Title: Insights from the reanalysis of high-throughput chemical genomics data for *Escherichia coli* K-12

Authors: Peter I-Fan Wu[1], Curtis Ross[1], Deborah A. Siegele[2] and James C. Hu[1,3]

Affiliations:

1. Department of Biochemistry and Biophysics, Texas A&M University and Texas Agrilife Research, College Station, TX 77843-2128

2. Department of Biology, Texas A&M University, College Station, TX 77843-3258

3. Deceased

Correspondence: siegele@bio.tamu.edu

Key words: phenotypic profiling, functional genomics, microbial genomics, biostatistics, *Escherichia coli*, bacterial genetics

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 qualitative 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.

INTRODUCTION

Genome-wide genetic screens and chemical genomic studies, pioneered in yeast (GIAEVER AND 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 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.

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. Fitness scores were obtained and normalized to a standard normal distribution based on the mean fitness for all strains in a given condition. Positive scores indicate increased fitness and negative scores indicate decreases fitness. Fitness scores were obtained and normalized to a standard normal distribution where positive scores indicate increased fitness and negative scores indicate decreased fitness, which was based on the mean fitness for all strains in a given growth condition.

Six annotation sets were used as sources of information about gene function. Annotations of *E. coli* genes to metabolic pathways and protein complexes were obtained from EcoCyc (Keseler et al., 2017); annotation of genes to operons and regulons were extracted from EcoCyc and RegulonDB (Gama-Castro et al., 2016); and annotations of genes to KEGG modules, 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 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 metabolic pathway or protein complex, etc. The number of genes annotated by each annotation set and the total number of annotations can be found in Materials and Methods.

The annotations of *E. coli* genes with Gene Ontology (GO) biological process terms (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). While it is possible to identify gene pairs that are co-annotated with the same GO term(s), automated methods will include co-annotations to high-level terms, such as ‘GO:0044237 cellular metabolic process’ or ‘GO:0051716 cellular response to stimulus’, which don’t provide very specific information about function. Also, co-annotation doesn’t capture instances where two genes are annotated with related, but not identical, terms. These limitation 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. The number of genes annotated with GO biological process terms set and the total number of annotations can be found in the Materials and Methods.

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 1). The first violin plot shows the distribution of |PCC| values for all possible gene pairs (mean |PCC| = 0.00016). The majority have a |PCC| value <0.25 and only 0.16% have a |PCC| value >0.75 (an arbitrarily chosen cut-off). 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 (0.032), 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| = 0.05). 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| = 0.19, 0.30, 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 increase also continued; this fraction increased from 0.16% for all gene pairs, to 3.2% for gene pairs in the same pathways, to 4.9% for gene pairs in the same protein complexes, to 19% for gene pairs in the same pathways and complexes, and to 30% for gene pairs that are co-annotated in pathways, 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 distribution of |PCC| values for all pairwise combinations of genes in each pathway or protein complex. Of the 366 pathways and 271 protein complexes analyzed, 72% of the pathways and 67% of the protein complexes had a median |PCC| value that was higher than the random expectation.

Phenotypic profile similarity is explained by functional annotations

To address the second question, which is to test 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). 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), 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 the auxotrophic phenotypes, 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 3, even though only a small fraction of conditions were excluded, this change resulted in dramatically higher precision overall, regardless of which functional annotation set was used to score co-annotation. A comparable increase in precision was also seen when auxotrophic mutants were excluded from the data set (Supplemental Figure S3).

Alternative metrics for measuring phenotypic profile similarity

There are other methods, besides the Pearson Correlation Coefficient, that can be used to assess 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 phenotypic profile similarity, and used the union of the five annotation sets 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 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 indicates 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. If we recalculate precision for all three metrics after excluding the 10 growth conditions where auxotrophic mutants don’t grow, there is very little difference in precision for the three metrics (Figure 4b).

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 and by the need for methods to combine quantitative and qualitative phenotypic information. Different quantitative datasets could be combined by renormalizing the data to make them interoperable. Alternatively, quantitative phenotypes could be converted to qualitative phenotypes, which would allow integration of both quantitative and qualitative data. We chose to test the second approach because, if successful, it would allow more datasets to be combined.

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| values used to separate the three phenotype classes were based on the 5% false discovery rate as described (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.

The first violin plot in Figure 5a shows the distribution of MI values for all possible gene pairs, followed, from left to right, by the distribution of MI values for gene pairs co-annotated to either the same pathway; the same protein complex; the same pathway and protein complex; or the same pathway, protein complex, operon, regulon, and KEGG module. Converting the continuous quantitative fitness values to discrete ternary scores reduced the variation in the data, reflected by the change in shape of the violin plots compared to the plots shown in Figure 1. However, as was seen for the mean |PCC| values in the analysis of the quantitative data (Figure 1), the mean MI values increased as the functional associations for a given gene pair increased (Figure 5a inset).

Many of the growth conditions used in the original chemical genomics study involved multiple tests of the same chemical present at different concentrations. To test the effect of further simplifying the phenotypes, the original 324 growth conditions were reduced to 114 unique stresses by including the score for only the most significant phenotype for each chemical treatment (1 or -1, as appropriate, or using a score of 0 if no significant phenotypes were seen for that treatment). The violin plots in Figure 5b show the distribution of MI values for all gene pairs and for different combinations of annotation sets for the reduced dataset. 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 5b inset). In addition, when the distributions of gene pairs in the same co-annotation group are compared between Figures 5a and 5b, very 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 quantitative phenotype scores were converted to the 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. Then precision was calculated, as described earlier, and was plotted versus ranking. As can be seen in Figure 6, precision is comparable for the top 100 gene pairs for both quantitative and 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 data sets, 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 there is still significant potential in using the discretized version of phenotypes to explain functions.

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. The Wang method (Wang et al., 2007) was used to compute semantic similarity, 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 7a, semantic similarity increased when only co-annotated gene pairs were considered. The mean pairwise semantic similarity increased from 0.217 for all genes with GO biological process annotations (first violin plot) to 0.543 for gene pairs co-annotated to the same EcoCyc pathway (second violin plot), and to 0.803 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=0.892) as well as for gene pairs that are co-annotated in all 5 annotation sets (mean=0.889), 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. The violin plots in Figure 7b show, from left to right, the distribution of semantic similarity values for all gene pairs with GO biological process annotations, the subset of gene pairs with |PCC| >0.75, the subset of gene pairs with MI >0.15 (where MI was determined using the ternary qualitative fitness scores for all growt conditions), and the subset of gene pairs with MI >0.32 (where . Comparison of the first two violin plots shows that gene pairs with |PCC| >0.75 are significantly enriched for higher semantic similarity. (The cutoff of |PCC| >0.75 was chosen arbitrarily to represent a moderate to high correlation (Hinkle et al., 2002).) Enrichment for higher semantic similarity scores was also seen for the next two subsets of gene pairs, where phenotypic profile similarity was calculated using the qualitative, ternary fitness values for either all 324 growth conditions (third violin plot) or for the collapsed set of 114 growth conditions (fourth violin plot). (The MI cutoffs of >0.15 for the third violin plot and >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.) These results are consistent with those in Figure 4b, which show higher phenotypic profile similarity enriches for co-annotated gene pairs.

In order to assess whether gene pairs that have higher semantic similarity also have higher phenotypic profile similarity, we chose an arbitrary cutoff of 0.5 for semantic similarity and used it to select a subset of gene pairs from the entire set of gene pairs with GO biological process annotations. We then compared the distribution of semantic similarity scores for the two sets of gene pairs. The violin plots are shown in Figure 8. Although the two distributions appeared almost identical, the subset of gene pairs with semantic similarity >0.5 is enriched for gene pairs with higher phenotypic profile similarity. The difference in the mean |PCC| values for the two distributions is small (0.093 vs 0.10), but it is statistically significant based on the Mann-Whitney test, p<0.0001. This is consistent with Figure 1, where co-annotated gene pairs show enriched phenotypic similarity.

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 associatons are enriched for high phenotypic profile similarity scores 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, 2014; 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, there was no significant difference in the semantic similarity of gene pairs whether electronic annotations were included (Figure 7b) or excluded (Figure S4).

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. We compared the performance of the metrics based on 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 the five annotation sets. 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). Because 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, but this wouldn’t be very informative.

Overall, there appeared to be little difference in the performance of |PCC|, |SRCC| or MI based on their precision scores for the top 500 gene pairs (Figure 4b). Initially, it appeared that |SRCC| and MI outperformed |PCC| (Figure 4a). However, when the analysis was repeated after removing the conditions involving growth on minimal media, the precision for gene pairs ranked based on |PCC| increased significantly (compare Figures 4a and 4b). We suggest that this difference is due to the sensitivity of the Pearson Correlation Coefficient to outliers in the data (Schober et al., 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. In contrast, these auxotrophic mutants didn't have a significant phenotype in most of the other growth conditions tested, which used rich media, so the large negative fitness scores on minimal media were essentially outliers. In our analysis, the sensitivity of PCC to outliers interfered with the measurement of precision because there were so many combinations of genes from different biosynthetic pathways that shared an auxotrophic phenotype but did not share a functional annotation in the annotation sets used.

However, this doesn't mean that |PCC| can't be used to measure phenotypic profile similarity in high-throughput phenotype screens. For most gene pairs that don'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.

In addition to showing comparable precision, the three metrics, |PCC|, |SRCC|, and MI, also produced comparable profile similarity scores for many, although not all, gene pairs. We conclude that there is no single best way to measure phenotypic profile similarity. Instead, it may be advantageous to use more than one correlation metric when searching for functional associations. For high-throughput experiments that measured growth of a large number of strains in many different environments, it may also be useful to preprocess the fitness data, such as filtering or combining results from certain growth conditions.

To make it easier to compare results for the different similarity metrics, we have made the data set 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|, |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 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, filtered out the gene pairs that have no co-annotations, and then manually examined the 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 1). 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 21.1. 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-coannotated 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, Vivijs et al., 2016, Luciano et al., 2019).

A significant conclusion from this study is that functional associations can still be inferred from phenotypic profiles after quantitative fitness scores are converted to qualitative, ternary fitness values. While some information was lost compared to using quantitative fitness scores, the precision based on qualitative fitness values was much greater than for randomly ordered gene pairs (Figure 6). This result suggests that inherently qualitative phenotypes, such as aspects of cell morphology, could be incorporated into phenotypic profiles and used to infer functional associations. It may also be possible to incorporate phenotype annotations into phenotypic profiles. These annotations typically capture information in a qualitative fashion and have previously been shown to be useful for inferring gene function (Hoehndorf et al., 2013; Ascensao et al., 2014). These results also suggest that using qualitative phenotypes may be a viable option for integrating phenotype information from different studies. Thus, we believe that using qualitative phenotypes to combine more *E. coli* datasets, or datasets from other microorganisms, will allow us to extract many more functional insights.

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.

Six annotation sets including GO annotations were obtained from various sources: From a downloaded version of EcoCyc version 21.1 (http://bioinformatics.ai.sri.com/ecocyc/dist/flatfiles-52983746/), the ECK identifiers in supplemental Table S2 from the original research paper (Nichols et al., 2011) were verified, corrected and mapped to EcoCyc gene identifiers and b numbers using information in the file genes.txt. EcoCyc Pathway annotations were mapped to each gene using information in the file pathways.col. EcoCyc Protein complex annotations were mapped to each gene using information in the file protcplxs.col. 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 file operon.txt was the source of operon annotations. The file object_synonym.txt 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 file gene_association.ecocyc 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 number of genes annotated by each annotation set and the total number of annotations are shown in Table 2.

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, 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). 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 plots associated with each violin plot, the middle lines in the boxes represent medians; the whiskers indicate the 1.5 interquartile range (IQR) away from either Q1 (lower box boundary) or Q3 (upper box boundary).

The code and data files used for calculations and reproducing the results are available on GitHub: https://github.com/peterwu19881230/Systematic-analyses-ecoli-phenotypes.

ACKNOWLEDGEMENTS: We mourn the unexpected death of JCH who led this project and passed away on 23 Jan 2020. We hope this publication will continue his scientific legacy. We thank Michelle Giglio, Matthew Sachs and Yen-Ting Lu 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.

REFERENCES

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

Ascensao, J.A., Dolan, M.E., Hill, D.P., and Blake, J.A. (2014) Methodology for the inference of gene function from phenotype data. *BMC Bioinformatics*, *15*(1), 405. doi:10.1186/s12859-014-0405-z

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

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. doi:10.15252/msb.20177573

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

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

Dedon, P.C. and 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*, 330−337. doi:10.1021/tx400438d

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), 3521-34. doi:10.1093/nar/12.8.3521

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muniz-Rascado, L., Garcia-Sotelo, J. S., . . . 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. doi:10.1093/nar/gkv1156

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

Guranowski, H., 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), 14784-9. PMID:6317672

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. doi: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. doi:10.1186/gb-2010-11-3-r30

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

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. doi: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. doi:https://doi.org/10.1007/978-1-4939-3743-1_9

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

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. doi:10.1093/nar/gkv1070

Karp, P. D., Ong, W. K., Paley, S., Billington, R., Caspi, R., Fulcher, C., . . . Paulsen, I. (2018). The EcoCyc database. *EcoSal Plus*, *8*(1), 10.1128/ecosalplus.ESP-0006-2018. doi:10.1128/ecosalplus.ESP-0006-2018

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martinez, C., Caspi, R., . . . Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res, 45*(D1), D543-D550. doi:10.1093/nar/gkw1003

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

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.e8. doi:10.1016/j.molcel.2019.05.031

Nichols, R. J., Sen, S., Choo, Y. J., Beltrao, P., Zietek, M., Chaba, R., . . . Gross, C. A. (2011). Phenotypic landscape of a bacterial cell. *Cell, 144*(1), 143-156. doi:10.1016/j.cell.2010.11.052

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

Pesquita, C. (2017). Semantic similarity in the Gene Ontology. *Methods Mol Biol*, *1446*, 161-173. doi:https://doi.org/10.1007/978-1-4939-3743-1_12

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., . . . Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature, 557*(7706), 503-509. doi: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. doi:10.1186/1471-2105-8-111

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

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. doi:10.1371/journal.pone.0118432

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

Shefchek, K. A., Harris, N. L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., . . . 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. doi:10.1093/nar/gkz997

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. doi:10.1186/s13326-019-0205-5

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. doi:10.3389/fmicb.2016.01672

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. doi:10.1093/bioinformatics/btm087

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Retrieved from https://ggplot2.tidyverse.org

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. doi:10.1093/bioinformatics/btq064

TABLES AND FIGURES

Table 1 Non-co-annotated gene pairs with |PCC| >0.75

Gene pair[1]	**Known or predicted functional association**
ECK0730-*pal*_ECK0725-*ybgC*[2]	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*)[3]	ubiquinol-8 biosynthesis (PWY-6708)
ECK3610-*rfaF*(*waaF*)_ECK3042-*rfaE*(*hldE*)[4]	superpathway of lipopolysaccharide biosynthesis (LPSSYN-PWY)
ECK3610-*rfaF*(*waaF*)_ECK0223-*lpcA*[4]	super pathway of lipopolysaccharide biosynthesis (LPSSYN-PWY)
ECK3852-*dsbA*_ECK1173-*dsbB*	periplasmic disulfide bond formation (PWY0-1599)[5]
ECK1544-*gnsB*_ECK2394-*gltX*	unknown
ECK2066-*yegK*(*pphC*)_ECK0345-*mhpB*	unknown
ECK3699-*mnmE*_ECK0050-*apaH*	unknown

[1] 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.

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

[3] *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

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

[5] PWY0-1599 was not present in EcoCyc version 21.1

Table 2. Annotation sets used in this study

Annotation set (source)	Subset of annotated genes tested[a]	Total no. of annotations for each subset[b]
Pathways (EcoCyc)	885	2,317
Heterooligomeric protein complexes (EcoCyc)	688[c]	871[c]
Operons (RegulonDB)	3,858	5,349
Regulons (RegulonDB)	1,572	3,886
Modules (KEGG)	333	524
Pathways or Protein complexes	1,385	3,269
Pathways and Protein Complexes	188	818[d]
Any (Union of all 5 annotation sets)	3,866	12,937
All (Intersection of all 5 annotation sets)	77	922[d]
GO biological process	2,609	5,775

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

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

[c] This excludes 681 genes annotated to protein complexes whose products form only homooligomeric complexes

[d] This is the number of annotations associated with any of the 77 genes that are annotated to all 5 annotation sets.

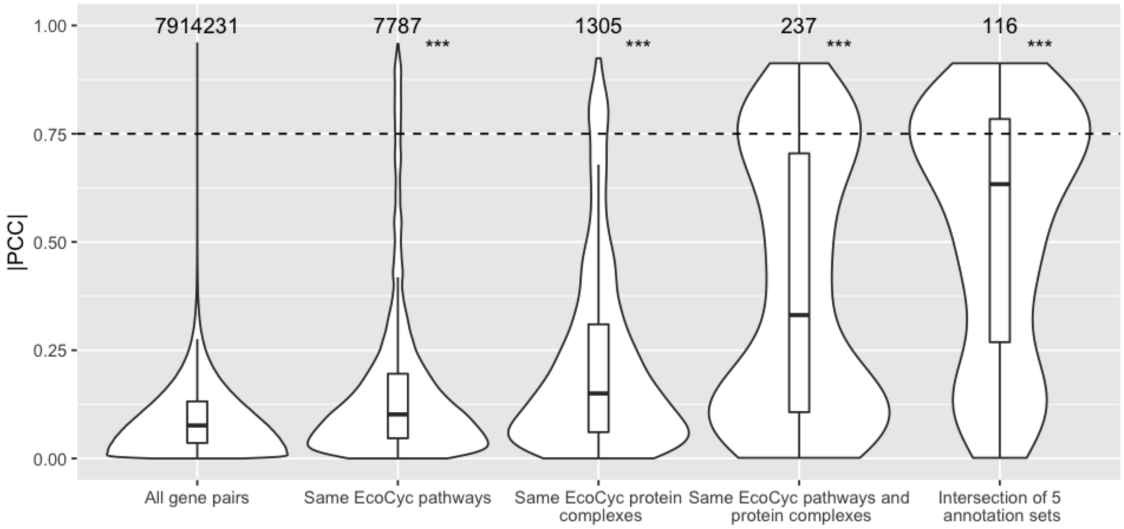
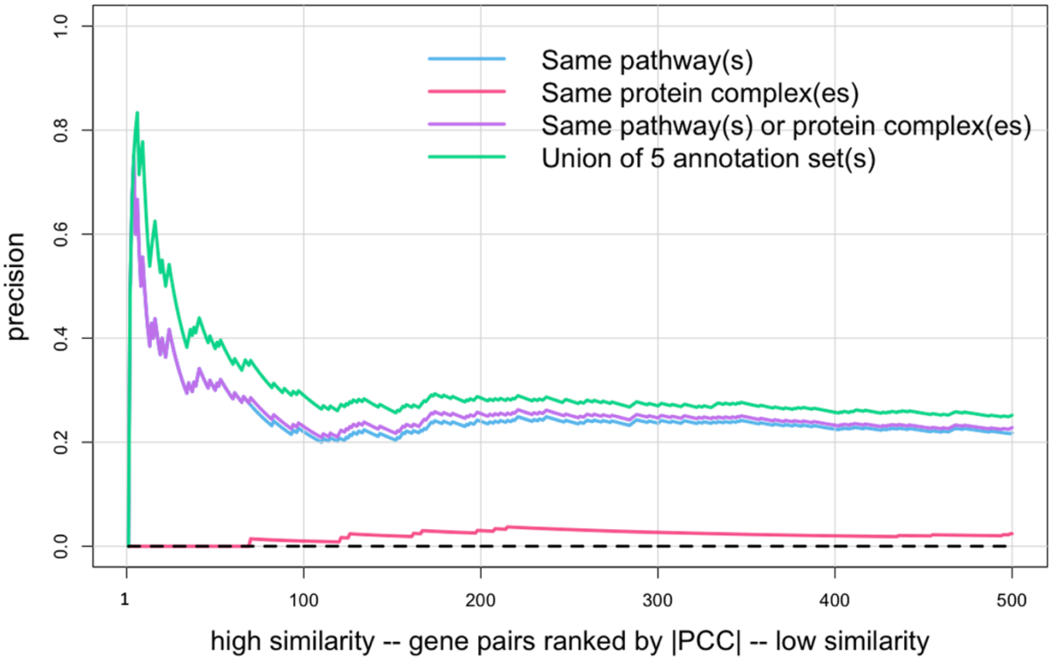
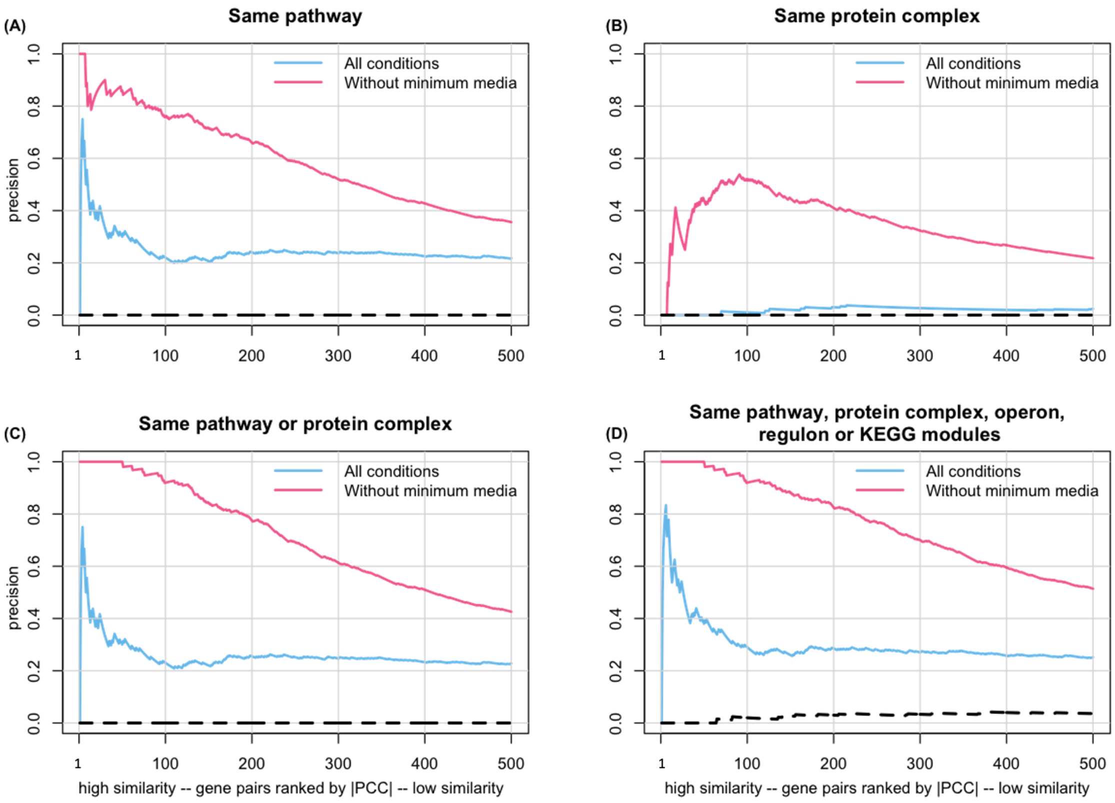


Figure 1. Higher phenotypic similarity was found for co-annotated gene pairs. Shown are violin plots of the distributions of |PCC| for the indicated groups of gene pairs. Numbers above each violin plot indicate the number of gene pairs in each plot. ***: p-value <0.001 was determined by 1-sided Mann-Whitney U test, compared to all gene pairs. The dashed line indicates |PCC| = 0.75, which was chosen as an arbitrary cut-off.



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

Figure 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 [TP/(TP+FP)], which was calculated as described in the text (only the first 500 gene pairs are shown). 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 (negative control). The correspondence between |PCC| and ranking is shown below the graph.



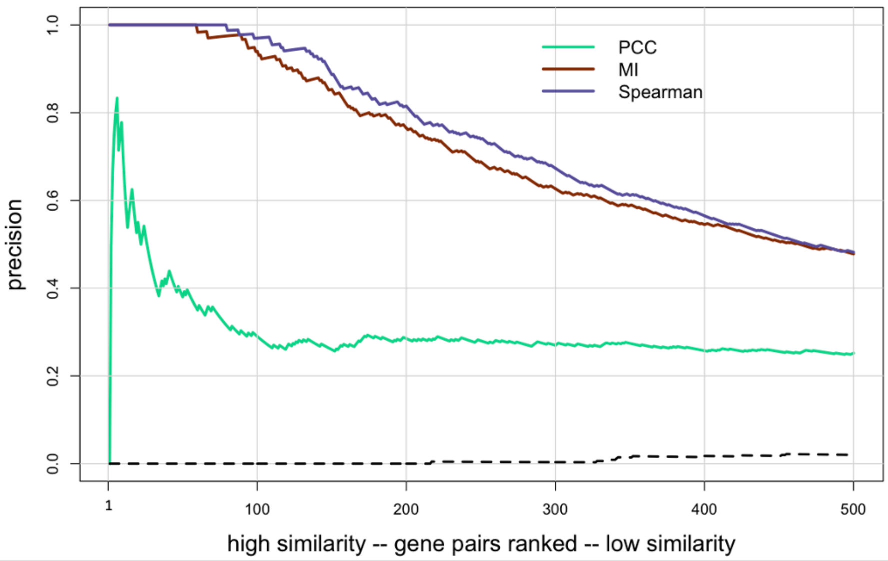
680

681

682 **Figure 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 dashed line shows precision for randomly ordered gene pairs (negative control). The correspondence between |PCC| and ranking is the same as in Figure 2.

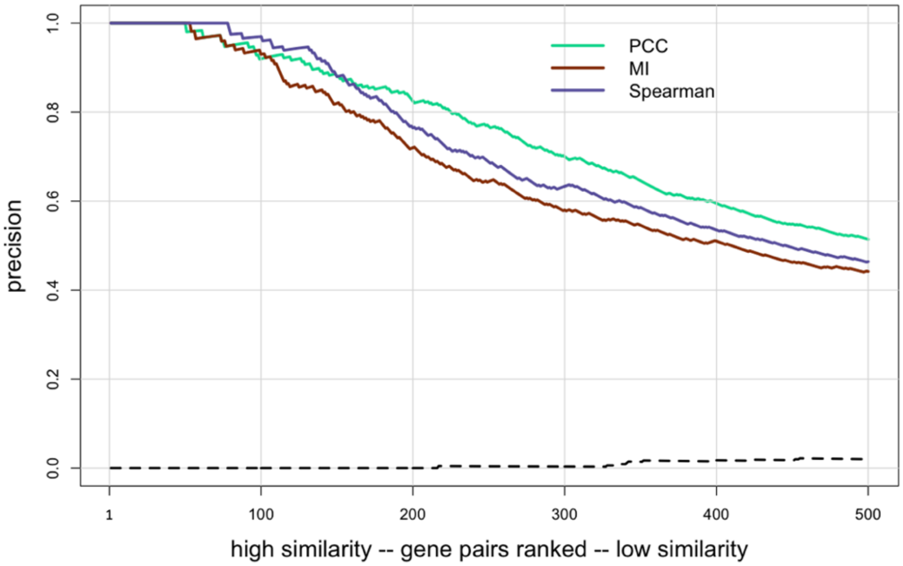
687

Figure 4a



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

Figure 4b



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 4. Precision versus ranking for different methods of measuring phenotype profile similarity. Gene pairs were ranked from high to low similarity and plotted versus precision, calculated as described in the text (only the first 500 gene pairs are shown). Phenotypic profile similarity was assessed using either |PCC|, MI, or |SRCC| with (a) all growth conditions used or (b) excluding growth conditions with minimal media. The dashed line shows precision for randomly ordered gene pairs (negative control). The correspondence between similarity scores and ranking is shown below each graph.

Figure 5a

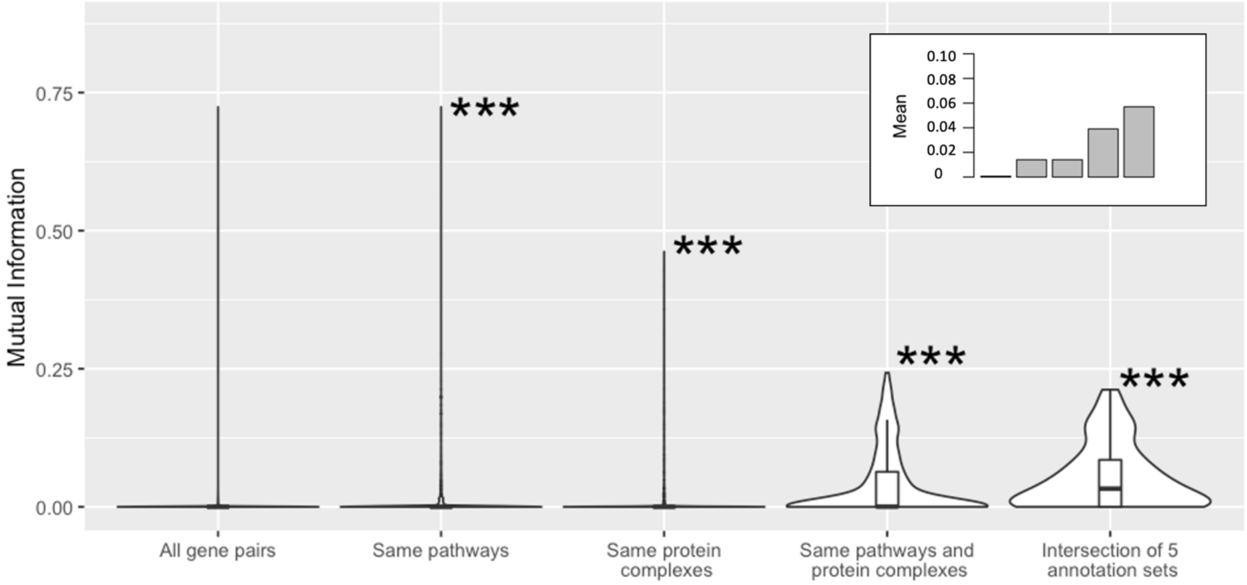


Figure 5b

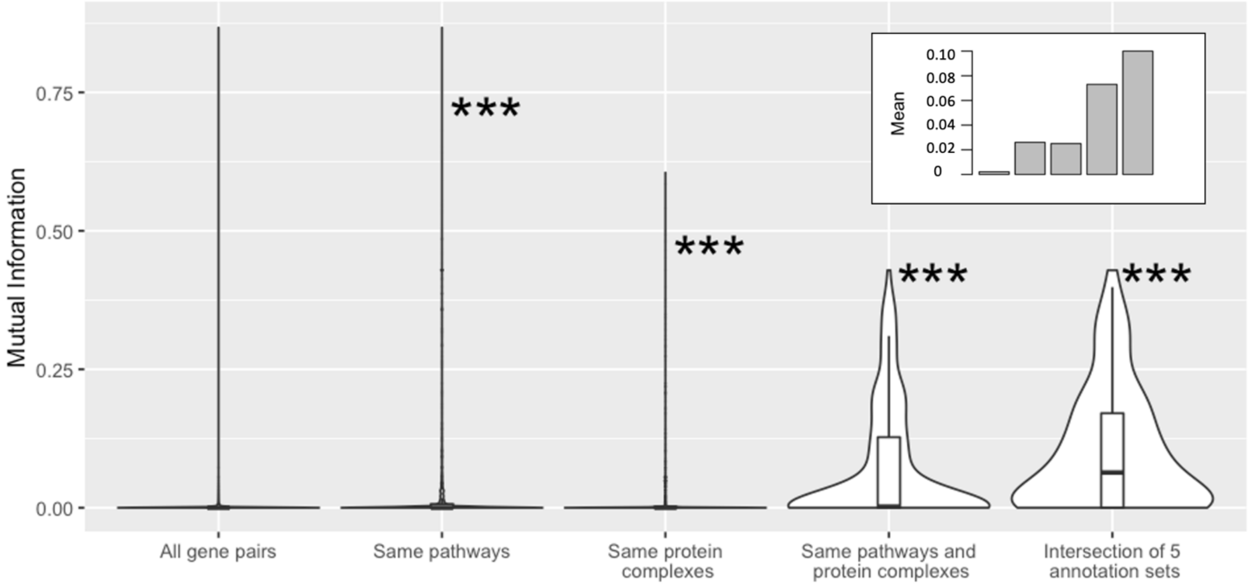
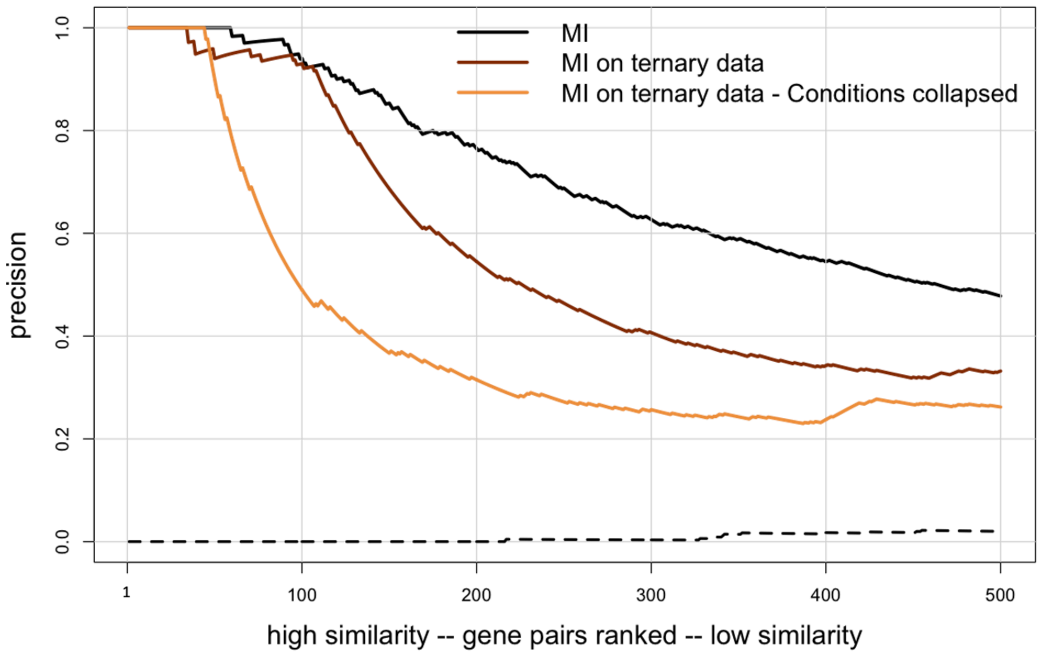


Figure 5. Phenotypic profile similarity after converting fitness scores from quantitative to qualitative, ternary values. Shown are violin plots of the distributions of phenotypic profile similarity based on Mutual Information for the indicated groups of gene pairs. Panel (a) shows results determined using all 324 growth conditions, and panel (b) shows results determined after collapsing the growth conditions to 114 unique stresses. The insets show the mean value for each distribution. For (a) the mean values are 0.0006, 0.014, 0.014, 0.039, and 0.057). For (b) the mean values are 0.0021, 0.026, 0.025, 0.073, and 0.1). ***: p-value <0.001 determined by 1-sided Mann-Whitney U test.



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 6. Precision versus ranking for quantitative versus qualitative, ternary fitness scores. Gene pairs were ranked from high to low similarity based on Mutual Information (MI) and plotted versus precision, calculated as described in the text (only the first 500 gene pairs are shown). The phenotypic profiles contained either the original quantitative data (black line), the discretized ternary values for all growth conditions (brown line), or the discretized, ternary values for growth conditions collapsed to 114 unique stresses (orange line). The dashed line shows precision for randomly ordered gene pairs (negative control). The correspondence between similarity scores and ranking is shown below each graph.

Figure 7a

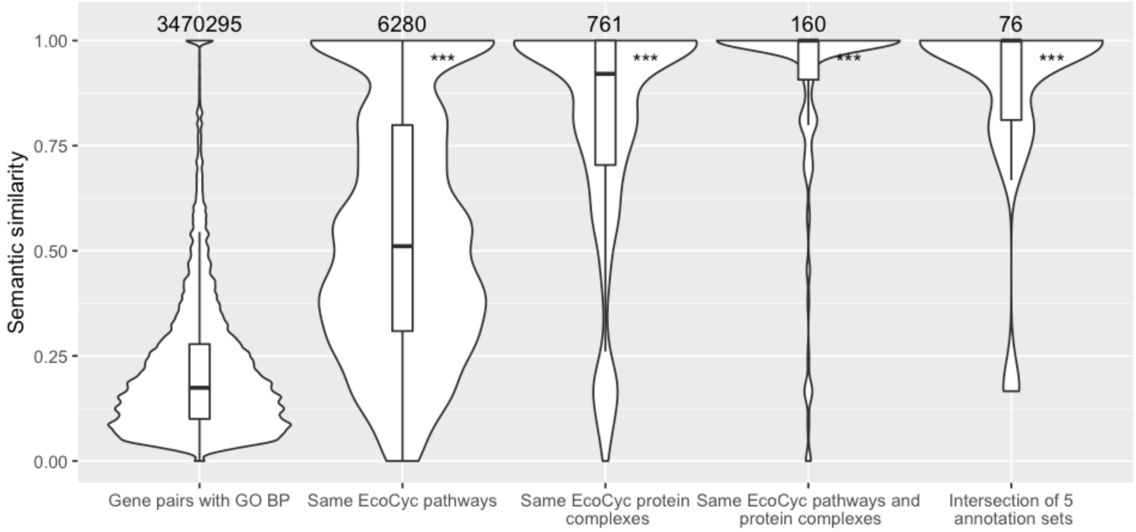


Figure 7b

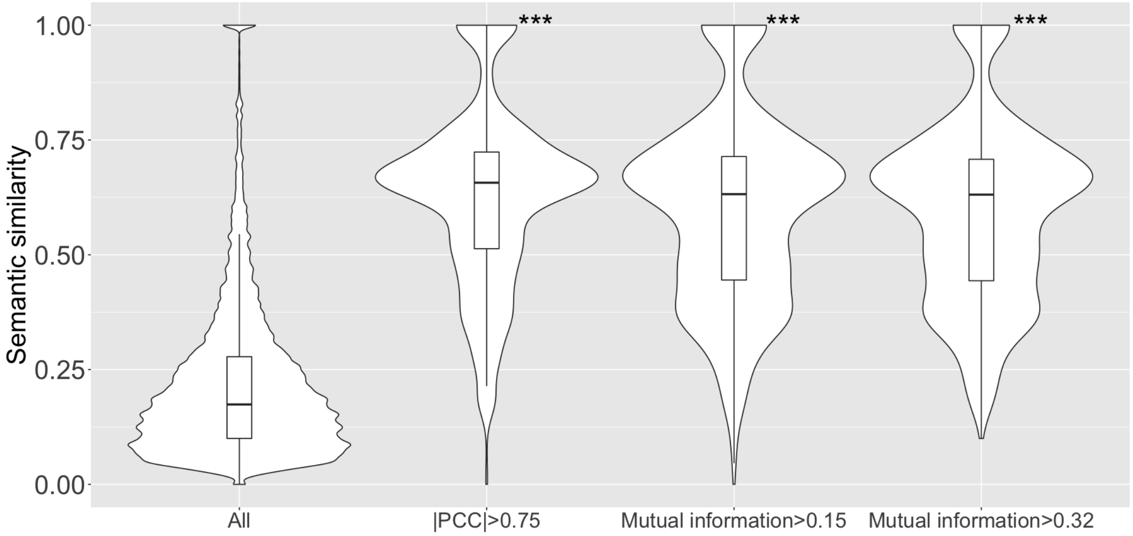


Figure 7. Higher semantic similarity and phenotypic profile similarity were found for co-annotated gene pairs. (a) Violin plots of the distributions of semantic similarity for the indicated groups of gene pairs. Numbers above each violin plot indicate the number of gene pairs in each plot. (b) Violin 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 (~1,000) of top-ranked gene pairs. ***: p-value <0.001 was determined by 1-sided Mann-Whitney U test, compared to all gene pairs.

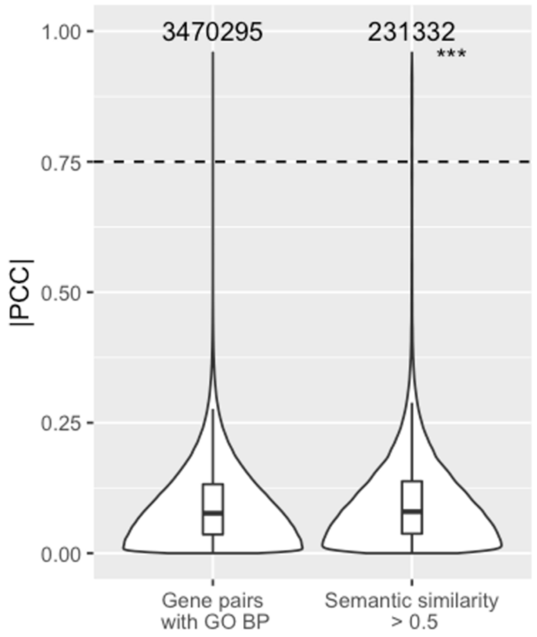


Figure 8. Higher phenotypic similarity was found for gene pairs that have higher GO semantic similarity. Violin plots of the distributions of the |PCC| values for all gene pairs with GO biological process annotations and the subset with semantic similarity is greater than an arbitrary cutoff of 0.5. 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.

Supplemental Tables and Figures

Supplemental Table S1. EcoCyc Pathways IDs and name of pathway for the labels used in Supplemental Figure S1.

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	preQ0 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

761

Supplemental Table S2. EcoCyc protein complex IDs and name of protein complex for the labels used in Supplemental Figure S2.

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	Zn2+ 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&mdash;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 sigma 54
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&mdash;[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

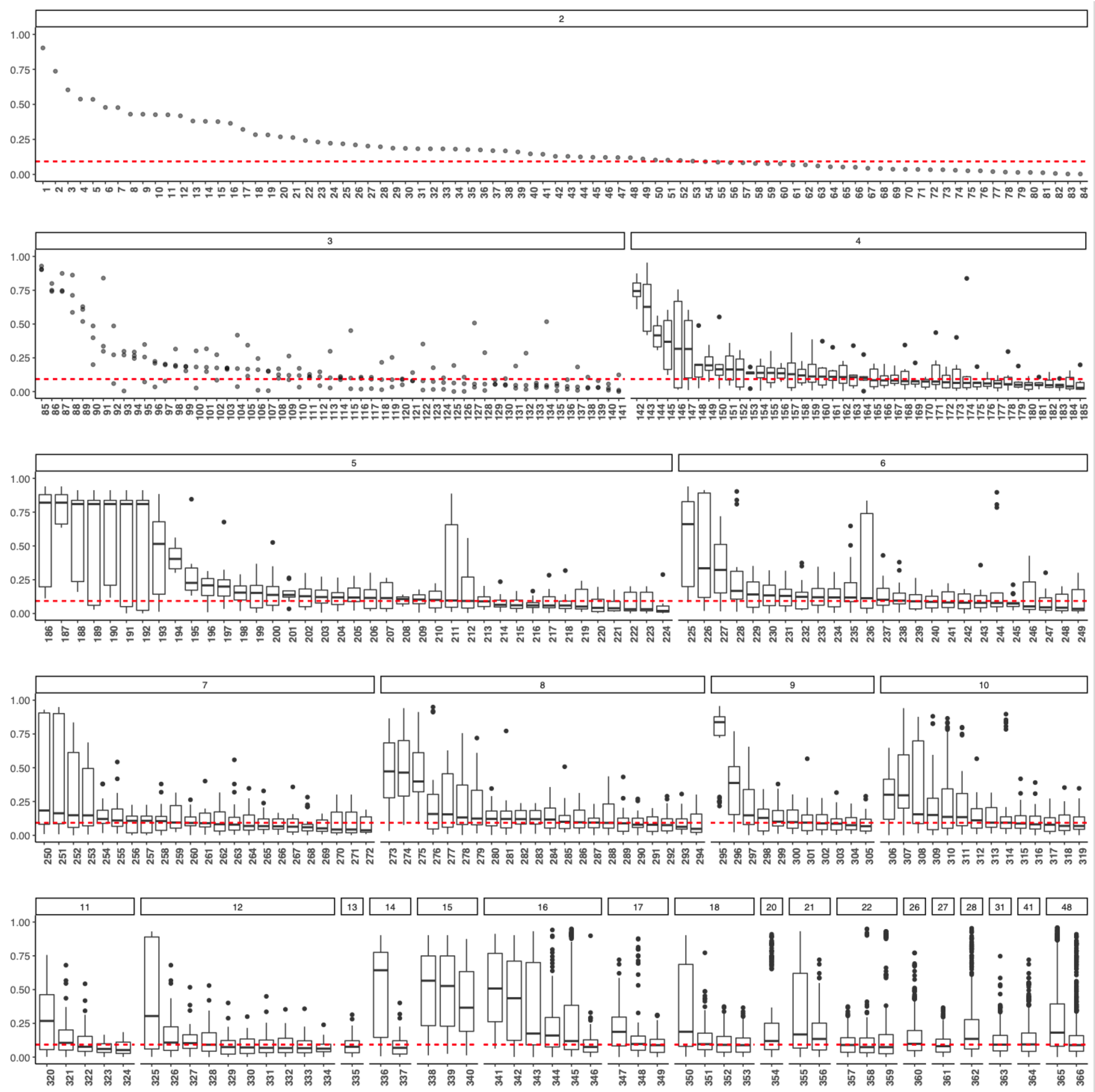
124	RUVABC-CPLX	resolvasome
125	CPLX0-7450	flagellar motor switch complex
126	ABC-18-CPLX	D-galactose / methyl-&beta;-D-galactoside ABC transporter
127	CPLX0-1923	energy transducing Ton complex
128	CPLX0-1924	vitamin B12 outer membrane transport complex
129	MUTHLS-CPLX	MutHLS complex, methyl-directed 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	Tsrglu-Me
152	TSR-GLN	Tsrgln
153	TSR-GLU	Tsrglu
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	Tapglu
162	TAP-CPLX	chemotaxis signaling complex - dipeptide sensing
163	TAP-GLUME	Tapglu-Me
164	TAP-GLN	Tapgln
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	Trgglu-Me
176	TRG-GLN	Trggln
177	TRG-GLU	Trgglu
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	Targlu-Me
198	TAR-GLN	Targln
199	TAR-GLU	Targlu
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

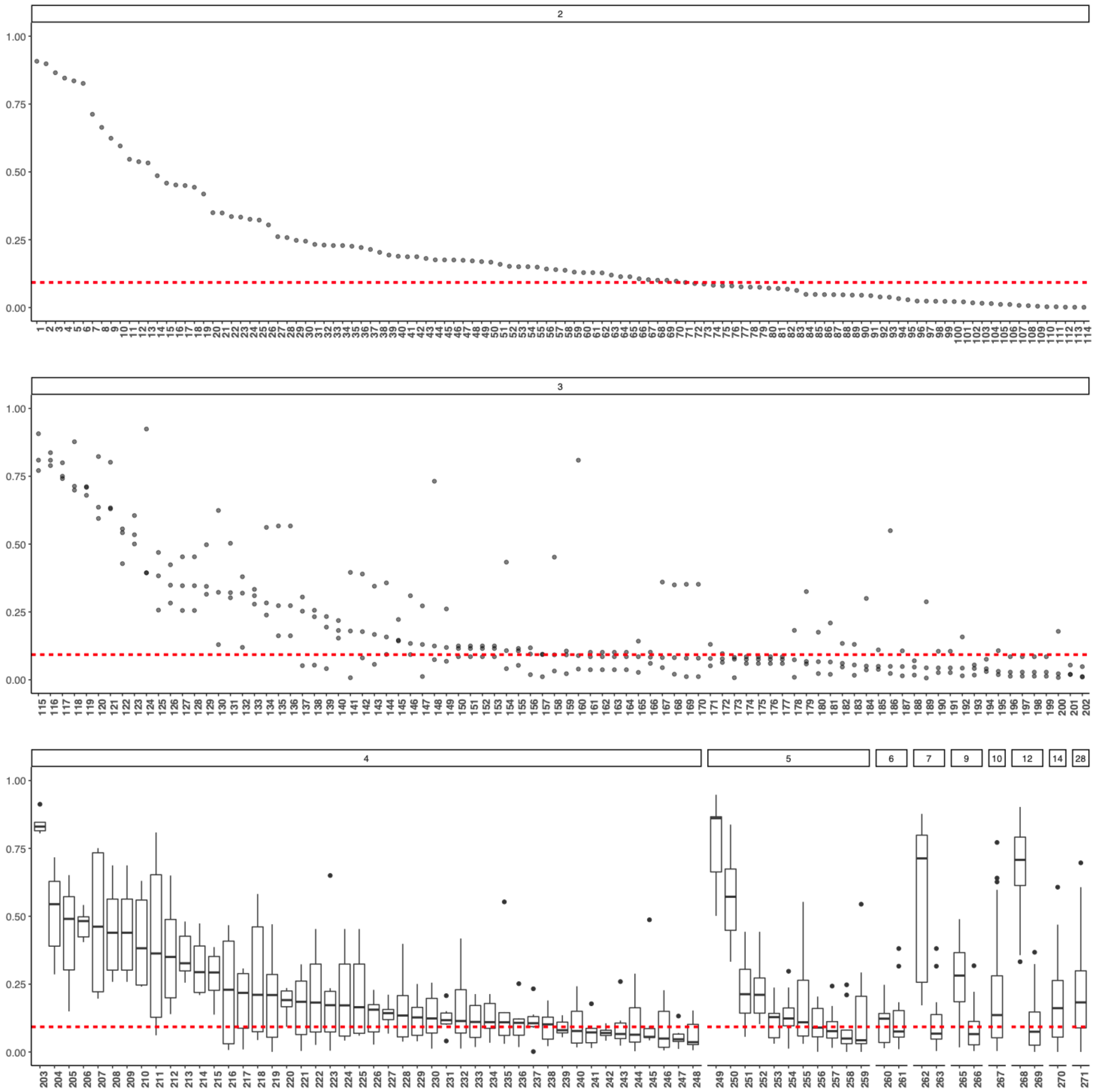
227	CPLX0-1762	phenylacetyl-CoA 1,2-epoxidase
228	ABC-29-CPLX	putrescine ABC exporter
229	ABC-55-CPLX	putative transport complex, ABC superfamily
230	CPLX0-7958	methylphosphonate degradation complex
231	HCAMULTI-CPLX	putative 3-phenylpropionate/cinnamate dioxygenase
232	CPLX0-7935	carbon-phosphorus lyase 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

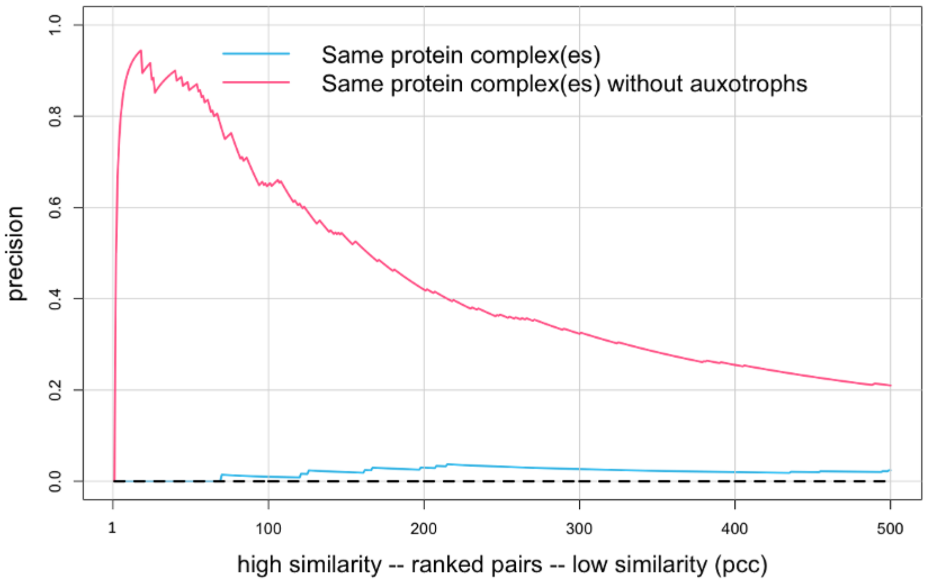


Supplemental Figure S1. Phenotypic profile similarity for genes in the same heteromeric protein complex. The distribution of phenotypic profile similarity values determined by |PCC| for all pairwise combinations of genes assigned to the same

EcoCyc pathway. In the figure, the pathways are sorted by (i) the number of genes in the pathway and then (ii) the median |PCC| value. The names of the pathways are indicated by numeric labels, which are defined in supplemental Table S1. The dashed line shows the average |PCC| value for random pairs of genes. 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.

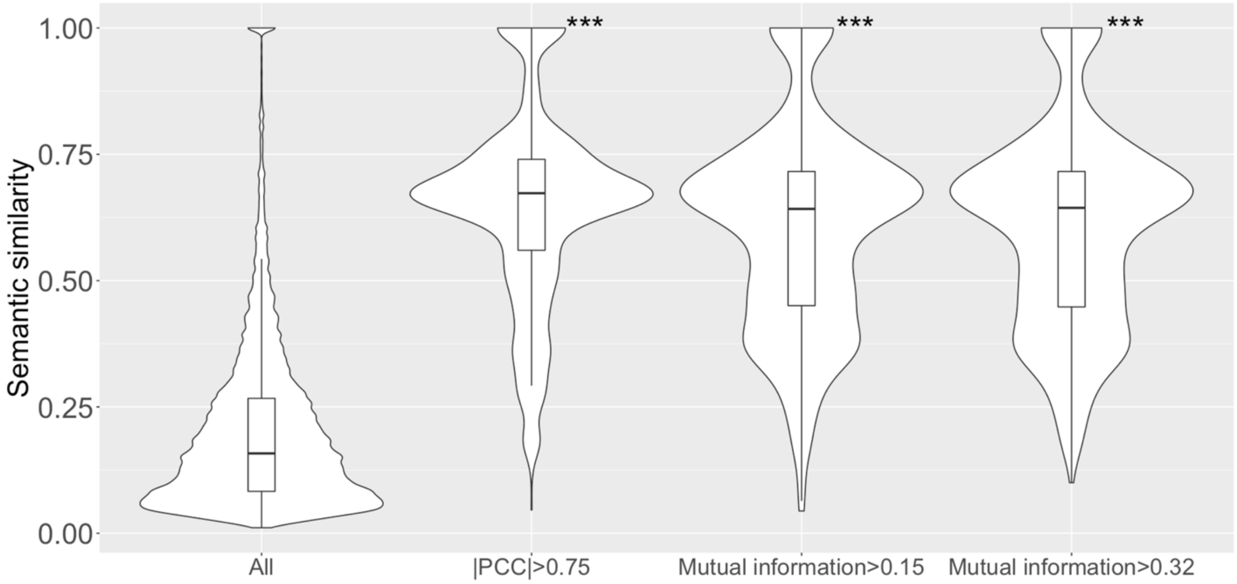


Supplemental Figure S2. Phenotypic profile similarity for genes in the same EcoCyc heteromeric protein complex. The distribution of phenotypic profile similarity values determined by |PCC| for all pairwise combinations of genes assigned to the same EcoCyc heteromeric protein complex. In the figure, the pathways are sorted by (i) the number of genes in the complex and then (ii) the median |PCC| value. The names of the complexes are indicated by numeric labels, which are defined in supplemental Table S2. The dashed line shows the average |PCC| value for random pairs of genes. 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.



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

Supplemental Figure S3. Precision increased when auxotrophic mutants 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 dashed line shows precision for randomly ordered gene pairs (negative control). The correspondence between phenotypic profile similarity based on |PCC| and ranking is shown below the graph.



Supplemental Figure S4. Higher semantic similarity and phenotypic profile similarity were still found when GO biological process annotations with an IEA evidence code were excluded. 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 (~1,000) of top-ranked gene pairs. ***: p-value <0.001 was determined by 1-sided Mann-Whitney U test, compared to all gene pairs.