SUPPLEMENTARY MATERIAL

1. SUPPLEMENTARY METHODS

1.1. Selection of GEO (Gene Expression Omnibus) Datasets: To scale GeneMANIA data processing and cope with the task of collecting and maintaining large datasets containing hundreds to thousands of networks for multiple organisms on a regular basis, we automate the process of data retrieval and processing to run with minimal manual intervention. A predefined set of parameters and data processing procedures were used to control this process. In the case of GEO, our data selection procedure is as follows:

1. A set of GEO platforms is specified from which to select studies based on an organism of interest, and the availability of platform metadata required for processing in standardized format and location.
2. For the given set of platforms, we select GSE data series containing a minimum of 12 samples. Interactions between genes are computed by correlation, and the above threshold on the number of samples is to aid the reliability of the inferred correlations.
3. Datasets are discarded for which standardized publication metadata is unavailable. Most notably, we require a PubMed reference providing attribution, automated naming, the description of networks, and link outs.
4. Any datasets not available in series matrix text format with standard naming and location for automated retrieval and ease of bulk processing are discarded.

By applying these criterion, we selected GEO platforms for "*E. coli* K-12": GPL73, GPL199, and GPL534 that were suitable for processing by our tools, with the resulting networks provided in our dataset (Supplementary Table S1). Future versions of GeneMANIA may be enhanced to broaden the collection of datasets automatically included. However, users may upload their own networks via the advanced options panel on the website or in the Cytoscape application.

1.2. AUC evaluation for STRING functional association networks: STRING does not directly predict a function for a gene; rather, it assigns a measure of functional similarity between pairs of genes. This contrasts with GeneMANIA, which ranks each gene according to their functional similairty to an entire set of input genes. Nonetheless, we did attempt to evaluate the relative accuracy of STRING compared to GeneMANIA by evaluating - for each gene in the STRING network - the maximum STRING score with an adjacent, annotated gene. To do so, we first transformed the functional association networks for the *E. coli* W3110 strain, fetched from http://string-db.org on Jul 12, 2014 (filename: protein.links.v9.1.txt.gz), into functional predictions as follows:

For each protein, P, in the *Escherichia coli* proteome, we evaluated the overall similarity to a Gene Ontology (GO) biological process, cellular component, or molecular function annotation, A, by taking the maximum STRING score of an adjacent edge that spanned that protein, and a protein with the annotation of interest. We denote this score STRING-max:

STRING-max(P, A) = max(0, string(P, P1), string(P, P2), ..., string(P, PN))

where $P_1 ... P_N$ are proteins annotated to A. We then evaluate the AUC as was done for the GeneMANIA scores.

While our particular method of converting the gene-gene STRING scores to a gene-annotation score may not be optimal, we emphasize that any researcher attempting to make direct functional predictions from STRING would face a similar difficulty. Investigating more advanced means of translating STRING networks to functional predictions is outside the scope of this manuscript, which discusses a GeneMANIA implementation aimed directly at making these kinds of functional predictions without further processing. We also evalated the value of using the STRING network as input evidence for the GeneMANIA algorithm, and found lower average error than the STRING-max predictions (data not shown). However, as STRING relies at least partly on text-mining of the same literature from which GO annotations are curated, this error may be underestimated, and we have therefore elected to exclude it. Nonetheless, we do plan to make the STRING network available as an optional evidence source in the GeneMANIA web application.

1.3. Integration of additional prokaryotes in the future: While we intend to extend GeneMANIA by integrating opportunistic pathogens such as *Helicobacter pylori*, *Campylobacter jejuni*, and *Mycobacterium tuberculosis* for which we have compiled large-sets of experimentally derived biochemical interactions (see Supplementary Table S7), integration of additional prokaryotes depends on how much genomic and interactomic data is available for them, and how close their evolutionary relationship with *E. coli* is. For example, if the organism is poorly studied and has very little information available, then any such predictions would rely almost exclusively on orthology mapping and GeneMANIA would likely offer little benefit. In contrast, assuming that the organism is at least partly or well-studied, then the question becomes how to integrate the *E. coli* input networks with the additional information specific to the other organisms. For this, the *E. coli* input networks can be "converted" into networks in the new organism according to sequence orthology. That is edges are only retained if both adjacent genes/proteins have a sequence ortholog. These orthologous networks can then be integrated with the other evidences (networks) available for the new organism using the GeneMANIA algorithm, which will weight the various networks appropriately according to their predictive power.

2. SUPPLEMENTARY TABLES

Supplementary Table S1 - List of input networks available for GeneMANIA function prediction.

Supplementary Table S2 - Average error (1-AUC) of GeneMANIA and STRING-max in recapitulating GO annotations.

Supplementary Table S3 - Change in AUC resulting from the omission of evidence categories.

Supplementary Table S4 - Top 100 GeneMANIA predictions for GO annotations.

Supplementary Table S5 - The mean value of each data point from three independent replicates along with their standard deviation per growth curve experiments.

Supplementary Table S6 - Comparison of experimentally confirmed novel predictions from GeneMANIA to STRING or eNET

Supplementary Table S7 - Compilation of literature curated interactions from various opportunistic pathogens can be used in GeneMANIA to predict gene functions either directly or through orthology mapping.

3. SUPPLEMENTARY FIGURES

Supplementary Fig. 1 - Overlap between evidence types supporting functional prediction

Supplementary Fig. 2 - Novel factors in ribosome biogenesis and cell adhesion. (**A**) Sub-network of YihD linked with the tRNA and rRNA methylation factors (i) based on SPD, Exp, and other network (ii) sources (e.g., phenomics and genomic context). Ribosome profiles (iii) of wild type

(WT) and *yihD* mutant strain is shown with their corresponding subunit peak ratios. Elevated translation errors (iv) of the *yihD* mutant strain (normalized to wild type) based on read-through using a β-galactosidase reporter system. Each data point represents the mean ± SD (error bars) of three independent biological measurements. Asterisks indicate a significant difference between the *yihD* mutant and wild type strains; *p*-value calculated using the Student's t-test. The EF4 (LepA) translation elongation factor served as positive control. (**B**) Sub-network of YdeT, YdhQ, and YhjY showed strong connectivity with the cell adhesion factors (i) based on SPD and Exp (ii). Crystal violet staining (iii) of indicated mutant strains showing the bacterial biofilm mass on the polystyrene surface just as the positive control biofilm mutant, *ompA*.

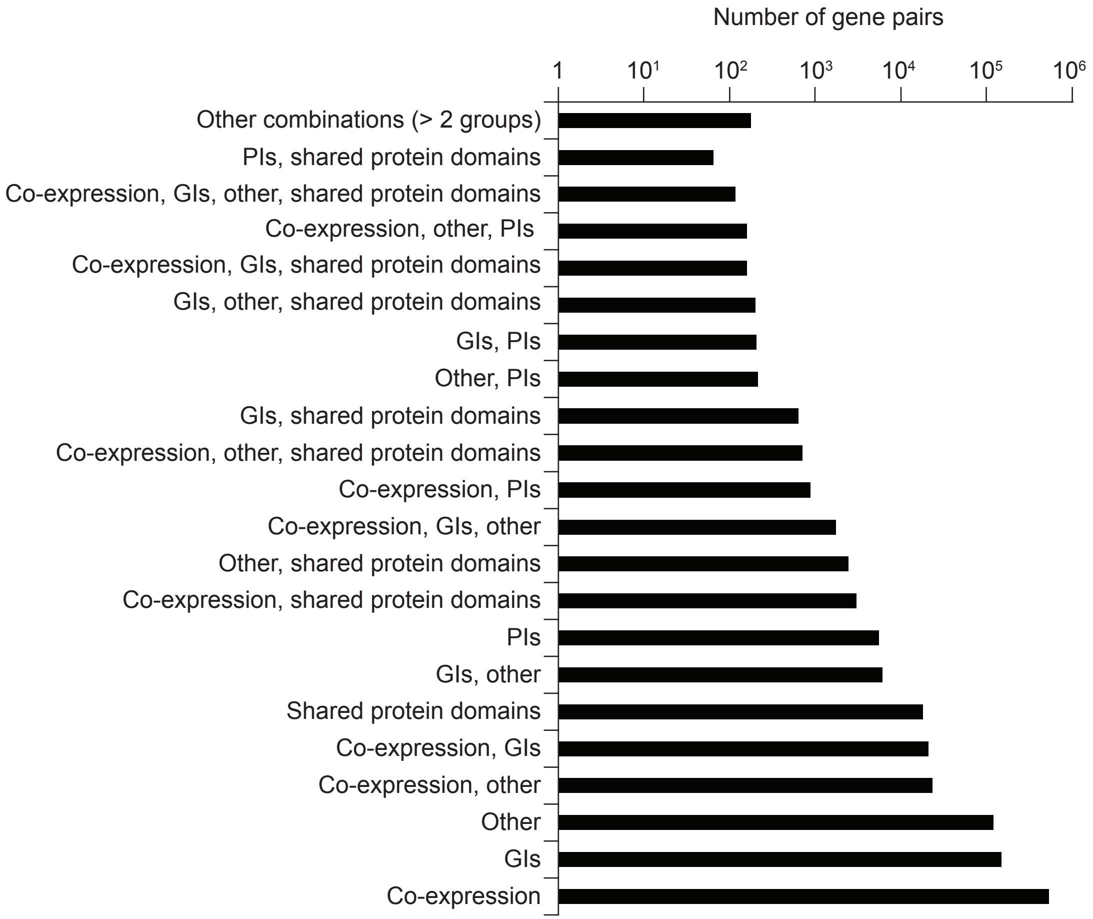
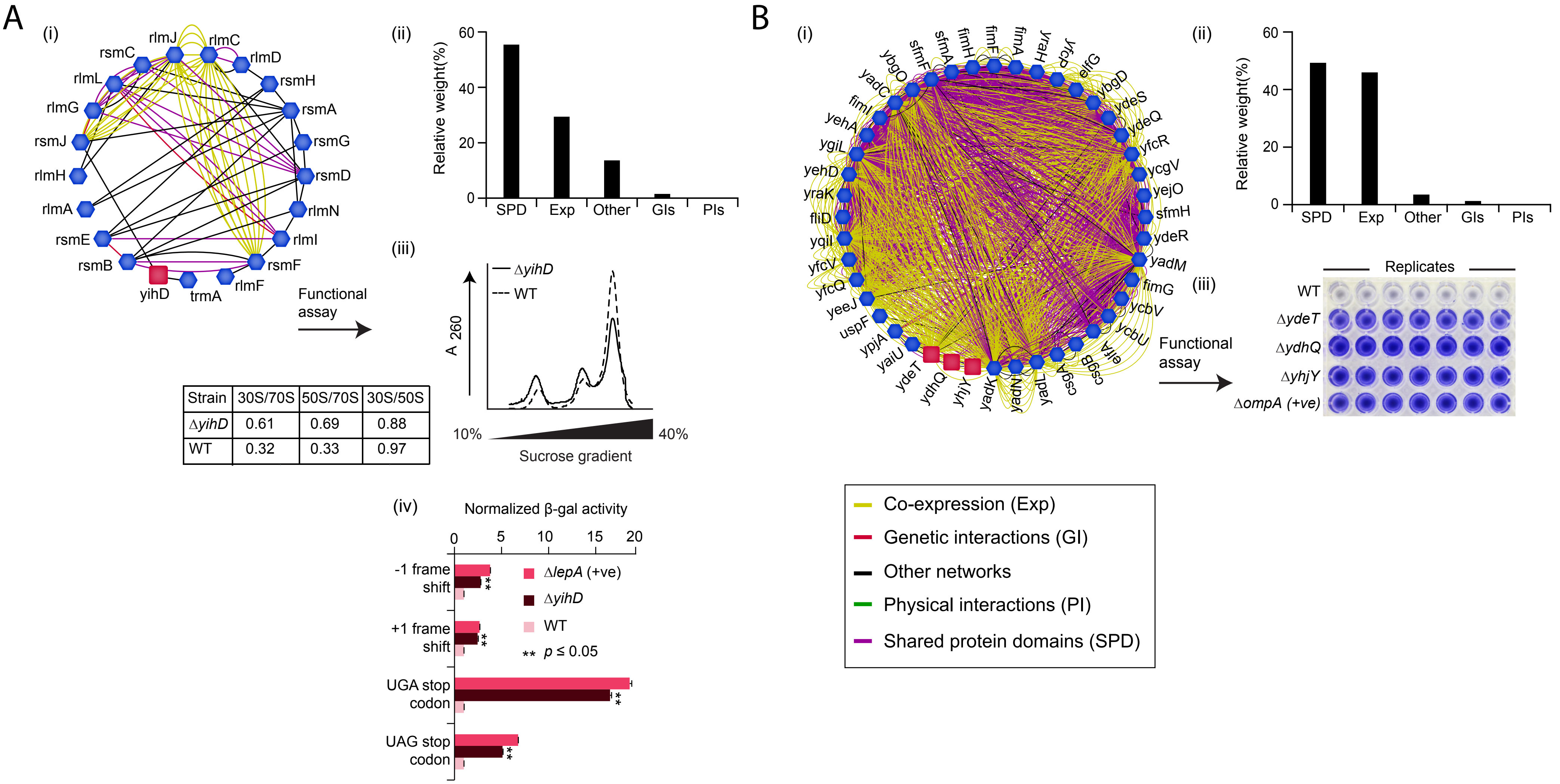


Figure S1. Overlap between evidence types supporting functional prediction. Exp: co-expression; GI: genetic interactions; PI: physical interactions; and SPD: shared protein domains.

Supplementary Figure 1



Strain	30S/70S	50S/70S	30S/50S
ΔyihD	0.61	0.69	0.88
WT	0.32	0.33	0.97

Figure S2. Novel factors in ribosome biogenesis and cell adhesion. (**A**) Sub-network of *yihD* linked with the tRNA and rRNA methylation factors (i) based on SPD, Exp, and other network (ii) sources (e.g., phenomics and genomic context). Ribosome profiles (iii) of wild type (WT) and *yihD* mutant strain is shown with their corresponding subunit peak ratios. Elevated translation errors (iv) of the *yihD* mutant strain (normalized to wild type) based on read-through using a β-galactosidase reporter system. Each data point represents the mean ± SD (error bars) of three independent biological measurements. Asterisks indicate a significant difference between the *yihD* mutant and wild type strains; *p*-value calculated using Student's t-test. The EF4 (LepA) translation elongation factor served as a positive control. (**B**) Sub-network of *ydeT*, *ydhQ*, and *yhjY* show strong connectivity with the cell adhesion factors (i) based on SPD and Exp (ii). Crystal violet staining (iii) of indicated mutant strains showing the bacterial biofilm mass on the polystyrene surface just as the positive control biofilm mutant, *ompA*.

Supplementary Figure 2