilvC	3	4	3	0	4	2
ECK3762-ilvE	3	2	1	4	0	2
ECK3764-ilvA	4	3	3	2	2	0

Here we also demonstrate the computation of the Pearson-Correlation-Coefficient-based distance metric, which is commonly used in high-throughput microbial phenotype data:

Command:

```
> pearson_dist=1- cor(t(phenotype_data),method="pearson")

> pearson_dist
```

Sample output:

	ECK3997-purD	ECK0516-purE	ECK3763-ilvD	ECK3766-ilvC	ECK3762-ilvE	ECK3764-ilvA
ECK3997-purD	0	0.01	0.67	0.79	0.61	0.71
ECK0516-purE	0.01	0	0.78	0.89	0.71	0.80
ECK3763-ilvD	0.67	0.78	0	0.01	0.01	0.02
ECK3766-ilvC	0.79	0.89	0.02	0	0.02	0.01
ECK3762-ilvE	0.61	0.71	0.01	0.02	0	0.01
ECK3764-ilvA	0.71	0.80	0.02	0.01	0.01	0

PARSE GENE ANNOTATIONS

In order to link the functional annotations associated with a gene to the phenotypic profile of the strain where that gene is mutated, by calculating correlation, three functions, attr_list(), one_attr() and generate_pairs_similarity_coannotation (), are provided where:

i) attr_list() generates a list from a table where the relevant gene in each mutant strain is associated with its functional annotations.

To illustrate this, we will load a sample table with the name ‘name_attribute’, a 2-column table associating strain names and metabolic pathway annotations:

```
> load(“name_attribute”)
```

```
>  name_attribute
```

Sample output:

ids	Pwy
ECK3762-ilvE	ALANINE-VALINESYN-PWY
ECK3762-ilvE	THREOCAT-PWY
ECK3762-ilvE	ALL-CHORISMATE-PWY
ECK3762-ilvE	PWY0-1061
ECK3762-ilvE	PHESYN
ECK3762-ilvE	ILEUSYN-PWY
ECK3762-ilvE	LEUSYN-PWY
ECK3762-ilvE	VALSYN-PWY
ECK3762-ilvE	COMPLETE-ARO-PWY
ECK3762-ilvE	BRANCHED-CHAIN-AA-SYN-PWY
ECK3764-ilvA	THREOCAT-PWY
ECK3764-ilvA	ILEUSYN-PWY
ECK3764-ilvA	BRANCHED-CHAIN-AA-SYN-PWY
ECK3997-purD	PRPP-PWY
ECK3997-purD	DENOVOPURINE2-PWY
ECK3997-purD	PWY-6121
ECK3997-purD	PWY-6122
ECK3997-purD	PWY-6277
ECK0516-purE	PRPP-PWY
ECK0516-purE	PWY-6123
ECK0516-purE	DENOVOPURINE2-PWY
ECK3763-ilvD	THREOCAT-PWY
ECK3763-ilvD	ILEUSYN-PWY
ECK3763-ilvD	VALSYN-PWY

ECK3763-ilvD	BRANCHED-CHAIN-AA-SYN-PWY
ECK3766-ilvC	THREOCAT-PWY
ECK3766-ilvC	PANTO-PWY
ECK3766-ilvC	ILEUSYN-PWY
ECK3766-ilvC	VALSYN-PWY
ECK3766-ilvC	PANTOSYN-PWY
ECK3766-ilvC	BRANCHED-CHAIN-AA-SYN-PWY

```
> attr_list(name_attribute) #output is a list


$`ECK3762-ilvE`

"ALANINE-VALINESYN-PWY"  "THREOCAT-PWY"     "ALL-CHORISMATE-PWY"     "PWY0-1061"

"PHESYN"      "ILEUSYN-PWY"       "LEUSYN-PWY"        "VALSYN-PWY"        "COMPLETE-

ARO-PWY"      "BRANCHED-CHAIN-AA-SYN-PWY"


$`ECK3764-ilvA`

"THREOCAT-PWY"      "ILEUSYN-PWY"        "BRANCHED-CHAIN-AA-SYN-PWY"


$`ECK3997-purD`

"PRPP-PWY"     "DENOVOPURINE2-PWY"       "PWY-6121"      "PWY-6122"      "PWY-6277"


$`ECK0516-purE`

"PRPP-PWY"     "PWY-6123"      "DENOVOPURINE2-PWY"


$`ECK3763-ilvD`

"THREOCAT-PWY"      "ILEUSYN-PWY"        "VALSYN-PWY"         "BRANCHED-CHAIN-AA-SYN-

PWY"


$`ECK3766-ilvC`
```

```
"THREOCAT-PWY"       "PANTO-PWY"  "ILEUSYN-PWY"          "VALSYN-PWY"          "PANTOSYN-
PWY"    "BRANCHED-CHAIN-AA-SYN-PWY"
```

ii) one_attr() takes the output from attr_list() as the input. It generates a table that contains all possible combination of mutants and whether they share annotations (0 stands for not having any same annotations, 1 for having at least 1 same annotation). For example:

```
> attribute_list=attr_list(name_attribute)
> one_attr(attribute_list)
```

mutant1	mutant2	sameORnot
ECK3762-ilvE	ECK3764-ilvA	1
ECK3762-ilvE	ECK3997-purD	0
ECK3762-ilvE	ECK0516-purE	0
ECK3762-ilvE	ECK3763-ilvD	1
ECK3762-ilvE	ECK3766-ilvC	1
ECK3764-ilvA	ECK3997-purD	0
ECK3764-ilvA	ECK0516-purE	0
ECK3764-ilvA	ECK3763-ilvD	1
ECK3764-ilvA	ECK3766-ilvC	1
ECK3997-purD	ECK0516-purE	1
ECK3997-purD	ECK3763-ilvD	0
ECK3997-purD	ECK3766-ilvC	0
ECK0516-purE	ECK3763-ilvD	0
ECK0516-purE	ECK3766-ilvC	0
ECK3763-ilvD	ECK3766-ilvC	1

iii) generate_pairs_similarity_coannotation() generates the pairwise similarity table that contains the strain pairs, similarity/distance value and a Boolean column of whether the corresponding strain pairs share the same annotation(s). It takes phenotype data, the result from attr_list() and a function as an argument to specify the similarity metric. For example:

```
> attribute_list<-one_attr(name_attribute)

> names(attribute_list)<-rownames(phenotype_data)  #(Must do) Synchronize the strain names

> generate_pairs_similarity_coannotation(data= phenotype_data,attribute_list=attribute_list,

+ dist_metric=pcc_dist)
```

Sample output:

mutant1	mutant2	similarity	same_annot
ECK3762-ilvE	ECK3764-ilvA	0.01	1
ECK3766-ilvC	ECK3764-ilvA	0.01	1
ECK3997-purD	ECK0516-purE	0.01	1
ECK3763-ilvD	ECK3762-ilvE	0.01	1
ECK3763-ilvD	ECK3766-ilvC	0.02	1
ECK3763-ilvD	ECK3764-ilvA	0.02	0
ECK3766-ilvC	ECK3762-ilvE	0.02	0
ECK3997-purD	ECK3762-ilvE	0.61	0
ECK3997-purD	ECK3763-ilvD	0.67	0
ECK0516-purE	ECK3762-ilvE	0.71	0
ECK3997-purD	ECK3764-ilvA	0.71	0
ECK0516-purE	ECK3763-ilvD	0.78	0
ECK3997-purD	ECK3766-ilvC	0.79	0
ECK0516-purE	ECK3764-ilvA	0.80	0
ECK0516-purE	ECK3766-ilvC	0.89	0

, where pcc_dist = 1- cor(t(your_data),method="pearson")

GET METRICS DERIVED OF THE CONFUSION MATRIX

After the similarity and co-annotation columns are computed as above, get_confusionMatrix() can be used to get the similarity-based confusion matrices and the derived metrics: sensitivity, specificity, precision, and accuracy. get_confusionMatrix() also permutate the co-annotation column and calculate the confusion matrices and other derived metrics as negative controls. For example:

```
> new <- generate_pairs_similarity_coannotation(data= phenotype_data,
attribute_list=attribute_list, dist_metric=pcc_dist)
```

```
> get_confusionMatrix_and_metrics (df=new, annot="same_annot",similarity="similarity")
```

Sample output :

(result columns are separated by 2 tables)

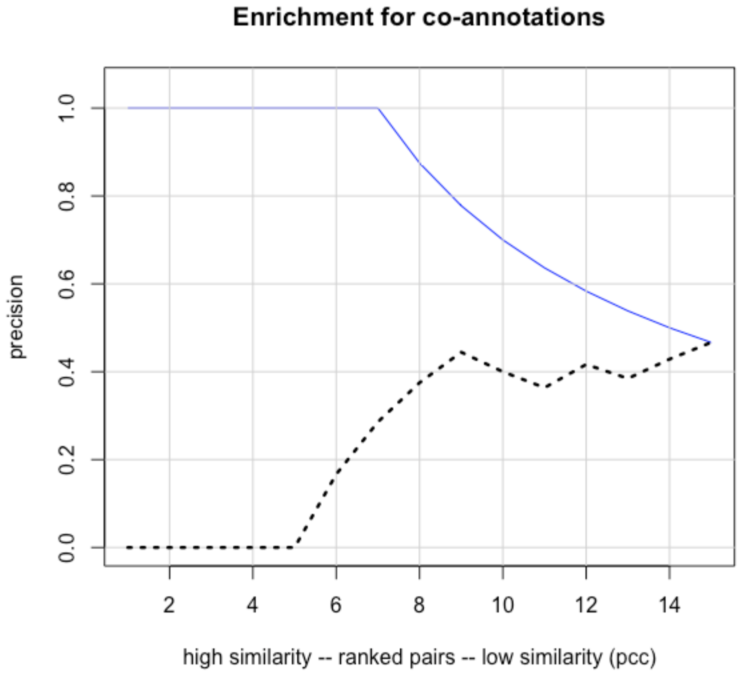
TP	FN	TN	FN	sensitivity	specificity	precision	accuracy
1	0	8	6	0.14	1	1	0.60
2	0	8	5	0.29	1	1	0.67
3	0	8	4	0.43	1	1	0.73
4	0	8	3	0.57	1	1	0.80
5	0	8	2	0.71	1	1	0.87
6	0	8	1	0.86	1	1	0.93
7	0	8	0	1	1	1	1
7	1	7	0	1	0.88	0.88	0.93
7	2	6	0	1	0.75	0.78	0.87
7	3	5	0	1	0.63	0.70	0.80
7	4	4	0	1	0.50	0.64	0.73
7	5	3	0	1	0.38	0.58	0.67
7	6	2	0	1	0.25	0.54	0.60
7	7	1	0	1	0.13	0.50	0.53
7	8	0	0	1	0	0.47	0.47

random_ TP	random_ FP	random_ TN	random_ FN	random_ sensitivity	random_ specificity	random_ precision	random_ accuracy
0	1	7	7	0	0.88	0	0.47
0	2	6	7	0	0.75	0	0.40
0	3	5	7	0	0.63	0	0.33
0	4	4	7	0	0.50	0	0.27
0	5	3	7	0	0.38	0	0.20
1	5	3	6	0.14	0.38	0.17	0.27
2	5	3	5	0.29	0.38	0.29	0.33
3	5	3	4	0.43	0.38	0.38	0.40
4	5	3	3	0.57	0.38	0.44	0.47
4	6	2	3	0.57	0.25	0.40	0.40
4	7	1	3	0.57	0.13	0.36	0.33
5	7	1	2	0.71	0.13	0.42	0.40
5	8	0	2	0.71	0	0.38	0.33
6	8	0	1	0.86	0	0.43	0.40
7	8	0	0	1	0	0.47	0.47

PLOT THE RESULTS

graph_corr_annot () takes the output from get_confusionMatrix() to plot the final result, where precisions were plotted against the ranked pairs of mutants. Enrichment for sensitivity, specificity, precision, and accuracy could be compared with the dotted line, which is the negative control. For example:

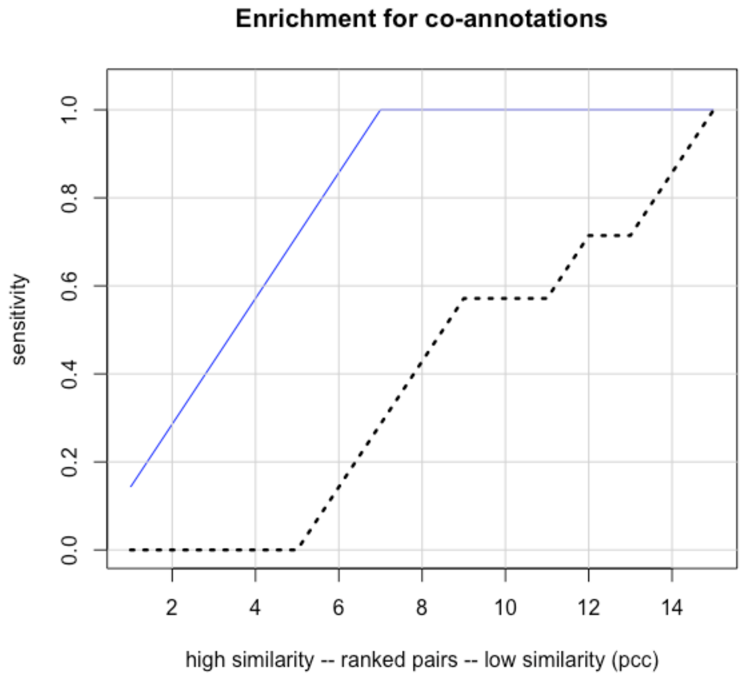
```
> confusionMatrix_obj <- get_confusionMatrix_and_metrics
(new,"same_annot","similarity",seed=103)
> metric="precision"; similarity="pcc"; subset=dim(confusionMatrix_obj)[1]; cols="blue";
ylim=c(0,1); lwd=1 # set graphing parameters
> graph_corr_annot(confusionMatrix_obj, metric, similarity, subset, cols, x_lab="",
ylim=c(0,1.05), lwd)
```



```
> metric="sensitivity" # use another metric for the y axis

> graph_corr_annot(confusionMatrix_obj, metric, similarity, subset, cols, x_lab="",

ylim=c(0,1.05), lwd)
```



HELPER FUNCTIONS

The following 8 functions are intended to speed up the phenotype data curation process:

i) any_incomplete() checks if the input matrix, data frame or data table contains any NA, NAN, NULL or "" (empty string). For example:

```
> incomplete_phenotype_data<- phenotype_data

> incomplete_phenotype_data[1:2, 1:2]<- NA   #introduce some NA

> incomplete_phenotype_data
```

STRAIN	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5
ECK3997-purD	-1.48	-12.30	3.33	-13.46	-13.79
ECK0516-purE	0.43	-6.75	2.63	-8.27	-9.81
ECK3763-ilvD	-1.41	-11.72	0.34	-0.69	0.32

ECK3766-ilvC	-0.58	-8.53	-0.81	0.06	0.51
ECK3762-ilvE	-0.01	-11.93	0.02	-0.68	-0.46
ECK3764-ilvA	-0.15	-11.55	-0.84	-0.04	-0.19

```
> any_incomplete(incomplete_phenotype_data)


$dimension

[1] "Dimension: 6 rows * 5 columns"



$na

STREPTOMYCIN.0.05        SUCCINATE

          2                         2



$null

named integer(0)



$nan

named integer(0)



$empty

named integer(0)



$completeness
```

```
[1] "4 of NA, NAN, NULL, or empty character is found from 30 data points. They constitute 13.3333333333333%"
```

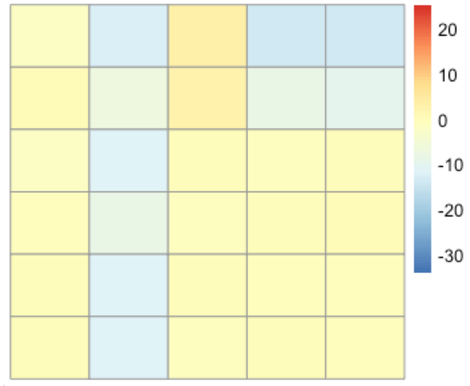
ii) filter_table() filters the input matrix, dataframe or datatable so that all rows and columns with NA/NAN/NULL/"" are removed. For example:

```
> filter_table(incomplete_phenotype_data)
```

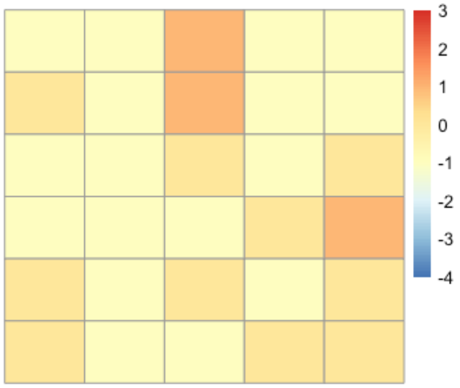
STRAIN	Cond. 1	Cond. 2	Cond. 3
ECK3997-purD	3.33	-13.46	-13.79
ECK0516-purE	2.63	-8.27	-9.81
ECK3763-ilvD	0.34	-0.69	0.32
ECK3766-ilvC	-0.81	0.06	0.51
ECK3762-ilvE	0.02	-0.68	-0.46
ECK3764-ilvA	-0.84	-0.04	-0.19

iii) graph_table() takes a matrix, dataframe or a table as the input and represents it using a heatmap. It also deals with continuous/categorical/mixed variables. For example:

```
>  graph_table(phenotype_data)
```



```
> graph_table(ter_phenotype)
```



iv) checkDuplicates_vect() checks if an input vector has duplicates. If so, it will return the frequency table. Otherwise, the text “Everything in this vector is unique” will be returned. For example:

```
> my_vector <- c(1,1,2,2,3)
> checkDuplicates_vect(my_vector)
[1] "Some duplicates are found:"
```

```
vect
1 2 3
2 2 1
```

```
> my_vector <- c(1,2,3)
> checkDuplicates_vect(my_vector)
[1] "Everything in this vector is unique"
```

v) change_names () changes row names or column names of a matrix, dataframe or datatable based on another matrix/dataframe/datatable. For example:

```
> new_col_names <-c("0.05 µg/ml streptomycin",
"0.3% succinate",
"100 µg/ml sulfamonomethoxine","0.1% taurocholate","0.5% taurocholate ")
```

```
> change_names(rowOrCol="col", phenotype_data,
 matrix(c(colnames(phenotype_data), new_col_names), ncol=2, byrow=FALSE))
```

STRAIN	**0.05 µg/ml streptomycin**	**0.3% succinate**	**100 µg/ml sulfamonomethoxine**	**0.1% taurocholate**	**0.5% taurocholate**
ECK3997-purD	-1.48	-12.30	3.33	-13.46	-13.79
ECK0516-purE	0.43	-6.75	2.63	-8.27	-9.81
ECK3763-ilvD	-1.41	-11.72	0.34	-0.69	0.32
ECK3766-ilvC	-0.58	-8.53	-0.81	0.06	0.51
ECK3762-ilvE	-0.01	-11.93	0.02	-0.68	-0.46
ECK3764-ilvA	-0.15	-11.55	-0.84	-0.04	-0.19

vi) melt_similarity() takes a similarity matrix or a distance object as the input and converts it to a long form dataframe, with an option to sort the molten dataframe by the 3rd column (the numeric column). For example:

```
> dist(phenotype_data) # the distance object of interest
```

STRAIN	ECK3997-purD	ECK0516-purE	ECK3763-ilvD	ECK3766-ilvC	ECK3762-ilvE
ECK0516-purE	8.82				
ECK3763-ilvD	19.27	13.91			
ECK3766-ilvC	20.48	13.85	3.58		
ECK3762-ilvE	18.82	13.37	1.65	3.75	
ECK3764-ilvA	19.62	13.99	1.92	3.13	1.18

```
> melt_similarity(dist(phenotype_data))
```

object_1	object_2	value

ECK3997-purD	ECK0516-purE	8.82
ECK3997-purD	ECK3763-ilvD	19.27
ECK3997-purD	ECK3766-ilvC	20.48
ECK3997-purD	ECK3762-ilvE	18.82
ECK3997-purD	ECK3764-ilvA	19.62
ECK0516-purE	ECK3763-ilvD	13.91
ECK0516-purE	ECK3766-ilvC	13.85
ECK0516-purE	ECK3762-ilvE	13.37
ECK0516-purE	ECK3764-ilvA	13.99
ECK3763-ilvD	ECK3766-ilvC	3.58
ECK3763-ilvD	ECK3762-ilvE	1.65
ECK3763-ilvD	ECK3764-ilvA	1.92
ECK3766-ilvC	ECK3762-ilvE	3.75
ECK3766-ilvC	ECK3764-ilvA	3.13
ECK3762-ilvE	ECK3764-ilvA	1.18

When compared with reshape2::melt(), the differences are: 1. melt_similarity() Can take a distance object as the main input 2. When a matrix is used as the main input, it has to be a similarity matrix 3. melt_similarity() remove the diagonal elements and the duplicated pairs that share the same similarity.

vii) convert_table() takes a matrix, dataframe or a datatable as the input and converts the types of elements to a designated type. For example:

```
> str(phenotype_data)
```

STREPTOMYCIN.0.05	-1.48	0.43	-1.41	-0.58
SUCCINATE	-12.3	-6.75	-11.72	-8.53

SULFAMONOMETHOXINE.100	3.33	2.63	0.34	-0.81
TAUROCHOLATE.0.1	-13.46	-8.27	-0.69	0.06
TAUROCHOLATE.0.5	-13.79	-9.81	0.32	0.51

```
> phenotype_data_chr <- convert_table(phenotype_data, as.character)

> str(phenotype_data_chr)
```

Sample output:

STREPTOMYCIN.0.05	"-1.48"	"0.43"	"-1.41"	"-0.58"
SUCCINATE	"-12.3"	"-6.75"	"-11.72"	"-8.53"
SULFAMONOMETHOXINE.100	"3.33"	"2.63"	"0.34"	"-0.81"
TAUROCHOLATE.0.1	"-13.46"	"-8.27"	"-0.69"	"0.06"
TAUROCHOLATE.0.5	"-13.79"	"-9.81"	"0.32"	"0.51"

viii) remove_NA() removes NA from an R vector object. For example:

```
> remove_NA (c(1,2,3,4,NA,NA))

[1] 1 2 3 4
```

SUMMARY

The MicrobialPhenotypes package is a pipeline to systematically parse and analyze the data produced by high-throughput phenotypic screens, with many functions as well as an example using a published *E. coli* dataset (Nichols et al., 2011). Although the motivation for this software is to process microbial phenotype data, we expect its usability to be easily extended to multivariate dataset with distinct annotation sets.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health [R01GM089636, U41HG008735]. PW wrote the package and drafted the manuscript. DS, JH supervised the project as well as testing the usability of the code.

CHAPTER 7: CONCLUSIONS, DISCUSSION AND FUTURE DIRECTIONS

Systematically measuring phenotypes on a large scale and using them to gain understanding into gene functions is not trivial. Even with well-studied and well annotated microorganisms, such as *Escherichia coli* K-12 or *Saccharomyces cerevisiae*, the amount of work to integrate and standardize existing phenotype data is expected to take a lot of time and effort. Expanding this effort to phenotypes of organisms across the domains of life will be challenging. However, unbiased, systematic reanalyses of high-throughput phenotypic profiles might help to improve and standardize strategies for extracting useful information, given that as new analytical strategies come out, it will be more obvious which kind of structured data is more useful. In addition, once the ability to predict function from phenotypic profiles is significantly improved, there could be many new hypotheses generated that may be helpful in guiding future biochemical and genetic studies of function.

There are many publicly available phenotypic datasets for model organisms such as *E. coli and S. cerevisiae* that could be made interoperable across species. I have located datasets that contain large numbers of phenotypes, curated them and systematically reanalyzed them. In Chapter 2, I report that a strong, genome-wide association between phenotypic profile similarity and functional similarity is found after systematic analysis of one of the curated datasets. There does not seem to be an obvious gold-standard method for calculating phenotypic profile similarity; association between phenotype and function was detected using three different similarity metrics: Pearson Correlation Coefficient (PCC), Mutual Information (MI) and Spearman's Rank

Correlation Coefficient (SRCC). Accumulation of new phenotype data and functional annotations are expected to improve prediction of functions. In addition, I have shown that at least some of the association between phenotype similarity and functional similarity is retained after converting quantitative phenotypes to qualitative, discretized phenotypes. This indicates that discretizing phenotype scores might be a good approach for integrating information from different experimental studies. I repeated the strategy to look for an association between phenotypic profiles and functions with a different dataset. The results are presented in Chapter 3, where I have demonstrated that analysis of a second high-throughput phenotypic profile dataset obtained from competitive growth assays with pools of knockout mutants also showed a strong correlation between phenotypic similarity and functional similarity. In addition, I found a strong association between phenotype similarity and functional similarity even when phenotype similarity was determined from annotations made using the Ontology of Microbial Phenotypes (OMP). This result indicates that OMP annotations can be used to integrate phenotype information from different studies, which may lead to the identification of new functional connections. In Chapter 4, it is reported that when Multiple Correspondence Analysis (MCA) is applied, the essential phenotypic dimensions from cell-cycle related genes can be determined. In addition, correlation between phenotypic similarity and functional similarity was also determined by using binary cell-cycle related phenotypes with GO annotations. More generally, as described in Chapter 5, many supervised or unsupervised machine learning methods can be directly applied to high-throughput phenotypic profiles, provided that the number of labels (the functional annotations) for training is adequate. Lastly, Chapter 6 provides potentially useful functions written in R for analyzing high-throughput phenotypic profiles. In summary, phenotypic profile similarity highly and systematically associates with functional similarity where the functional annotations mostly made by manual curation from biocurators. I have also shown that it is possible to use some classic machine learning solutions to predict functions of genes, if there are enough training samples of phenotypic profiles. The written software performing the analyses are made publicly online for future analyses on high-throughput phenotypic profiles.

As reported in chapters 2 and 3, we observed that gene pairs enriched for phenotypic similarity have high functional similarity, and vice versa. However, not every gene pair sharing one or more functional annotations had enriched phenotypic profile similarity. For example, approximately 25% of co-annotated gene pairs did not share enriched phenotypic profile similarity when the Nichols dataset (Nichols et al., 2011) and pathway annotations were used. There are a variety of reasons why this might occur. Some genes may be part of more than one functional pathway and consequently would have a different phenotypic profile than genes that are involved in only a single pathway. Another possible explanation is that the mutation in one of the strains affects the function or expression of another gene or genes. In addition, some genes may not show a mutant phenotype under the growth conditions examined in a particular study. When seeking to associate phenotypic profiles with gene functions, it is typically assumed that only a single gene is altered in each of the mutant strains being studied, and, therefore, the phenotypes observed for a particular strain can be attributed to alteration of that specific gene. However, this assumption is not always correct. If the genome annotation is inaccurate or incomplete, a mutation designed to disrupt one gene may have unexpected effects if there is an unidentified coding sequence or regulatory region overlapping with the targeted gene. While we envision that single mutants are still going to be one of the major sources for phenotypic profiling, because of the availability of current libraries and the ease of interpreting the results in a gene-centric way, it is possible that once the scalability of profiling phenotypes increases, more systematic approaches that use strains with double mutations or other more complicated genotypes can come into place.

Even for a relatively well-studied bacterium such as *E. coli*, approximately one-third of the genes have unknown functions (Price et al., 2018). This indicates that there is a large room for many computational functional predictions to take place. Since causal relationship can be assumed between phenotypes and functions, computational and statistical approaches that accompany biochemical/genetics experiments predicting functions from phenotypes or, in reverse, predicting phenotypes from functions would both have the potential in providing new biological insights.

There are a number of studies that aim to predict specific phenotypes from genotype information, such as the following studies (Guzzetta et al., 2010; Lees et al., 2020; Mahfouz et al., 2020; Stoesser et al., 2013; Tang et al., 2020), where these approaches were based on the input of multiple single nucleotide polymorphism sites (SNPs), sets of loci associated with phenotypes, genomic sequence homology, or antimicrobial resistance markers. For microorganisms in specific, genome sequencing or multi-omics approaches have been used to predict phenotypes. For example, Karr *et al.* (Karr et al., 2012) used DNA-seq, RNA-seq and other molecular methods to compute a whole-cell model of the bacterium *Mycoplasma genitalium*, an intracellular human pathogen with only 500 genes; Stoesser *et al.* (Stoesser et al., 2013) compiled a reference gene database from public resources and used more than 100 known antibiotic resistant markers to predict the susceptibilities of 74 *E. coli* and 69 *K. pneumoniae* isolates to seven antimicrobial compounds; Feldbauer *et al*. (Feldbauer et al., 2015) describe a comparative genomics approach using Clusters of Orthologous Groups (COG) to predict whether a bacterium is obligate intracellular, facultative intracellular or free living; Lees *et al*. (Lees et al., 2020) used elastic regression methods to predict phenotypes based on pangenome data; and Aun *et al*. (Aun et al., 2018) developed a k-mer based prediction algorithm for classifying phenotypes based on genomic sequence of bacterial isolates. However, computationally predicting gene functions from large-scale phenotype observations in a systematic way has been much less developed.

When high-throughput phenotypic profiles are available, identifying the phenotypes most relevant to particular functions is essential, not only because it helps to build models for predicting functions, but also relate new connections from the key phenotypes with a specific function. There are abundant sources of high-throughput phenotype data available for model microorganisms, obtained by observing growth under diverse conditions, which indicates that there are probably more than enough phenotypic variables, i.e. different phenotypes, that can be used to assign functions to genes, whether through the "guilt by association" approach, which attributes functions to genes based on calculating phenotypic profile similarity, or through machine learning methods that identify features (phenotypic variables) by model training and testing. It should be noted that defining the universe of phenotypes that can account for the universe of functions is not trivial.

Different metrics can be used to calculate phenotypic profile similarity for the “guilt by association” approach. The Pearson Correlation Coefficient (PCC) is useful for quantitative data. Spearman’s Rank Correlation Coefficient (SRCC) uses the same formula as PCC except that it uses the ranking of the variables as input rather than the original values. This allows SRCC to capture non-linear correlations and it is not affected by outliers as PCC is (Schober et al., 2018). Another phenotypic similarity metric that was tested, Mutual Information, works well for associating discrete and sparse phenotypic profiles, based on the entropy and mutual dependence of the input data.

For unsupervised machine learning methods, the availability of large numbers of phenotypes to use as variables will allow the separation of functions by dimensional reduction. For supervised machine learning methods, having a large number of phenotypes will allow automatic selection of the key phenotypes, or selection for dependence of multiple phenotypes. However, in order to build effective classification models to predict functions, training samples with mutually exclusive labels are needed. In many cases, a metabolic pathway, complex or operon only has a limited number of gene members, e.g., three to five genes, which is inadequate to build any strong supervised classification model. If a larger number of genes within and without a functional class can be identified, it may be possible to build a classification model and identify the key phenotypes. Since there may be many correlated conditions within a study, modified regression methods like ridge regression, LASSO or elastic net may be helpful in identifying the key phenotypes that are associated with particular functions, because they tend to shrink or delete relatively insignificant coefficients of variables, or remove redundant variables (Tibshirani, 1996; Zou & Hastie, 2005). To achieve higher prediction accuracy, ensemble learning methods or neural network may be used, because, empirically, these models tend to give better overall scores, although they sometimes have limited interpretability (Carvalho et al., 2019).

A frequent method for obtaining high-throughput phenotypes information is to use a collection of mutants whose phenotypes are measured under many different growth conditions. There are many commonly tested conditions including nutrient sources like carbon, nitrogen and phosphorus; stresses like antibiotics, drugs and chemicals. These conditions are particularly useful because they are known to be related to many existing functions. However, in terms of predicting a certain function, not all the tested conditions need to be used when a portion of the key conditions predictive of that function is identified.

There are many ways to measure phenotypes. Of the three high-throughput phenotypic profiles analyzed in this work, two measured fitness during growth under specific conditions, while the third study measured morphological features derived from microscopic images. Image-based phenotypes can be measured at the population level (Nichols et al., 2011) or on single cells (Bougen-Zhukov et al., 2017). Kritikos *et al*. describes the Iris software (Kritikos et al., 2017), which can computationally capture three-dimensional morphological qualities from two-dimensional images of colonies. Image-based phenotypes, if measured in time series, can be even more informative for elucidating functions (Zahir et al., 2019), because multiple stages of phenotypes are observed. In addition, there have also been approaches to differentiate phenotypes based on mass cytometry (Georgopoulou et al., 2021).

It is not yet clear how much phenotype data will be required to form stronger classification or clustering models to predict the functions of genes across many microbial species. Methods and scalability of collecting phenotypic profile data for different species can vary significantly. In addition, there are different levels of interest in different microorganisms, resulting in imbalanced amounts of available data. Even for *E. coli* and *S. cerevisiae*, which are the species that have the largest quantities of phenotype data, measuring new phenotypes is still expected to be useful for interpreting gene functions, given the number of genes in these organisms whose functions are still not completely understood.

Systematically collecting and maintaining sufficient phenotype data to predict functions for many different organisms is very challenging. Thessen *et al.* (Thessen, Walls, et al., 2020) summarize some of the difficulties: i. The names of phenotypes are inconsistent. Sometimes, one entity will have several phenotypic descriptions, while the same phenotype term may be used to describe several different entities; ii. Definitions of phenotypes may change over time; iii. The process of recording the same phenotypes from different sources are very different, resulting in non-interoperable data; iv. It is sometimes unclear whether a phenotype came from an individual organism or from a population; v. The definitions of species to which phenotype data are attached may change. Use of phenotype ontology, such as OMP, can help to overcome some of these problems. Because OMP is designed to be used for many microorganisms, including bacteria, archaea, fungi, and viruses, it aims to minimize ambiguity and maximize interoperability of phenotypic descriptions. In OMP annotations, phenotypes are associated with specified genotypes, which makes species annotation immune to taxonomic change. There are separate OMP terms for phenotypes observed at the level of an individual cell and phenotypes observed for a population of cells.

Using evidence codes as part of annotations is important (Giglio et al., 2019) not only for phenotype annotations but also for the field of biocuration as a whole, because it gives additional support to the conclusion embedded in the annotation. Specifically, the evidence code indicates the type of experiment or assay the annotation is based on. It also gives a sense of confidence for the annotations made. Evidence codes from the Evidence Code Ontology (ECO), which is used by the Gene Ontology and other ontologies associated with the OBO-Foundry, also indicate whether an annotation was made manually or was assigned automatically without review by a biocurator (Giglio et al., 2019). By using evidence codes from ECO, it is easily seen whether an annotation came from direct experimental results with manual assertion, a high-throughput assay, or was generated entirely by computational approaches. In Bastian *et al.*, (Bastian et al., 2015), they even describe a tailored controlled vocabulary, The Confidence Information Ontology (CIO), whose terms can be used to indicate the level of confidence in the assertion being made by an annotation based on the experimental evidence. In addition to incorporating terms from the CIO as part of individual annotations, the CIO contains terms that can be used in summary annotations to report, for example, that an assertion is supported by experimental results from multiple types of experiments, which would indicate a high level of confidence.

Currently the most debated type of annotations are annotations with an evidence code indicating there was minimal curator's effort (Škunca et al., 2017). Examples of this type of evidence code are Inferred from Sequence Orthology (ISO) or Inferred from Sequence Alignment (ISA). Since sequencing technology has become routine, the propagation of computational annotations based on sequence similarity has its own limitations. For example, a pair of very similar sequences might encode proteins of completely different functions, because of point mutations or frameshift mutations. Although computational annotations are not the most favored type of annotation, their quality has been increasing over the years, and including automatic annotations has been shown to increase specificity, reliability and coverage of gene functions (Škunca et al., 2012). In Chapter 2, where I report the results of using GO biological process annotations to associate phenotypic profiles with gene functions, no significant difference was found whether automatic annotations (annotations with the Inferred from Electronic Annotation evidence code, now obsolete) were included or excluded, showing that at least in this analysis including automatic annotations was neutral. One way to determine how much computational annotations have improved would be to go through many history versions of the annotation database with the current version, repeat the experiments describe in chapter 2 and 3, correlate the annotation similarity with existing phenotypic profile similarity, and see if the precision increases in general.

There are many tools and advanced prediction algorithms that can predict functions. However, many of these predict only one or a few specific functions (Anahtar et al., 2021; Van Camp et al., 2020; Yang et al., 2019). What is needed are tools that can predict, for example, complete metabolic pathways or protein complexes. It is worth noting that no matter how advanced the algorithms for predicting gene function are, they can "get more from something", but cannot "get anything from nothing". For example, if there is an orphan gene whose function is never experimentally characterized, and thus does not exist in the functional annotation pool, there will never be a tool that can magically identify that new function. Therefore, the work presented in this dissertation is limited to the available amount of structured phenotype data, and the completeness of major functional annotations to train reliable prediction models.

As discussed in Chapter 1 and Chapter 5, there are many machine learning methods already available that can be used to predict gene functions. These methods fall into two major categories: supervised methods, where labels are required for each sample, e.g. a phenotypic profile of a mutant, and unsupervised methods, where there is no need to label the samples. Given enough samples and a large enough number of phenotypes as variables, many machine learning models will be able to select key interacting features (phenotypes) associated with a function; something that would be difficult to do by other methods. However, it is also important to avoid the problem of overfitting, as mentioned in Teschendorff *et al*. (Teschendorff, 2019). If a prediction model has overfit the training data, it will very likely fail to give accurate predictions for future data, thus invalidating the effort. A key reminder to ensure the robustness of any model is to make sure there is always an unseen portion of data left out of the training data that can be used to best estimate the errors of the model.

Envisioning the integration of different microbial phenotypes, OMP might be a powerful base tool. To expand the power of OMP, many more phenotype annotations using the ontology need to be brought into the database in a timely manner. The majority of annotations currently in the OMP database are for three model organisms: *Escherichia coli*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. Adding additional annotations for these organisms and adding annotations for other microbes will increase the usefulness of the Microbial Phenotypes Wiki as a resource for microbial phenotype information.

A direction that can potentially bring OMP to medical applications is to annotate phenotypes from pathogenic microorganisms. One example of a phenotypic dataset that could be annotated and incorporated into the OMP database is Dragset *et al*. (Dragset et al., 2019), which describes genome-wide phenotypic profiling of *M. tuberculosis*, the pathogen that causes tuberculosis. A number of pathogens have been studied using Biolog phenotype microarrays (Bochner et al., 2001; Mackie et al., 2014). References for these papers can be found on the Biolog website: https://www.biolog.com/support/bibliography/.

Most of the phenotype curation effort using OMP has been done by members of the OMP group. The Microbial Phenotypes Wiki (https://microbialphenotypes.org) is the repository for displaying these annotations. In order to efficiently curate more phenotype data to provide timely help for microbiologists, I propose the following: i. Prioritize the curation of data from papers that have the highest number of phenotypes, or papers that contain phenotypes that provide the most functional insights; ii. Reframe the front-end user interface to a more modern design; iii.

Establish a batch phenotype submission system following the principles described in making annotations (Siegele et al., 2019); iv. Adopt a curation effort estimator to monitor the productivity of the biocuration process. (Rodriguez-Esteban, 2015).

Overall, the major future directions the OMP project is facing can be divided into 3 main aspects: i. The biocuration perspective: the need to curate large amounts of phenotype data in a timely manner; ii. The engineering perspective: To facilitate increased usage of the OMP system, a better website framework is needed; iii. The Data Science perspective: The phenotype data captured with the ontology will be much more powerful when advanced Biostatistics/Artificial intelligence methods are applied. Hopefully, with the results in this dissertation, more curated microbial phenotypic profiles coming in the future, and the OMP annotation platform, my established analytical pipeline can be extended, and would thus co-evolve with future phenotype data and lead to improvement in predicting functions.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433-459. https://doi.org/https://doi.org/10.1002/wics.101

Abernathy, M. H., Yu, J., Ma, F., Liberton, M., Ungerer, J., Hollinshead, W. D., Gopalakrishnan, S., He, L., Maranas, C. D., Pakrasi, H. B., Allen, D. K., & Tang, Y. J. (2017). Deciphering cyanobacterial phenotypes for fast photoautotrophic growth via isotopically nonstationary metabolic flux analysis. *Biotechnol Biofuels*, *10*, 273. https://doi.org/10.1186/s13068-017-0958-y

Acin-Albiac, M., Filannino, P., Gobbetti, M., & Di Cagno, R. (2020). Microbial high throughput phenomics: The potential of an irreplaceable omics. *Comput Struct Biotechnol J*, *18*, 2290-2299. https://doi.org/10.1016/j.csbj.2020.08.010

Akman, O., Comar, T., Hrozencik, D., & Gonzales, J. (2019). Chapter 11 - Data Clustering and Self-Organizing Maps in Biology. In R. Robeva & M. Macauley (Eds.), *Algebraic and Combinatorial Computational Biology* (pp. 351-374). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-814066-6.00011-8

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, 21-23 Aug. 2017). Understanding of a convolutional neural network. 2017 International Conference on Engineering and Technology (ICET),

Alfred, S. E., Surendra, A., Le, C., Lin, K., Mok, A., Wallace, I. M., Proctor, M., Urbanus, M. L., Giaever, G., & Nislow, C. (2012). A phenotypic screening platform to identify small molecule modulators of Chlamydomonas reinhardtii growth, motility and photosynthesis. *Genome Biol*, *13*(11), R105. https://doi.org/10.1186/gb-2012-13-11-r105

Alonezi, S., Tusiimire, J., Wallace, J., Dufton, M. J., Parkinson, J. A., Young, L. C., Clements, C. J., Park, J. K., Jeon, J. W., Ferro, V. A., & Watson, D. G. (2016). Metabolomic profiling of the effects of melittin on cisplatin resistant and cisplatin sensitive ovarian cancer cells using mass spectrometry and biolog microarray technology. *Metabolites*, *6*(4). https://doi.org/10.3390/metabo6040035

Alonezi, S., Tusiimire, J., Wallace, J., Dufton, M. J., Parkinson, J. A., Young, L. C., Clements, C. J., Park, J. K., Jeon, J. W., Ferro, V. A., & Watson, D. G. (2017). Metabolomic profiling of the synergistic effects of melittin in combination with cisplatin on ovarian cancer cells. *Metabolites*, *7*(2). https://doi.org/10.3390/metabo7020014

Anahtar, M. N., Yang, J. H., & Kanjilal, S. (2021). Applications of machine learning to the problem of antimicrobial resistance: An emerging model for translational research. *J Clin Microbiol*. https://doi.org/10.1128/JCM.01260-20

Anderson, R. P., Jin, R., & Grunkemeier, G. L. (2003). Understanding logistic regression analysis in clinical reports: an introduction. *Ann Thorac Surg*, *75*(3), 753-757. https://doi.org/10.1016/s0003-4975(02)04683-0

Arnoldo, A., Kittanakom, S., Heisler, L. E., Mak, A. B., Shukalyuk, A. I., Torti, D., Moffat, J., Giaever, G., & Nislow, C. (2014). A genome scale overexpression screen to reveal drug activity in human cells. *Genome Med*, *6*(4), 32. https://doi.org/10.1186/gm549

Ascensao, J. A., Dolan, M. E., Hill, D. P., & Blake, J. A. (2014). Methodology for the inference of gene function from phenotype data. *BMC Bioinformatics*, *15*, 405. https://doi.org/10.1186/s12859-014-0405-z

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, *25*(1), 25-29. https://doi.org/10.1038/75556

Aun, E., Brauer, A., Kisand, V., Tenson, T., & Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput Biol*, *14*(10), e1006434. https://doi.org/10.1371/journal.pcbi.1006434

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, *2*, 2006 0008. https://doi.org/10.1038/msb4100050

Babajide Mustapha, I., & Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. *Molecules*, *21*(8). https://doi.org/10.3390/molecules21080983

Bastian, F. B., Chibucos, M. C., Gaudet, P., Giglio, M., Holliday, G. L., Huang, H., Lewis, S. E., Niknejad, A., Orchard, S., Poux, S., Skunca, N., & Robinson-Rechavi, M. (2015). The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. *Database (Oxford)*, *2015*, bav043. https://doi.org/10.1093/database/bav043

Bermudez, R. M., Wu, P. I., Callerame, D., Hammer, S., Hu, J. C., & Polymenis, M. (2020). Phenotypic associations among cell cycle genes in Saccharomyces cerevisiae. *G3 (Bethesda)*, *10*(7), 2345-2351. https://doi.org/10.1534/g3.120.401350

Biocuration, I. S. f. (2018). Biocuration: Distilling data into knowledge. *PLoS Biol*, *16*(4), e2002846. https://doi.org/10.1371/journal.pbio.2002846

Blank, H. M., Callahan, M., Pistikopoulos, I. P. E., Polymenis, A. O., & Polymenis, M. (2018). Scaling of G1 duration with population doubling time by a cyclin in Saccharomyces cerevisiae. *Genetics*, *210*(3), 895-906. https://doi.org/10.1534/genetics.118.301507

Bochner, B. R. (2009). Global phenotypic characterization of bacteria. *FEMS Microbiol Rev*, *33*(1), 191-205. https://doi.org/10.1111/j.1574-6976.2008.00149.x

Bochner, B. R., Gadzinski, P., & Panomitros, E. (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res*, *11*(7), 1246-1255. https://doi.org/10.1101/gr.186501

Bougen-Zhukov, N., Loh, S. Y., Lee, H. K., & Loo, L. H. (2017). Large-scale image-based screening and profiling of cellular phenotypes. *Cytometry A*, *91*(2), 115-125. https://doi.org/10.1002/cyto.a.22909

Braberg, H., Echeverria, I., Bohn, S., Cimermancic, P., Shiver, A., Alexander, R., Xu, J., Shales, M., Dronamraju, R., Jiang, S., Dwivedi, G., Bogdanoff, D., Chaung, K. K., Huttenhain, R., Wang, S., Mavor, D., Pellarin, R., Schneidman, D., Bader, J. S., Fraser, J. S., Morris, J., Haber, J. E., Strahl, B. D., Gross, C. A., Dai, J., Boeke, J. D., Sali, A., & Krogan, N. J. (2020). Genetic interaction mapping informs integrative structure determination of protein complexes. *Science*, *370*(6522). https://doi.org/10.1126/science.aaz4910

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., Richardson, J. E., & Mouse Genome Database, G. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Res*, *47*(D1), D801-D806. https://doi.org/10.1093/nar/gky1056

Camon, E. B., Barrell, D. G., Dimmer, E. C., Lee, V., Magrane, M., Maslen, J., Binns, D., & Apweiler, R. (2005). An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, *6 Suppl 1*, S17. https://doi.org/10.1186/1471-2105-6-S1-S17

Campos, M., Govers, S. K., Irnov, I., Dobihal, G. S., Cornet, F., & Jacobs-Wagner, C. (2018). Genomewide phenotypic analysis of growth, cell morphogenesis, and cell cycle events in Escherichia coli. *Mol Syst Biol*, *14*(6), e7573. https://doi.org/10.15252/msb.20177573

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning Interpretability: a survey on methods and metrics. *Electronics*, *8*(8). https://doi.org/ARTN 832
10.3390/electronics8080832

Ceapa, C., Lambert, J., van Limpt, K., Wels, M., Smokvina, T., Knol, J., & Kleerebezem, M. (2015). Correlation of Lactobacillus rhamnosus genotypes and carbohydrate Utilization signatures determined by phenotype profiling. *Appl Environ Microbiol*, *81*(16), 5458-5470. https://doi.org/10.1128/AEM.00851-15

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., & Wong, E. D. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res*, *40*(Database issue), D700-705. https://doi.org/10.1093/nar/gkr1029

Chibucos, M. C., Zweifel, A. E., Herrera, J. C., Meza, W., Eslamfam, S., Uetz, P., Siegele, D. A., Hu, J. C., & Giglio, M. G. (2014). An ontology for microbial phenotypes. *BMC Microbiol*, *14*, 294. https://doi.org/10.1186/s12866-014-0294-3

Child, D. (1990). *The essentials of factor analysis, 2nd ed*. Cassell Educational.

Collins, S. R., Schuldiner, M., Krogan, N. J., & Weissman, J. S. (2006). A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol*, *7*(7), R63. https://doi.org/10.1186/gb-2006-7-7-r63

Cooper, L., Meier, A., Laporte, M. A., Elser, J. L., Mungall, C., Sinn, B. T., Cavaliere, D., Carbon, S., Dunn, N. A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J. H., Stevenson, D. W., Arnaud, E., & Jaiswal, P. (2018). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res*, *46*(D1), D1168-D1180. https://doi.org/10.1093/nar/gkx1152

Daley, D. O., Rapp, M., Granseth, E., Melen, K., Drew, D., & von Heijne, G. (2005). Global topology analysis of the Escherichia coli inner membrane proteome. *Science*, *308*(5726), 1321-1323. https://doi.org/10.1126/science.1109730

Dedon, P. C., & Begley, T. J. (2014). A system of RNA modifications and biased codon use controls cellular stress response at the level of translation. *Chem Res Toxicol*, *27*(3), 330-337. https://doi.org/10.1021/tx400438d

Deng, Y., Gao, L., Wang, B., & Guo, X. (2015). HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLoS One*, *10*(2), e0115692. https://doi.org/10.1371/journal.pone.0115692

Dragset, M. S., Ioerger, T. R., Zhang, Y. J., Maerk, M., Ginbot, Z., Sacchettini, J. C., Flo, T. H., Rubin, E. J., & Steigedal, M. (2019). Genome-wide phenotypic profiling identifies and categorizes genes required for mycobacterial low iron fitness. *Sci Rep*, *9*(1), 11394. https://doi.org/10.1038/s41598-019-47905-y

Elseviers, D., Petrullo, L. A., & Gallagher, P. J. (1984). Novel E. coli mutants deficient in biosynthesis of 5-methylaminomethyl-2-thiouridine. *Nucleic Acids Res*, *12*(8), 3521-3534. https://doi.org/10.1093/nar/12.8.3521

Engel, S. R., Balakrishnan, R., Binkley, G., Christie, K. R., Costanzo, M. C., Dwight, S. S., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Hong, E. L., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R., Oughtred, R., Park, J., Skrzypek, M. S., Weng, S., Wong, E. D., Dolinski, K., Botstein, D., & Cherry, J. M. (2010). Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res*, *38*(Database issue), D433-436. https://doi.org/10.1093/nar/gkp917

Fabris, F., Doherty, A., Palmer, D., de Magalhaes, J. P., & Freitas, A. A. (2018). A new approach for interpreting Random Forest models and its application to the biology of ageing. *Bioinformatics*, *34*(14), 2449-2456. https://doi.org/10.1093/bioinformatics/bty087

Farooq, A. S., Greetham, D., Somani, A., Marvin, M. E., Louis, E. J., & Du, C. (2018). Identification of ethanologenic yeast strains from wild habitats. *Journal of Applied Microbiology and Biochemistry*, *2*. https://doi.org/10.21767/2576-1412.100025

Feldbauer, R., Schulz, F., Horn, M., & Rattei, T. (2015). Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics*, *16 Suppl 14*, S1. https://doi.org/10.1186/1471-2105-16-S14-S1

Fey, P., Dodson, R. J., Basu, S., Hartline, E. C., & Chisholm, R. L. (2019). DictyBase and the Dicty Stock Center (version 2.0) - a progress report. *Int J Dev Biol*, *63*(8-9-10), 563-572. https://doi.org/10.1387/ijdb.190226pf

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., & Morris, Q. (2018). GeneMANIA update 2018. *Nucleic Acids Res*, *46*(W1), W60-W64. https://doi.org/10.1093/nar/gky311

Fuhrer, T., Zampieri, M., Sevin, D. C., Sauer, U., & Zamboni, N. (2017). Genomewide landscape of gene-metabolome associations in Escherichia coli. *Mol Syst Biol*, *13*(1), 907. https://doi.org/10.15252/msb.20167150

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muniz-Rascado, L., Garcia-Sotelo, J. S., Alquicira-Hernandez, K., Martinez-Flores, I., Pannier, L., Castro-Mondragon, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martinez, C., Perez-Rueda, E., Alquicira-Hernandez, S., Porron-Sotelo, L., Lopez-Fuentes, A., Hernandez-Koutoucheva, A., Del Moral-Chavez, V., Rinaldi, F., & Collado-Vides, J. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res*, *44*(D1), D133-143. https://doi.org/10.1093/nar/gkv1156

Gene Ontology, C. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*, *49*(D1), D325-D334. https://doi.org/10.1093/nar/gkaa1113

Georgopoulou, D., Callari, M., Rueda, O. M., Shea, A., Martin, A., Giovannetti, A., Qosaj, F., Dariush, A., Chin, S. F., Carnevalli, L. S., Provenzano, E., Greenwood, W., Lerda, G., Esmaeilishirazifard, E., O'Reilly, M., Serra, V., Bressan, D., Consortium, I., Mills, G. B., Ali, H. R., Cosulich, S. S., Hannon, G. J., Bruna, A., & Caldas, C. (2021). Landscapes of cellular phenotypic diversity in breast cancer xenografts and their impact on drug response. *Nat Commun*, *12*(1), 1998. https://doi.org/10.1038/s41467-021-22303-z

Ghatak, S., King, Z. A., Sastry, A., & Palsson, B. O. (2019). The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Res*, *47*(5), 2446-2454. https://doi.org/10.1093/nar/gkz030

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K. D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C. Y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., & Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, *418*(6896), 387-391. https://doi.org/10.1038/nature00935

Giaever, G., Flaherty, P., Kumm, J., Proctor, M., Nislow, C., Jaramillo, D. F., Chu, A. M., Jordan, M. I., Arkin, A. P., & Davis, R. W. (2004). Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc Natl Acad Sci U S A*, *101*(3), 793-798. https://doi.org/10.1073/pnas.0307490100

Giaever, G., & Nislow, C. (2014). The yeast deletion collection: a decade of functional genomics. *Genetics*, *197*(2), 451-465. https://doi.org/10.1534/genetics.114.161620

Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., Mitraka, E., Schriml, L. M., Gaudet, P., Hobbs, E. T., Erill, I., Siegele, D. A., Hu, J. C., Mungall, C., & Chibucos, M. C. (2019). ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Res*, *47*(D1), D1186-D1194. https://doi.org/10.1093/nar/gky1036

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabasi, A. L. (2007). The human disease network. *Proc Natl Acad Sci U S A*, *104*(21), 8685-8690. https://doi.org/10.1073/pnas.0701361104

Greene, D., Richardson, S., & Turro, E. (2017). ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*, *33*(7), 1104-1106. https://doi.org/10.1093/bioinformatics/btw763

Grys, B. T., Lo, D. S., Sahin, N., Kraus, O. Z., Morris, Q., Boone, C., & Andrews, B. J. (2017). Machine learning and computer vision approaches for phenotypic profiling. *J Cell Biol*, *216*(1), 65-71. https://doi.org/10.1083/jcb.201610026

Guranowski, A., Jakubowski, H., & Holler, E. (1983). Catabolism of diadenosine 5',5'''-P1,P4-tetraphosphate in procaryotes. Purification and properties of diadenosine 5',5'''-P1,P4-tetraphosphate (symmetrical) pyrophosphohydrolase from Escherichia coli K12. *J Biol Chem*, *258*(24), 14784-14789. https://www.ncbi.nlm.nih.gov/pubmed/6317672

Guzzetta, G., Jurman, G., & Furlanello, C. (2010). A machine learning pipeline for quantitative phenotype prediction from genotype data. *BMC Bioinformatics*, *11 Suppl 8*, S3. https://doi.org/10.1186/1471-2105-11-S8-S3

Ha, C. W. Y., Martin, A., Sepich-Poore, G. D., Shi, B., Wang, Y., Gouin, K., Humphrey, G., Sanders, K., Ratnayake, Y., Chan, K. S. L., Hendrick, G., Caldera, J. R., Arias, C., Moskowitz, J. E., Ho Sui, S. J., Yang, S., Underhill, D., Brady, M. J., Knott, S., Kaihara, K., Steinbaugh, M. J., Li, H., McGovern, D. P. B., Knight, R., Fleshner, P., & Devkota, S. (2020). Translocation of viable gut microbiota to mesenteric adipose drives formation of creeping fat in humans. *Cell*, *183*(3), 666-683 e617. https://doi.org/10.1016/j.cell.2020.09.009

Hanauer, D. I., Graham, M. J., Sea, P., Betancur, L., Bobrownicki, A., Cresawn, S. G., Garlena, R. A., Jacobs-Sera, D., Kaufmann, N., Pope, W. H., Russell, D. A., Jacobs, W. R., Jr., Sivanathan, V., Asai, D. J., & Hatfull, G. F. (2017). An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. *Proc Natl Acad Sci U S A*, *114*(51), 13531-13536. https://doi.org/10.1073/pnas.1718188115

Harris, M. A., Lock, A., Bahler, J., Oliver, S. G., & Wood, V. (2013). FYPO: the fission yeast phenotype ontology. *Bioinformatics*, *29*(13), 1671-1678. https://doi.org/10.1093/bioinformatics/btt266

Harris, T. W., Arnaboldi, V., Cain, S., Chan, J., Chen, W. J., Cho, J., Davis, P., Gao, S., Grove, C. A., Kishore, R., Lee, R. Y. N., Muller, H. M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Rodgers, F. H., Russell, M., Schindelman, G., Auken, K. V., Wang, Q., Williams, G., Wright, A. J., Yook, K., Howe, K. L., Schedl, T., Stein, L., & Sternberg, P. W. (2020). WormBase: a modern model organism information resource. *Nucleic Acids Res*, *48*(D1), D762-D767. https://doi.org/10.1093/nar/gkz920

Hentchel, K. L., Reyes Ruiz, L. M., Curtis, P. D., Fiebig, A., Coleman, M. L., & Crosson, S. (2019). Genome-scale fitness profile of Caulobacter crescentus grown in natural freshwater. *ISME J*, *13*(2), 523-536. https://doi.org/10.1038/s41396-018-0295-6

Hill, D. P., Davis, A. P., Richardson, J. E., Corradi, J. P., Ringwald, M., Eppig, J. T., & Blake, J. A. (2001). Program description: Strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. *Genomics*, *74*(1), 121-128. https://doi.org/10.1006/geno.2001.6513

Hillenmeyer, M. E., Ericson, E., Davis, R. W., Nislow, C., Koller, D., & Giaever, G. (2010). Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome Biol*, *11*(3), R30. https://doi.org/10.1186/gb-2010-11-3-r30

Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., Proctor, M., St Onge, R. P., Tyers, M., Koller, D., Altman, R. B., Davis, R. W., Nislow, C., & Giaever, G. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, *320*(5874), 362-365. https://doi.org/10.1126/science.1150021

Hinkle, D. E., Wiersma , W., & Jurs, S. G. (2002). *Applied Statistics for the Behavioral Sciences (5th Edition)* [Book]. Houghton Mifflin.

Hinton, L. v. d. M. a. G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605. http://jmlr.org/papers/v9/vandermaaten08a.html

Hoehndorf, R., Hardy, N. W., Osumi-Sutherland, D., Tweedie, S., Schofield, P. N., & Gkoutos, G. V. (2013). Systematic analysis of experimental phenotype data reveals gene functions. *PLoS One*, *8*(4), e60847. https://doi.org/10.1371/journal.pone.0060847

Holliday, G. L., Davidson, R., Akiva, E., & Babbitt, P. C. (2017). Evaluating functional annotations of enzymes using the Gene Ontology. *Methods Mol Biol*, *1446*, 111-132. https://doi.org/10.1007/978-1-4939-3743-1_9

Hoose, S. A., Rawlings, J. A., Kelly, M. M., Leitch, M. C., Ababneh, Q. O., Robles, J. P., Taylor, D., Hoover, E. M., Hailu, B., McEnery, K. A., Downing, S. S., Kaushal, D., Chen, Y., Rife, A., Brahmbhatt, K. A., Smith, R., 3rd, & Polymenis, M. (2012). A systematic analysis of cell cycle regulators in yeast reveals that most factors act independently of cell size to control initiation of division. *PLoS Genet*, *8*(3), e1002590. https://doi.org/10.1371/journal.pgen.1002590

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: the next challenge. *Nat Rev Genet*, *11*(12), 855-866. https://doi.org/10.1038/nrg2897

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., & Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature*, *455*(7209), 47-50. https://doi.org/10.1038/455047a

Howe, D. G., Ramachandran, S., Bradford, Y. M., Fashena, D., Toro, S., Eagle, A., Frazer, K., Kalita, P., Mani, P., Martin, R., Moxon, S. T., Paddock, H., Pich, C., Ruzicka, L., Schaper, K., Shao, X., Singer, A., Van Slyke, C. E., & Westerfield, M. (2021). The Zebrafish Information Network: major gene page and home page updates. *Nucleic Acids Res*, *49*(D1), D1058-D1064. https://doi.org/10.1093/nar/gkaa1010

International Society for Biocuration. (2018). Biocuration: Distilling data into knowledge. *PLoS Biol*, *16*(4), e2002846. https://doi.org/10.1371/journal.pbio.2002846

J.-P., B. (1992). *Correspondence analysis handbook*. CRC Press.

Jiang, D., Zhao, Y., Wang, X., Fan, J., Heng, J., Liu, X., Feng, W., Kang, X., Huang, B., Liu, J., & Zhang, X. C. (2013). Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. *Proc Natl Acad Sci U S A*, *110*(36), 14664-14669. https://doi.org/10.1073/pnas.1308127110

Jorgensen, P., Nishikawa, J. L., Breitkreutz, B. J., & Tyers, M. (2002). Systematic identification of pathways that couple cell growth and division in yeast. *Science*, *297*(5580), 395-400. https://doi.org/10.1126/science.1070850

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*, *44*(D1), D457-462. https://doi.org/10.1093/nar/gkv1070

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881-892. https://doi.org/10.1109/TPAMI.2002.1017616

Karp, P. D. (2016). How much does curation cost? *Database (Oxford)*, *2016*. https://doi.org/10.1093/database/baw110

Karp, P. D., Ong, W. K., Paley, S., Billington, R., Caspi, R., Fulcher, C., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P. E., Subhraveti, P., Gama-Castro, S., Muniz-Rascado, L., Bonavides-Martinez, C., Santos-Zavaleta, A., Mackie, A., Collado-

Vides, J., Keseler, I. M., & Paulsen, I. (2018). The EcoCyc Database. *EcoSal Plus*, *8*(1). https://doi.org/10.1128/ecosalplus.ESP-0006-2018

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Jr., Assad-Garcia, N., Glass, J. I., & Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, *150*(2), 389-401. https://doi.org/10.1016/j.cell.2012.05.044

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martinez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velazquez-Ramirez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I., & Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res*, *45*(D1), D543-D550. https://doi.org/10.1093/nar/gkw1003

Keseler, I. M., Skrzypek, M., Weerasinghe, D., Chen, A. Y., Fulcher, C., Li, G. W., Lemmer, K. C., Mladinich, K. M., Chow, E. D., Sherlock, G., & Karp, P. D. (2014). Curation accuracy of model organism databases. *Database (Oxford), 2014*. https://doi.org/10.1093/database/bau058

Khatri, P., Done, B., Rao, A., Done, A., & Draghici, S. (2005). A semantic analysis of the annotations of the human genome. *Bioinformatics*, *21*(16), 3416-3421. https://doi.org/10.1093/bioinformatics/bti538

Ko, M., & Park, C. (2000). H-NS-Dependent regulation of flagellar synthesis is mediated by a LysR family protein. *J Bacteriol*, *182*(16), 4670-4672. https://doi.org/10.1128/jb.182.16.4670-4672.2000

Kohler, S., Oien, N. C., Buske, O. J., Groza, T., Jacobsen, J. O. B., McNamara, C., Vasilevsky, N., Carmody, L. C., Gourdine, J. P., Gargano, M., McMurry, J. A., Danis, D., Mungall, C. J., Smedley, D., Haendel, M., & Robinson, P. N. (2019). Encoding clinical data with the Human Phenotype Ontology for computational differential diagnostics. *Curr Protoc Hum Genet*, *103*(1), e92. https://doi.org/10.1002/cphg.92

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480. https://doi.org/10.1109/5.58325

Koo, B. M., Kritikos, G., Farelli, J. D., Todor, H., Tong, K., Kimsey, H., Wapinski, I., Galardini, M., Cabal, A., Peters, J. M., Hachmann, A. B., Rudner, D. Z., Allen, K. N., Typas, A., & Gross, C. A. (2017). Construction and analysis of two genome-scale deletion libraries for Bacillus subtilis. *Cell Syst*, *4*(3), 291-305 e297. https://doi.org/10.1016/j.cels.2016.12.013

Kritikos, G., Banzhaf, M., Herrera-Dominguez, L., Koumoutsi, A., Wartel, M., Zietek, M., & Typas, A. (2017). A tool named Iris for versatile high-throughput phenotyping in microorganisms. *Nat Microbiol*, *2*, 17014. https://doi.org/10.1038/nmicrobiol.2017.14

Krupke, D. M., Begley, D. A., Sundberg, J. P., Richardson, J. E., Neuhauser, S. B., & Bult, C. J. (2017). The Mouse Tumor Biology Database: A comprehensive resource for mouse models of human cancer. *Cancer Res*, *77*(21), e67-e70. https://doi.org/10.1158/0008-5472.CAN-17-0584

Kurihara, S., Oda, S., Kato, K., Kim, H. G., Koyanagi, T., Kumagai, H., & Suzuki, H. (2005). A novel putrescine utilization pathway involves gamma-glutamylated intermediates of Escherichia coli K-12. *J Biol Chem*, *280*(6), 4602-4608. https://doi.org/10.1074/jbc.M411114200

Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., Dos Santos, G., Garapati, P. V., Goodman, J. L., Gramates, L. S., Millburn, G., Strelets, V. B., Tabone, C. J., Thurmond, J., & FlyBase, C. (2021). FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res*, *49*(D1), D899-D907. https://doi.org/10.1093/nar/gkaa1026

Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S., & Qi, L. S. (2013). CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc*, *8*(11), 2180-2196. https://doi.org/10.1038/nprot.2013.132

Lee, A. Y., St Onge, R. P., Proctor, M. J., Wallace, I. M., Nile, A. H., Spagnuolo, P. A., Jitkova, Y., Gronda, M., Wu, Y., Kim, M. K., Cheung-Ong, K., Torres, N. P., Spear, E. D., Han, M. K., Schlecht, U., Suresh, S., Duby, G., Heisler, L. E., Surendra, A., Fung, E., Urbanus, M. L., Gebbia, M., Lissina, E., Miranda, M., Chiang, J. H., Aparicio, A. M., Zeghouf, M., Davis, R. W., Cherfils, J., Boutry, M., Kaiser, C. A., Cummins, C. L., Trimble, W. S., Brown, G. W., Schimmer, A. D., Bankaitis, V. A., Nislow, C., Bader, G. D., & Giaever, G. (2014). Mapping the cellular response to small molecules using chemogenomic fitness signatures. *Science*, *344*(6180), 208-211. https://doi.org/10.1126/science.1250217

Lees, J. A., Mai, T. T., Galardini, M., Wheeler, N. E., Horsfield, S. T., Parkhill, J., & Corander, J. (2020). Improved prediction of bacterial benotype-phenotype associations using interpretable pangenome-spanning regressions. *MBio*, *11*(4). https://doi.org/10.1128/mBio.01344-20

Lian, J., Schultz, C., Cao, M., HamediRad, M., & Zhao, H. (2019). Multi-functional genome-wide CRISPR system for high throughput genotype–phenotype mapping. *Nature Communications*, *10*(1), 5794. https://doi.org/10.1038/s41467-019-13621-4

Liu, X., & Winey, M. (2012). The MPS1 family of protein kinases. *Annu Rev Biochem*, *81*, 561-585. https://doi.org/10.1146/annurev-biochem-061611-090435

Lock, A., Rutherford, K., Harris, M. A., Hayles, J., Oliver, S. G., Bahler, J., & Wood, V. (2019). PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Res*, *47*(D1), D821-D827. https://doi.org/10.1093/nar/gky961

Luciano, D. J., Levenson-Palmer, R., & Belasco, J. G. (2019). Stresses that raise Np4A levels induce protective nucleoside tetraphosphate capping of bacterial RNA. *Mol Cell*, *75*(5), 957-966 e958. https://doi.org/10.1016/j.molcel.2019.05.031

Mackie, A. M., Hassan, K. A., Paulsen, I. T., & Tetu, S. G. (2014). Biolog phenotype microArrays for phenotypic characterization of microbial cells. In I. T. Paulsen & A. J. Holmes (Eds.), *Environmental Microbiology: Methods and Protocols* (pp. 123-130). Humana Press. https://doi.org/10.1007/978-1-62703-712-9_10

Mahfouz, N., Ferreira, I., Beisken, S., von Haeseler, A., & Posch, A. E. (2020). Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *J Antimicrob Chemother*, *75*(11), 3099-3108. https://doi.org/10.1093/jac/dkaa257

Marquardt, D. W., & Snee, R. D. (1975). Ridge Regression in Practice. *The American Statistician*, *29*(1), 3-20. https://doi.org/10.1080/00031305.1975.10479105

Montojo, J., Zuberi, K., Rodriguez, H., Bader, G. D., & Morris, Q. (2014). GeneMANIA: Fast gene network construction and function prediction for Cytoscape. *F1000Res*, *3*, 153. https://doi.org/10.12688/f1000research.4572.1

Montojo, J., Zuberi, K., Rodriguez, H., Kazi, F., Wright, G., Donaldson, S. L., Morris, Q., & Bader, G. D. (2010). GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, *26*(22), 2927-2928. https://doi.org/10.1093/bioinformatics/btq562

Mutalik, V. K., Adler, B. A., Rishi, H. S., Piya, D., Zhong, C., Koskella, B., Kutter, E. M., Calendar, R., Novichkov, P. S., Price, M. N., Deutschbauer, A. M., & Arkin, A. P. (2020). High-throughput mapping of the phage resistance landscape in E. coli. *PLoS Biol*, *18*(10), e3000877. https://doi.org/10.1371/journal.pbio.3000877

Mutalik, V. K., Novichkov, P. S., Price, M. N., Owens, T. K., Callaghan, M., Carim, S., Deutschbauer, A. M., & Arkin, A. P. (2019). Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria. *Nat Commun*, *10*(1), 308. https://doi.org/10.1038/s41467-018-08177-8

Nakamura, H., Saiki, K., Mogi, T., & Anraku, Y. (1997). Assignment and functional roles of the cyoABCDE gene products required for the Escherichia coli bo-type quinol oxidase. *J Biochem*, *122*(2), 415-421. https://doi.org/10.1093/oxfordjournals.jbchem.a021769

Nichols, R. J., Sen, S., Choo, Y. J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K. M., Lee, K. J., Wong, A., Shales, M., Lovett, S., Winkler, M. E., Krogan, N. J., Typas, A., & Gross, C. A. (2011). Phenotypic landscape of a bacterial cell. *Cell*, *144*(1), 143-156. https://doi.org/10.1016/j.cell.2010.11.052

Noble, W. S. (2006). What is a support vector machine? *Nat Biotechnol*, *24*(12), 1565-1567. https://doi.org/10.1038/nbt1206-1565

Noinaj, N., Guillier, M., Barnard, T. J., & Buchanan, S. K. (2010). TonB-dependent transporters: regulation, structure, and function. *Annu Rev Microbiol*, *64*, 43-60. https://doi.org/10.1146/annurev.micro.112408.134247

Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., Jaiswal, P., Kalberer, S. R., Lloyd, J. P., Meinke, D., Menda, N., Moore, L., Nelson, R. T., Pujar, A., Lawrence, C. J., & Huala, E. (2015). An ontology approach to comparative phenomics in plants. *Plant Methods*, *11*, 10. https://doi.org/10.1186/s13007-015-0053-y

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572. https://doi.org/10.1080/14786440109462720

Pesquita, C. (2017). Semantic similarity in the Gene Ontology. The Gene Ontology Handbook, Chapter 12 [Online book]. *MIMB*, *1446*. https://doi.org/https://doi.org/10.1007/978-1-4939-3743-1_12

Peters, J. M., Colavin, A., Shi, H., Czarny, T. L., Larson, M. H., Wong, S., Hawkins, J. S., Lu, C. H., Koo, B. M., Marta, E., Shiver, A. L., Whitehead, E. H., Weissman, J. S., Brown, E. D., Qi, L. S., Huang, K. C., & Gross, C. A. (2016). A comprehensive, CRISPR-based functional analysis of essential benes in bacteria. *Cell*, *165*(6), 1493-1506. https://doi.org/10.1016/j.cell.2016.05.003

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., Blow, M. J., Arkin, A. P., & Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, *557*(7706), 503-509. https://doi.org/10.1038/s41586-018-0124-0

Priness, I., Maimon, O., & Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, *8*, 111. https://doi.org/10.1186/1471-2105-8-111

Raetz, C. R., & Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annu Rev Biochem*, *71*, 635-700. https://doi.org/10.1146/annurev.biochem.71.110601.135414

Reimer, L. C., Vetcininova, A., Carbasse, J. S., Sohngen, C., Gleim, D., Ebeling, C., & Overmann, J. (2019). BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. *Nucleic Acids Res*, *47*(D1), D631-D636. https://doi.org/10.1093/nar/gky879

Reynolds, D. A. a. R., R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*, 72-83. https://doi.org/10.1109/89.365379

Rishi, H. S., Toro, E., Liu, H., Wang, X., Qi, L. S., & Arkin, A. P. (2020). Systematic genome-wide querying of coding and non-coding functional elements in E. coli using CRISPRi. https://doi.org/10.1101/2020.03.04.975888

Rodriguez-Esteban, R. (2015). Biocuration with insufficient resources and fixed timelines. *Database (Oxford)*, *2015*. https://doi.org/10.1093/database/bav116

Rokach, L. (2019). *Ensemble learning: Pattern classification using ensemble methods (second edition)* [Book].

Rousseeuw, L. K. P. J. (1990). Divisive analysis (program DIANA). Chapter 6, Finding groups in data: An introduction to cluster analysis. https://doi.org/doi.org/10.1002/9780470316801.ch6

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salimi, N., & Vita, R. (2006). The biocurator: connecting and enhancing scientific data. *PLoS Comput Biol*, *2*(10), e125. https://doi.org/10.1371/journal.pcbi.0020125

Schapire Robert, E., & Freund, Y. (2013). Boosting: foundations and algorithms. *Kybernetes*, *42*(1), 164-166. https://doi.org/10.1108/03684921311295547

Schneider, B. L., Yang, Q. H., & Futcher, A. B. (1996). Linkage of replication to start by the Cdk inhibitor Sic1. *Science*, *272*(5261), 560-562. https://doi.org/10.1126/science.272.5261.560

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesth Analg*, *126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Schober, P. B., Christa; Schwarte, Lothar A. (2018). Correlation coefficients: appropriate use and interpretation. *126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Shefchek, K. A., Harris, N. L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., Keith, D., Conlin, T., Vasilevsky, N., Zhang, X. A., Balhoff, J. P., Babb, L., Bello, S. M., Blau, H., Bradford, Y., Carbon, S., Carmody, L., Chan, L. E., Cipriani, V., Cuzick, A., Della Rocca, M., Dunn, N., Essaid, S., Fey, P., Grove, C., Gourdine, J. P., Hamosh, A., Harris, M., Helbig, I., Hoatlin, M., Joachimiak, M., Jupp, S., Lett, K. B., Lewis, S. E., McNamara, C., Pendlington, Z. M., Pilgrim, C., Putman, T., Ravanmehr, V., Reese, J., Riggs, E., Robb, S., Roncaglia, P., Seager, J., Segerdell, E., Similuk, M., Storm, A. L., Thaxon, C., Thessen, A., Jacobsen, J. O. B., McMurry, J. A., Groza, T., Kohler, S., Smedley, D., Robinson, P. N., Mungall, C. J., Haendel, M. A., Munoz-Torres, M. C., & Osumi-Sutherland, D. (2020). The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*, *48*(D1), D704-D715. https://doi.org/10.1093/nar/gkz997

Shiver, A. L., Osadnik, H., Peters, J. M., Mooney, R. A., Wu, P. I., Hu, J. C., Landick, R., Huang, K. C., & Gross, C. A. (2020). Chemical-genetic interrogation of RNA polymerase mutants reveals structure-function relationships and physiological tradeoffs. *bioRxiv*. https://doi.org/10.1101/2020.06.16.155770

Siegele, D. A., LaBonte, S. A., Wu, P. I., Chibucos, M. C., Nandendla, S., Giglio, M. G., & Hu, J. C. (2019). Phenotype annotation with the ontology of microbial phenotypes (OMP). *J Biomed Semantics*, *10*(1), 13. https://doi.org/10.1186/s13326-019-0205-5

Škunca, N., Altenhoff, A., & Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. *PLoS Comput Biol*, *8*(5), e1002533. https://doi.org/10.1371/journal.pcbi.1002533

Škunca, N., Roberts, R. J., & Steffen, M. (2017). Evaluating computational Gene Ontology annotations. In *The Gene Ontology Handbook* (Vol. 1446). https://doi.org/https://doi.org/10.1007/978-1-4939-3743-1_12

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Consortium, O. B. I., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, *25*(11), 1251-1255. https://doi.org/10.1038/nbt1346

Smith, C. M., Hayamizu, T. F., Finger, J. H., Bello, S. M., McCright, I. J., Xu, J., Baldarelli, R. M., Beal, J. S., Campbell, J., Corbani, L. E., Frost, P. J., Lewis, J. R., Giannatto, S. C., Miers, D., Shaw, D. R., Kadin, J. A., Richardson, J. E., Smith, C. L., & Ringwald, M. (2019). The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res*, *47*(D1), D774-D779. https://doi.org/10.1093/nar/gky922

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*, *27*(2), 130-135. https://doi.org/10.11919/j.issn.1002-0829.215044

Stoesser, N., Batty, E. M., Eyre, D. W., Morgan, M., Wyllie, D. H., Del Ojo Elias, C., Johnson, J. R., Walker, A. S., Peto, T. E., & Crook, D. W. (2013). Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. *J Antimicrob Chemother*, *68*(10), 2234-2244. https://doi.org/10.1093/jac/dkt180

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., & von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, *43*(Database issue), D447-452. https://doi.org/10.1093/nar/gku1003

Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics*, *2*(3 Suppl), S75-83. https://www.ncbi.nlm.nih.gov/pubmed/15130820

Tang, Z. Y., Xu, Y. Y., Jin, L., Aibaidula, A., Lu, J. F., Jiao, Z. C., Wu, J. S., Zhang, H., & Shen, D. G. (2020). Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients. *Ieee Transactions on Medical Imaging*, *39*(6), 2100-2109. https://doi.org/10.1109/Tmi.2020.2964310

Tarasava, K., Oh, E. J., Eckert, C. A., & Gill, R. T. (2018). CRISPR-enabled tools for engineering microbial genomes and phenotypes. *Biotechnol J*, *13*(9), e1700586. https://doi.org/10.1002/biot.201700586

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J., & Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*, *47*(D1), D941-D947. https://doi.org/10.1093/nar/gky1015

Teschendorff, A. E. (2019). Avoiding common pitfalls in machine learning omic data science. *Nat Mater*, *18*(5), 422-427. https://doi.org/10.1038/s41563-018-0241-z

The Gene Ontology, C. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*, *45*(D1), D331-D338. https://doi.org/10.1093/nar/gkw1108

Thessen, A. E., Grondin, C. J., Kulkarni, R. D., Brander, S., Truong, L., Vasilevsky, N. A., Callahan, T. J., Chan, L. E., Westra, B., Willis, M., Rothenberg, S. E., Jarabek, A. M., Burgoon, L., Korrick, S. A., & Haendel, M. A. (2020). Community approaches for integrating environmental exposures into human models of disease. *Environ Health Perspect*, *128*(12), 125002. https://doi.org/10.1289/EHP7215

Thessen, A. E., Walls, R. L., Vogt, L., Singer, J., Warren, R., Buttigieg, P. L., Balhoff, J. P., Mungall, C. J., McGuinness, D. L., Stucky, B. J., Yoder, M. J., & Haendel, M. A. (2020). Transforming the study of organisms: Phenomic data models and knowledge bases. *PLoS Comput Biol*, *16*(11), e1008376. https://doi.org/10.1371/journal.pcbi.1008376

Thompson, M. G., Blake-Hedges, J. M., Cruz-Morales, P., Barajas, J. F., Curran, S. C., Eiben, C. B., Harris, N. C., Benites, V. T., Gin, J. W., Sharpless, W. A., Twigg, F. F., Skyrud, W., Krishna, R. N., Pereira, J. H., Baidoo, E. E. K., Petzold, C. J., Adams, P. D., Arkin, A. P., Deutschbauer, A. M., & Keasling, J. D. (2019). Massively parallel fitness profiling reveals multiple novel enzymes in Pseudomonas putida lysine metabolism. *MBio*, *10*(3). https://doi.org/10.1128/mBio.02577-18

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*, *58*(1), 267-288. https://doi.org/DOI 10.1111/j.2517-6161.1996.tb02080.x

Tohsato, Y., & Mori, H. (2008). Phenotype profiling of single gene deletion mutants of E. coli using Biolog technology. *Genome Inform*, *21*, 42-52. https://www.ncbi.nlm.nih.gov/pubmed/19425146

Tong, M., French, S., El Zahed, S. S., Ong, W. K., Karp, P. D., & Brown, E. D. (2020). Gene dispensability in Escherichia coli grown in thirty different carbon environments. *MBio*, *11*(5). https://doi.org/10.1128/mBio.02259-20

Typas, A., Nichols, R. J., Siegele, D. A., Shales, M., Collins, S. R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B. L., Mori, H., Weissman, J. S., Krogan, N. J., & Gross, C. A. (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. *Nat Methods*, *5*(9), 781-787. https://doi.org/10.1038/nmeth.1240

UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, *47*(D1), D506-D515. https://doi.org/10.1093/nar/gky1049

Vaas, L. A., Sikorski, J., Hofner, B., Fiebig, A., Buddruhs, N., Klenk, H. P., & Goker, M. (2013). opm: an R package for analysing OmniLog(R) phenotype microarray data. *Bioinformatics*, *29*(14), 1823-1824. https://doi.org/10.1093/bioinformatics/btt291

Valens, M., Thiel, A., & Boccard, F. (2016). The MaoP/maoS Site-Specific System Organizes the Ori Region of the E. coli Chromosome into a Macrodomain. *PLoS Genet*, *12*(9), e1006309. https://doi.org/10.1371/journal.pgen.1006309

Van Camp, P. J., Haslam, D. B., & Porollo, A. (2020). Bioinformatics approaches to the understanding of molecular mechanisms in antimicrobial resistance. *Int J Mol Sci*, *21*(4). https://doi.org/10.3390/ijms21041363

Vehkala, M., Shubin, M., Connor, T. R., Thomson, N. R., & Corander, J. (2015). Novel R pipeline for analyzing Biolog Phenotypic MicroArray data. *PLoS One*, *10*(3), e0118392. https://doi.org/10.1371/journal.pone.0118392

Vivijs, B., Aertsen, A., & Michiels, C. W. (2016). Identification of genes required for growth of Escherichia coli MG1655 at moderately low pH. *Front Microbiol*, *7*, 1672. https://doi.org/10.3389/fmicb.2016.01672

Vuckovic, D., Gasparini, P., Soranzo, N., & Iotchkova, V. (2015). MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies. *Bioinformatics*, *31*(16), 2754-2756. https://doi.org/10.1093/bioinformatics/btv222

Wang, C. W. (2006). New ensemble machine learning method for classification and prediction on gene expression data. *Conf Proc IEEE Eng Med Biol Soc*, *2006*, 3478-3481. https://doi.org/10.1109/IEMBS.2006.259893

Wang, D. (2020). Using genetics to reveal protein structure. *Science*, *370*(6522), 1269-1270. https://doi.org/10.1126/science.abf3863

Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R., & Church, G. M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, *460*(7257), 894-898. https://doi.org/10.1038/nature08187

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, *23*(10), 1274-1281. https://doi.org/10.1093/bioinformatics/btm087

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, *38*(Web Server issue), W214-220. https://doi.org/10.1093/nar/gkq537

Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P., & Deutschbauer, A. (2015). Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio*, *6*(3), e00306-00315. https://doi.org/10.1128/mBio.00306-15

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D. J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J. L., Riles, L., Roberts, C. J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R. K., Veronneau, S., Voet, M., Volckaert, G., Ward, T. R., Wysocki, R., Yen, G. S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M., & Davis, R. W. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, *285*(5429), 901-906. https://doi.org/10.1126/science.285.5429.901

Wrighton, K. H. (2018). Synthetic biology: Multiplex genome engineering in eukaryotes. *Nat Rev Genet*, *19*(1), 6-7. https://doi.org/10.1038/nrg.2017.103

Wu, P. I. F., Ross, C., Siegele, D. A., & Hu, J. C. (2021). Insights from the reanalysis of high-throughput chemical genomics data for Escherichia coli K-12. *G3 Genes|Genomes|Genetics*, *11*(1). https://doi.org/10.1093/g3journal/jkaa035

Xu, C., & Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biol*, *20*(1), 76. https://doi.org/10.1186/s13059-019-1689-0

Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrubbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., Walker, G. C., & Collins, J. J. (2019). A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell*, *177*(6), 1649-1661 e1649. https://doi.org/10.1016/j.cell.2019.04.016

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugan, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J. G., Maurel, T., McDowall, M., McMahon, A., Mohanan, S., Moore, B., Nuhn, M., Oheh, D. N., Parker, A., Parton, A., Patricio, M., Sakthivel, M. P., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sycheva, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Flint, B., Frankish, A., Hunt, S. E., G, I. I., Kostadima, M., Langridge, N., Loveland, J. E., Martin, F. J., Morales, J., Mudge, J. M., Muffato, M., Perry, E., Ruffier, M., Trevanion, S. J., Cunningham, F., Howe, K. L., Zerbino, D. R., & Flicek, P. (2020). Ensembl 2020. *Nucleic Acids Res*, *48*(D1), D682-D688. https://doi.org/10.1093/nar/gkz966

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, *26*(7), 976-978. https://doi.org/10.1093/bioinformatics/btq064

Zahir, T., Camacho, R., Vitale, R., Ruckebusch, C., Hofkens, J., Fauvart, M., & Michiels, J. (2019). High-throughput time-resolved morphology screening in bacteria reveals phenotypic responses to antibiotics. *Communications Biology*, *2*(1), 269. https://doi.org/10.1038/s42003-019-0480-9

Zhu, B., & Stulke, J. (2018). SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis. *Nucleic Acids Res*, *46*(D1), D743-D748. https://doi.org/10.1093/nar/gkx908

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005). *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *67*, 768-768. https://doi.org/DOI 10.1111/j.1467-9868.2005.00527.x