4-2009

Global Functional Atlas of *Escherichia coli* Encompassing Previously Uncharacterized Proteins

Pingzhao Hu
University of Toronto

Mohan Babu
University of Toronto

Sarah Chandra Janga
University of Toronto

Gareth Butland
University of Toronto

Wenhong Yang
University of Toronto

See next page for additional authors

Recommended Citation

Hu, Pingzhao; Babu, Mohan; Janga, Sarah Chandra; Butland, Gareth; Yang, Wenhong; Pogoutse, Oxana; Guao, Xinghua; Phanse, Sadhna; Wong, Peter; Chandran, Shamanta; Christopolous, Constantine; Nazarians-Armavil, Anaies; Musso, Gabriela; Ali, Mehrab; Nazemof, Nazila; Eroukova, Veronika; Golshani, Ashkan; Paccanaro, Alberto; Greenblatt, Jack F.; Moreno-Hagelsieb, Gabriel; and Emili, Andrew, "Global Functional Atlas of *Escherichia coli* Encompassing Previously Uncharacterized Proteins" (2009). *Biology Faculty Publications*. 4.
https://scholars.wlu.ca/biol_faculty/4

Authors

Pingzhao Hu, Mohan Babu, Sarah Chandra Janga, Gareth Butland, Wenhong Yang, Oxana Pogoutse, Xinghua Guao, Sadhna Phanse, Peter Wong, Shamanta Chandran, Constantine Christopolous, Anaies Nazarians-Armavil, Gabriela Musso, Mehrab Ali, Nazila Nazemof, Veronika Eroukova, Ashkan Golshani, Alberto Paccanaro, Jack F. Greenblatt, Gabriel Moreno-Hagelsieb, and Andrew Emili

Global Functional Atlas of *Escherichia coli* Encompassing Previously Uncharacterized Proteins

Pingzhao Hu[1☯], Sarath Chandra Janga[1,2☯], Mohan Babu[1☯], J. Javier Díaz-Mejía[1,3☯], Gareth Butland[1☯¤], Wenhong Yang[1], Oxana Pogoutse[1], Xinghua Guo[1], Sadhna Phanse[1], Peter Wong[1], Shamanta Chandran[1], Constantine Christopoulos[1], Anaies Nazarians-Armavil[1], Negin Karimi Nasseri[1], Gabriel Musso[1], Mehrab Ali[1], Nazila Nazemof[4], Veronika Eroukova[4], Ashkan Golshani[4], Alberto Paccanaro[5], Jack F. Greenblatt[1], Gabriel Moreno-Hagelsieb[3*], Andrew Emili[1*]

1 Banting and Best Department of Medical Research, Terrence Donnelly Center for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada, **2** Medical Research Council Laboratory of Molecular Biology, Cambridge, United Kingdom, **3** Department of Biology, Wilfrid Laurier University, Waterloo, Ontario, Canada, **4** Department of Biology and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Canada, **5** Department of Computer Science, Royal Holloway, University of London, Egham, United Kingdom

One-third of the 4,225 protein-coding genes of *Escherichia coli* K-12 remain functionally unannotated (orphans). Many map to distant clades such as Archaea, suggesting involvement in basic prokaryotic traits, whereas others appear restricted to *E. coli*, including pathogenic strains. To elucidate the orphans' biological roles, we performed an extensive proteomic survey using affinity-tagged *E. coli* strains and generated comprehensive genomic context inferences to derive a high-confidence compendium for virtually the entire proteome consisting of 5,993 putative physical interactions and 74,776 putative functional associations, most of which are novel. Clustering of the respective probabilistic networks revealed putative orphan membership in discrete multiprotein complexes and functional modules together with annotated gene products, whereas a machine-learning strategy based on network integration implicated the orphans in specific biological processes. We provide additional experimental evidence supporting orphan participation in protein synthesis, amino acid metabolism, biofilm formation, motility, and assembly of the bacterial cell envelope. This resource provides a "systems-wide" functional blueprint of a model microbe, with insights into the biological and evolutionary significance of previously uncharacterized proteins.

Citation: Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, et al. (2009) Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. PLoS Biol 7(4): e1000096. doi:10.1371/journal.pbio.1000096

Introduction

Because of its central position in the microbial research community, the Gram-negative bacterium *Escherichia coli* plays a leading role in investigations of the fundamental molecular biology of bacteria [1–8]. This experimentally tractable microbe is a workhorse in basic and applied research aimed at elucidating the mechanistic basis of prokaryotic processes and traits, including those of pathogens. The ever-expanding availability of genomic resources makes *E. coli* particularly well-suited to systematic investigations of microbial protein components and functional relationships on a global scale. These include a genome-wide collection of single-gene deletion strains [2] along with extensive knowledge of regulatory circuits [3,5,7,9] and metabolic pathways [6,10,11].

Yet despite being the most highly studied model bacterium, a recent comprehensive community annotation effort for the fully sequenced reference K-12 laboratory strains [8] indicated that only half (~54%) of the protein-coding gene products of *E. coli* currently have experimental evidence indicative of a biological role. The remaining genes have either only generic, homology-derived functional attributes (e.g., "predicted DNA-binding") or no discernable physiological significance. Some of these functional "orphans" (not to be confused with ORFans, which are genes present within only single or closely related species) may have eluded characterization in part because they exhibit mild mutant phenotypes, are expressed at low or undetectable levels, or have limited homology to annotated genes. This suggests more-sensitive analytical procedures are warranted.

A key feature of the molecular organization of all organisms, including bacteria, is the tendency of gene products to associate into macromolecular complexes, bio-

Academic Editor: Andre Levchenko, Johns Hopkins University, United States of America

Received October 21, 2008; **Accepted** March 16, 2009; **Published** April 28, 2009

Abbreviations: BIND, Biomolecular Interaction Network Database; CAI, codon adaptation index; COG, Clusters of Orthologous Group; DIP, Database of Interacting Proteins; GC, genomic context; GO, Gene Ontology; LCMS, liquid chromatography-tandem mass spectrometry; LPS, lipopolysaccharide; MALDI, matrix-assisted laser desorption/ionization mass spectrometry; MCL, Markov clustering algorithm; PI, physical interactions; STRING, Search Tool for the Retrieval of Interacting Genes/Proteins

* To whom correspondence should be addressed. E-mail: gmoreno@wlu.ca (GM-H); andrew.emili@utoronto.ca (AE)

☯ These authors contributed equally to this work.

¤ Current address: Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America

Author Summary

One goal of modern biology is to chart groups of proteins that act together to perform biological processes via direct and indirect interactions. Such groupings are sometimes called functional modules. The types of protein interactions within modules include physical interactions that generate protein complexes and biochemical associations that make up metabolic pathways. We have combined proteomic and bioinformatic tools, and used them to decipher a large number of protein interactions, complexes, and functional modules with high confidence. In addition, exploring the topology of the resulting interaction networks, we successfully predicted specific biological roles for a number of proteins with previously unknown functions, and identified some potential drug targets. Although our work is focused on *E. coli*, our phylogenetic projections suggest that a considerable fraction of our observations and predictions can be extrapolated to many other bacterial taxa. As all the data derived from this study are publicly available, others may build on our work for further hypothesis-driven studies of gene function discovery.

chemical pathways, and functional modules that in turn mediate all the major cellular processes. Elaboration of these interaction networks via proteomic, genomic, and bioinformatic approaches can reveal previously overlooked components and unanticipated functional associations [12]. For example, a recent integrative analysis of phenotypic, phylogenetic, and physical interaction data led to the discovery of an evolutionarily conserved set of novel bacterial motility-related proteins [13]. However, although systematic integration of diverse high-throughput interaction datasets is routinely performed to reveal new functional relationships in model eukaryotes such as yeast, worm, and fly [14–19], few analogous studies of the global functional architecture of *E. coli*, and any prokaryote for that matter, have been reported to date [20–22].

To this end, we have combined complementary, highly sensitive computational and experimental procedures to derive extensive high-quality maps of the functional interactions inferred by genomic context (GC) methods and physical interactions (PI) deduced by proteomics of *E. coli*. Our results indicate that many previously unannotated bacterial proteins are components of functionally cohesive modules and multiprotein complexes linked to well-known biological processes. A substantive fraction of these associations could be verified by independent experimentation and were found to be broadly conserved across prokaryotic phyla, indicating homologous systems in other microbes, whereas others are seemingly restricted to the *E. coli* lineage. The entire data collection is publicly accessible via a searchable Web-browser interface to stimulate exploration of both conserved and specialized bacterial proteins within the context of biological processes of particular interest.

Results

The Extent of Existing Functional Annotation for *E. coli* Proteins

Since the functional characterization of *E. coli*, and bacteria in general, has largely been guided historically by scientific interests and technical considerations, some bias is expected in terms of the coverage and depth of existing biological knowledge as reflected in current gene annotations. To evaluate the degree to which the physiological functions of the 4,225 putative protein-coding sequences of *E. coli* K-12 are characterized presently, we examined the scope of literature reference records curated in the UniProt annotation system [23]. After excluding PubMed references corresponding to genomic mapping studies, the average total number of papers associated with each of the proteins of *E. coli* K-12 is surprisingly limited (Figure 1A), with many proteins apparently still uncited.

We next examined recent *E. coli* K-12 (substrains W3110 and MG1655) gene annotations in the public databases RefSeq [24], MultiFun [25], and EcoCyc [11]. Since W3110 is commonly used for high-throughput studies, we devoted the bulk of our subsequent analysis to this substrain. However, to make sure that relevant functional attributes were not overlooked, we cross-mapped the corresponding gene accessions in both substrains and compiled an inclusive set of functional annotations accordingly (Table S1). In total, we found that 2,794 (66%) of *E. coli*'s proteins had either proper mnemonic names [26], experimentally derived annotations in the MultiFun multifunction schema, or literature documentation to a well-defined pathway or multiprotein complex in EcoCyc (Figure 1B). This left 1,431 proteins (34%) as currently functionally uncharacterized (which constitute our orphans set, listed in Table S1; see Protocol S1 for details). Of these, 446 (31%) have at least one putative molecular function defined on the basis of sequence (such as the presence of a predicted DNA-binding domain or an enzymatic motif) in the Clusters of Orthologous Groups (COGs) of proteins catalog [27].

Properties of the Functional Orphans of *E. coli*

The genes lacking annotation appear to be translated into bona fide proteins as their corresponding transcripts [28] were not significantly ($p = 0.36$) less stable than the products of annotated genes (Figure 1C). However, some differences were evident in terms of their biophysical attributes and evolutionary scope relative to annotated genes (Table S1). Most notably, only 21 orphans (1.5%) are required for viability under standard laboratory conditions [2] in contrast with the 280 annotated genes (10%) previously deemed essential. The orphans were also significantly ($p <$ 1e–10) less abundant at both the transcript (Figure 1D; average normalized mRNA expression over 400 microarray experiments [5]: 8.0 [orphans] vs. 8.9 [annotated]) and protein levels (Figure 1E; average codon adaptation index [CAI]: 0.41 vs. 0.47). Furthermore, they tend to encode somewhat smaller proteins (Figure 1F; average M_W: 29.4 vs. 38.2 kDa; $p <$ 1e–10) with fewer domain assignments (44%) than for annotated proteins (74%) according to the SUPERFAMILY database [29] (Table S1).

Orphans also generally find fewer orthologs in a non-redundant genome dataset, filtered at 90% similarity based on the frequency of shared orthologs (Figure 1G), with an average of 0.22 as compared with 0.48 for annotated genes ($p <$ 1e–10) using a maximum-score E-value cutoff of 1×10^{-6} for BLAST bidirectional best hits (BDBHs; Table S2 and Protocol S5 for details). Nevertheless, broader sequence comparisons against currently available metagenomes (Figure 1H) indicated that orphan homologs (one-way BLAST hits) are often widely distributed in diverse environments (Table

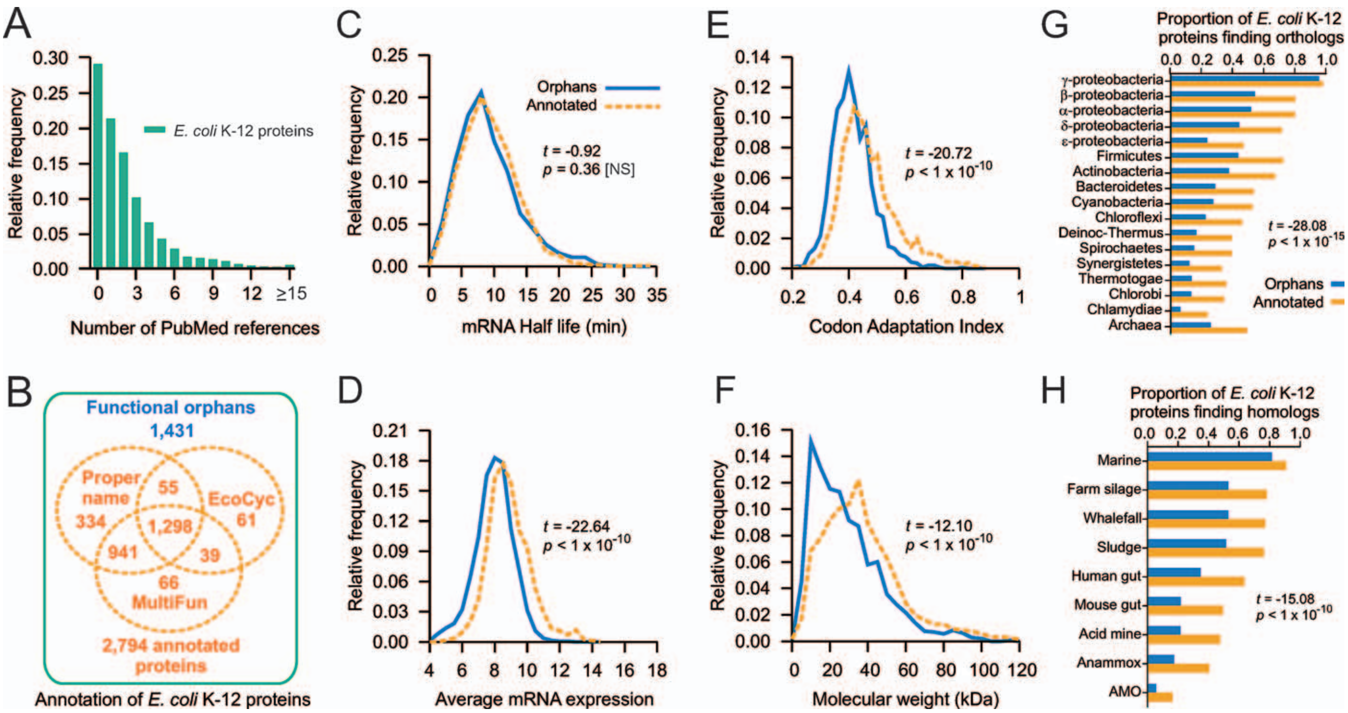


Figure 1. Annotated and Functional Orphan Genes of the *E. coli* K-12 Reference Strain

(A) Frequency distribution of supporting publications per *E. coli* protein-coding gene.

(B) Summary of existing annotations for *E. coli*, showing proteins of unknown function (orphans) lacking proper names and functional annotations in MultiFun [25] or EcoCyc [11].

(C–F) Although the functional orphans are encoded by transcripts with half-lives comparable to those of annotated genes, they tend to be expressed at lower levels based on (D) microarray analysis of mRNA and (E) CAI scores, and (F) have lower molecular weights on average. The *x*-axis in (D) represents the average of the log-scale mRNA expression level of each gene for all the arrays using the Robust Multi-Array Average normalized data obtained from the M3D database [103]. NS, not statistically significant; *p*, *p*-value; *t*, *t*-test.

(G) Orthologs of orphans are also less prevalent in sequenced genomes than those of annotated genes.

(H) However, examination of environmental metagenomic libraries indicates that the orphans are not necessarily exclusive to the *Escherichia* lineage. *Anammox*, anaerobic ammonium oxidation bacteria; *AMO*, methane oxidizing Archaea.

doi:10.1371/journal.pbio.1000096.g001

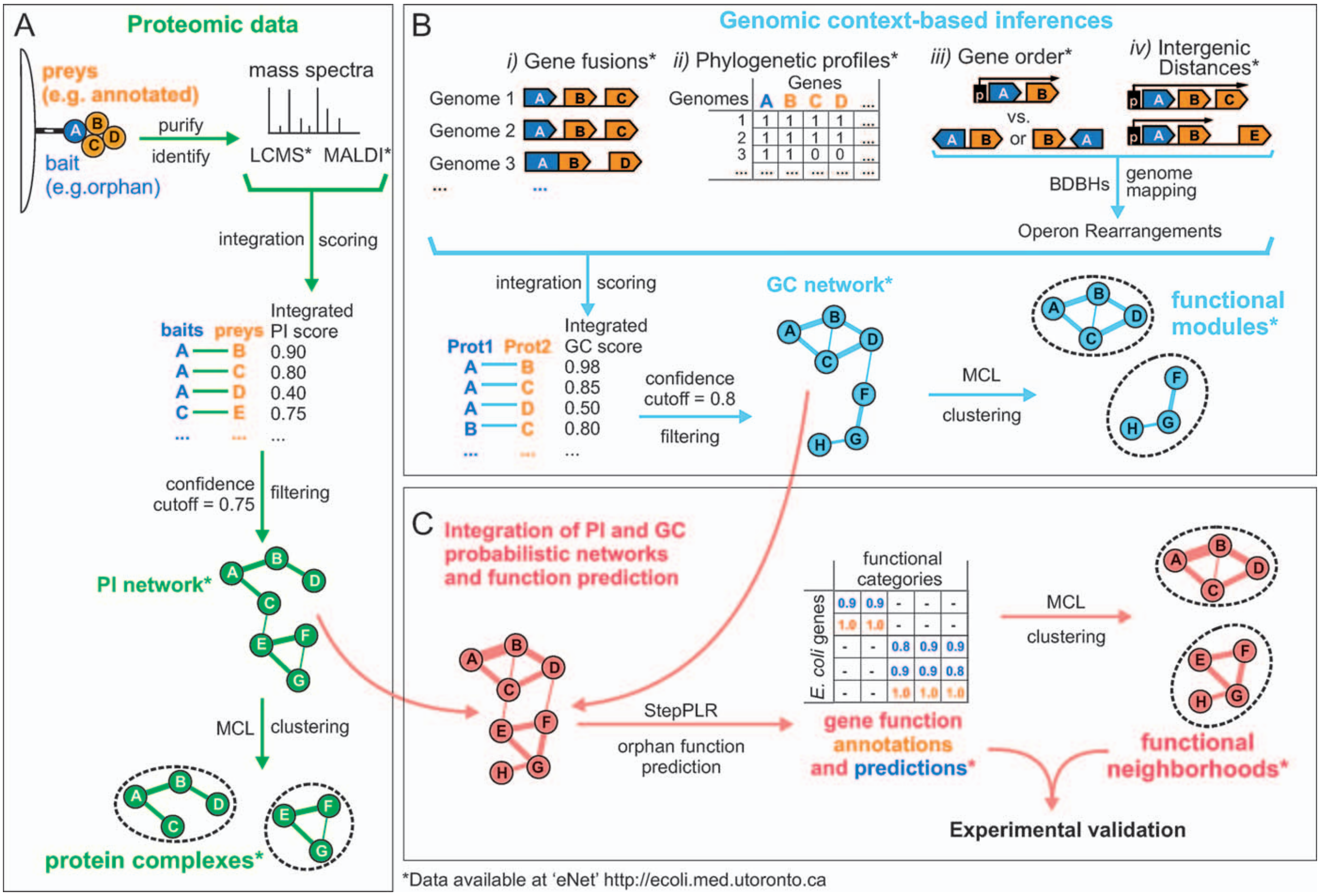
S2 and Protocol S2); for example, a high proportion (0.80) of orphans have homologs present in marine metagenomes, anaerobic bacterial populations (farm silage, 0.51; whalefall, 0.50; sludges, 0.49), and even in the residents of the mammalian gut (union of human and mouse, 0.35), implying participation in core bacterial processes. Furthermore, the same high proportion (~99%) of orphan and annotated genes have orthologs in the other sequenced *E. coli* isolates, including pathogenic variants and closely related *Shigella* strains (Table S2). Taken together, this argues that the functional significance of the orphans is more pervasive than the current annotations suggest.

A Systematic Approach to Elucidate Biological Function

The scarcity of the existing knowledge regarding the biological roles of the orphans is likely due to multiple reasons, ranging from the lower expression, nonessentiality, or smaller sizes of certain orphan proteins to their lack of obvious homologs in other organisms including humans. Accordingly, integration of multiple data sources is warranted to decipher the specific biological roles of this uncharacterized repertory. Since the elucidation of physical and functional interaction networks can provide insights into bacterial protein function based on the concept of guilt by association [30], we took a multipronged approach. We performed large-scale proteomic analysis to determine orphan participation as components of multimeric protein complexes, and inferred functional relationships based on genomic context inference, which exploits the patterns of gene conservation across bacterial genomes [31]. We then predicted the functions of the orphans using an integrative machine-learning procedure with extensive benchmarking. Finally, we performed independent experiments to validate a subset of high-confidence predictions related to core biological processes. Key steps in our pipeline are outlined schematically in Figure 2.

Experimental Definition of the Physical Interaction Network of the Soluble Proteome

We performed systematic large-scale tandem-affinity purifications of all endogenous soluble orphan and annotated proteins detectably expressed in *E. coli* W3110 under standard culture conditions (see Materials and Methods and Protocol S3 for details). We used an optimized Sequential Peptide Affinity (SPA)-tagging system to isolate multiprotein complexes [32]. This procedure is based on the integration of a marker cassette bearing a dual-affinity tag, consisting of three FLAG sequences and a calmodulin-binding peptide separated by a protease cleavage site, fused to the C-termini of targeted open reading frames in *E. coli* DY330 (W3110 background) via λ-phage "Red"–mediated homologous recombination. This system enables recovery of native bacterial protein complexes



*Data available at 'eNet' http://ecoli.med.utoronto.ca

Figure 2. Generation and Integration of Physical and Functional Networks and Orphan Function Prediction

(A) Construction of a PI network based on protein copurification and detection by mass spectrometry. The confidence scoring of the LCMS and MALDI networks was conducted using a logistic regression with datasets consisting of PI from low-throughput studies curated in DIP, BIND, and IntAct (gold positives) and proteins in different subcellular localizations (gold negatives). The two networks were integrated using a probabilistic model [61] (Protocol S6). The resulting PI network, with edge weights corresponding to likelihood ratios, was clustered using MCL to delimit "multiprotein complexes."

(B) Integration of four GC methods into a single functional interaction network using the same probabilistic model [61] and resulting scores (edge weights) were input to MCL to delimit "functional modules."

(C) Orphan function prediction was conducted using a "guilt-by-association" procedure. After integration of PI and GC interactions into a single probabilistic network [61], a machine learning algorithm (StepPLR) newly developed for this study was used to assign functions based on the binary associations of orphans with annotated proteins, the respective interaction edge weights, and the overall network topology. Correlations between vectors of these function predictions (orphans), and the annotations were then used as input to delimit "functional neighborhoods" by clustering using MCL.

doi:10.1371/journal.pbio.1000096.g002

at near-endogenous levels [4], minimizing spurious nonspecific protein associations. Stably interacting polypeptides were subsequently detected using a highly sensitive combination of tandem mass spectrometry (LCMS) and peptide mass fingerprinting procedures (MALDI) to increase detection coverage and accuracy (Protocol S3), just as we had previously done in a focused investigation of highly conserved essential *E. coli* proteins [4]. We successfully chromosomally tagged 1,241 new baits, aiming to verify putative interactions by reciprocal tagging where possible, for a total of 1,476 large-scale protein purifications (after including the 235 reported previously), of which 552 represented orphans (Protocol S3).

Since proteomic datasets typically contain noise in the form of nonspecific associations, we performed a careful statistical analysis and quality filtering to determine biologically meaningful PI. We considered that the specificity and affinity between any two putatively interacting proteins should be correlated with the consistency of copurification over all the experiments in which the proteins were identified (i.e., co-complexed). We therefore used an established copurification metric [33] to assess interaction specificity based on the similarity of the protein copurification patterns (Protocol S3). We then generated a single consolidated confidence score for each putative pairwise physical interaction based on the copurification metric together with the primary interaction evidence to penalize inconsistent or promiscuous binders (i.e., possible false positives) using alternatively a logistic regression model and Bayesian inference [34] (Protocol S3).

The logistic regression model was trained using a reference set of curated gold-standard PI (Protocol S3), which represents the union of experimentally verified PIs derived from low-throughput experiments extracted from the Database of Interacting Proteins (DIP) [35], the Biomolecular Interaction Network Database (BIND) [36], and the IntAct database [37]. For the negative gold standards, we compiled pairs of proteins annotated with different subcellular localizations (i.e., one cytoplasmic, the other periplasmic or outer membrane-bound [38].

Despite its relative simplicity, the logistic regression model offered better performance than the Bayesian method (see Figure 3A and Table S3). We therefore applied the former to our global PI network, assigning a probabilistic confidence score for each pair of putatively interacting proteins (Protocol S3). To minimize false positives without incurring excessive false negatives, we further filtered our network using a stringent minimum confidence cutoff of ≥ 0.75 as a high proportion (71%) of PI verified by reciprocal purification (Table S4) had likelihood scores at or above this threshold (Figure S1A). Finally, we removed the ten most-highly connected "hub" proteins that were deemed particularly abundant nonspecific contaminants (Table S5).

The resulting final network consisted of 5,993 high-confidence, nonredundant pairwise interactions among 1,757 distinct *E. coli* proteins, including 451 orphans, or roughly two-thirds of the predicted soluble cytoplasmic proteome. As summarized in Figure 3B, most (3,193, or 53%) of these PI are novel (Table S6), whereas only 47% were already reported in either the DIP, BIND, or IntAct interaction databases, or previous large-scale proteomic studies [1,4]. Importantly, our filtered dataset had a comparable level of accuracy (median confidence of 0.79) as for the much smaller set of 716 "validated" PI previously reported by our group [4] and a genome-scale dataset of 7,123 PI (median confidence of 0.69) generated using an analogous affinity purification schema in yeast [39].

The reliability of our dataset was also evident by two additional independent criteria. First, the mRNA expression patterns of the putatively interacting proteins were nearly as highly correlated as those of PI determined by low-throughput experiments (Figure 3C), even when these last are formed by presumably more abundant proteins (Figures S1E and S1F). Second, despite the more limited evolutionary distribution of the orphans, the putatively interacting proteins exhibited an elevated degree of co-occurrence of the respective orthologs across other bacterial species, evident in the high mutual information of the corresponding phylogenetic profiles (Protocol S5), again comparable to that of interacting pairs derived from low-throughput experiments (Figure 3D).

Collectively, these results indicate that our physical interaction network is very likely to be informative about orphan protein function.

Orphan Membership within Multiprotein Complexes

Since macromolecular assemblies mediate biological function in cells, we partitioned our high-confidence physical interaction network using the Markov clustering algorithm (MCL; see Materials and Methods and Protocol S4) to define orphan membership as subunits of discrete multiprotein complexes. MCL simulates random walks (i.e., flux) to delimit highly connected subnetworks based on both the connectivity and the weight of the graph edges [40]. In this case, the weights reflect the integrated PI scores obtained by logistic regression (Figure 2A). The higher the flux within in a region of the PI network, the more likely MCL will delimit the region as a cluster (in this case, a putative multimeric protein complex). A recent comparative study [41] found that MCL is often superior to other clustering algorithms in identifying functionally related groupings in probabilistic molecular interaction graphs and is remarkably resilient to spurious graph perturbations (e.g., missing edges).

We optimized the MCL parameters (see Materials and Methods and Protocol S4) to partition the 5,993 PI network, generating a set of 443 putative multiprotein complexes (Figure 3E), most of which consist of two to four polypeptides (Table S7). In agreement with previous reports [41], alternative clustering algorithms comparable to MCL in terms of accuracy, such as the Restricted Neighborhood Search Cluster algorithm [42], produced similar groupings (unpublished data). Moreover, as was found in a proteomic survey of yeast multiprotein complexes [39], both the subunit number and degree of connectivity of the MCL clusters followed a power-law distribution (Figure S2B and S2C). More telling, 244 (55%) of these *E. coli* multiprotein complexes contained at least one orphan as a putative subunit, with linkages suggestive of concerted biological functions (Figure 3E). The complexes also showed a significant ($p < 0.001$) enrichment in terms of functional homogeneity, compared with null random models (Figure S2A), implying that both the annotated components and the associated orphans tend to participate in the same biological processes.

For example, 25 orphans were detected as part of a large subnetwork of putative complexes involved in protein synthesis (Figure 3F). These include the orphans YbcJ and YncE, which physically interacted with the pseudouridylate synthase RluB, the RNA helicases SrmB and DeaD, the exoribonucleases E (Rne) and R (Rnr), and other components of the ribonucleolytic "degradosome" responsible for mRNA degradation, suggesting a probable role in RNA processing and/or turnover. Likewise, YfgB copurified with three translation-related complexes, including ribosomal proteins. Consistent with these observations, the expression of YncE, which has similarity to the nonribosomal peptide synthase AfuA of *Aspergillus fumigatus*, is reduced more than 9-fold upon exposure of *E. coli* to the translational inhibitor puromycin [43]. We also determined that deletion of *ybcJ* results in a significant reduction in the incorporation of ^{35}S-labeled methionine in vivo relative to wild type (Figure 3G), indicating a decrease in the global rate of protein synthesis. Similarly, ribosome profile analysis (Figure 3H) showed that inactivation of *yfgB* decreased the level of mature polysomes actively engaged in mRNA translation and altered the cellular ratios of 30S and 50S ribosomal subunits relative to 70S monosomes. Moreover, both the *ybcJ* and *yfgB* mutants exhibited reduced translation fidelity (Figure 3I) as assayed by four reporter plasmids that measure the frequency of frameshifts and stop codon readthrough.

Other orphans in this translation subnetwork include YibL, which copurified both with YfgB and YbcJ, and with RNA processing factors involved in ribosome biogenesis, such as the RNA pseudouridine synthetases RluB/RluC and the RNA helicase DeaD, and with RppH (formerly NudH), which was recently identified as a regulator of 5′-end–dependent mRNA degradation [44–46]. Similarly, the orphan YdhQ copurified

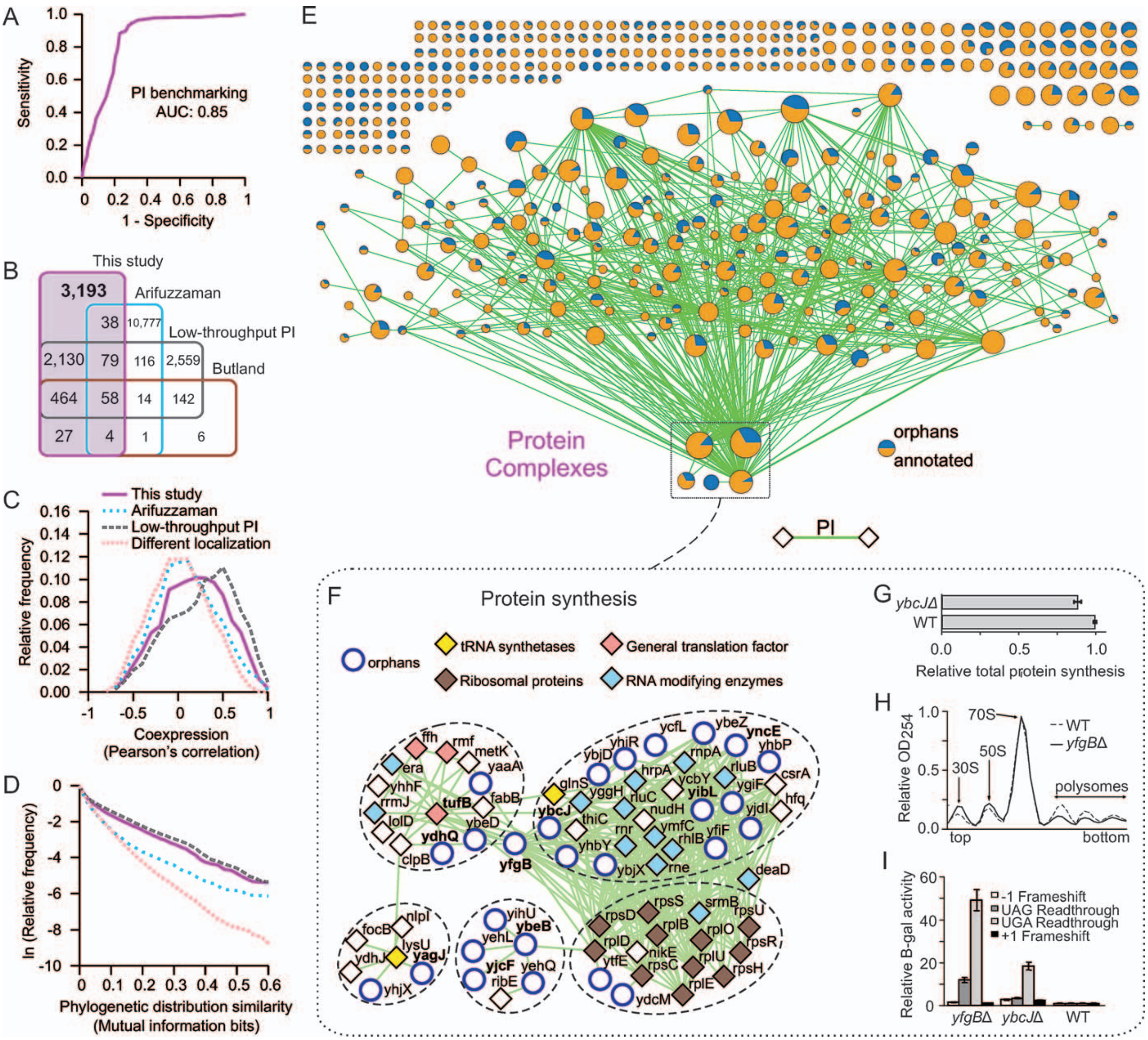


Figure 3. High-Confidence Physical Interactions and Putative Multiprotein Complexes

(A) Benchmarking of the experimentally derived PI network in *E. coli* against positive and negative gold standards by receiver operating characteristic (ROC)-curve analysis; cumulative area-under-the-curve (AUC) is shown as an overall performance measure.

(B) Overlap of PI identified in this study with previous proteomic reports [1,4] and low-throughput PI obtained from DIP, BIND, and IntAct.

(C and D) Putatively interacting proteins have highly correlated gene expression patterns (C) and similar phylogenetic profiles (D) based on mutual information as for low-throughput curated PI and in contrast to control protein pairs derived from different subcellular compartments.

(E) Graphical schematic of putative stable, soluble multiprotein complexes, drawn using the GenePRO Cytoscape plugin [104] (see Table S7 for listing). Each node represents a complex, whose size reflects the number of contained proteins; edge widths reflect the number of interactions between subunits of different complexes.

(F) Multiprotein complexes implicated in the bacterial translation apparatus; orphan and annotated genes mentioned in the main text are highlighted in bold.

(G) Reduced rate of total protein synthesis in a strain lacking *ybcJ* relative to wild-type cells (WT).

(H) Perturbed ribosome profiles in an *yfgB* deletion strain.

(I) Elevated rates of frame-shifting and stop-codon readthrough in *yfgB* and *ybcJ* deletion strains relative to wild-type (WT). β-gal activity is only produced after the corresponding translational defect has occurred; error bars indicate standard deviation.

doi:10.1371/journal.pbio.1000096.g003

with translation elongation factor Tu, whereas YagJ interacted with lysine tRNA synthetase (LysU); and YjcF, which has similarity to phenylalanyl-tRNA synthetase PheT of *Bacteroides vulgatus*, bound ribosomal release factor 2 and another orphan, YbeB, which in turn was found to associate with the 50S ribosome subunit, as recently reported [47]. These results confirm that our high-confidence physical interaction network is informative about the function of at least certain orphans.

Functional Interactions Predicted by Genomic Context Methods

Although we attempted to tag and purify the entire soluble *E. coli* interactome, we failed to detect 469 orphan proteins by MALDI or LCMS, presumably because they are membrane-associated (~35%; Figure S1B) and hence not soluble, or are of particularly low abundance (~40%), as reflected by their CAI and mRNA levels (Figures S1C and S1D, respectively). To bypass this limitation, we applied computational methods to discern a network of high-confidence pairwise functional interactions for all *E. coli* proteins, including those not detectable by proteomic methods, by examining the natural chromosomal clustering of bacterial genes. As illustrated in Figure 2B, we used four different GC methods, namely: (1) gene fusions [48,49]; (2) similarity between phylogenetic profiles [27,50,51]; (3) evolutionary conservation of gene order [52–54]; and (4) intergenic distances [55–57] (see Materials and Methods and Protocol S5 for details). The latter two methods are independent approaches to detect operons and their subsequent rearrangements across prokaryotic genomes. In particular, the intergenic distances method, leads to considerably more high-quality predicted functional associations compared with the first three classic GC methods [55], and does not depend critically on the detection of orthologs in evolutionarily distant genomes, making it potentially better suited for detecting functional interactions involving orphans.

The pairwise interactions generated by each of these prediction methods were independently evaluated by benchmarking using suitable gold standards. Positive gold standards were defined as pairs of *E. coli* genes belonging to the same biological pathway as defined in EcoCyc, while the negative gold standards represented pairs of annotated *E. coli* genes whose products participate in different pathways (see Protocol S5 for details). The results of each GC method were subsequently combined to create a single unified functional association score (Protocol S6 and Figure 2B). Although different data integration algorithms have been developed [58–61], most of these have a similar probabilistic basis and assumptions. For this study, we opted for the integration procedure used by von Mering and colleagues [61] to construct the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database. This approach treated the reliability of the associations generated by each GC method as independent probabilities, such that the likelihood of an interaction is proportional to the number of times it was observed and the degree to which each GC method contributed to the overall network reliability (Protocol S6). Finally, we applied a stringent filter to the unified functional network to obtain a set of 74,776 high-confidence (probabilities ≥0.80) nonredundant interactions (Figure 4A and Table S8).

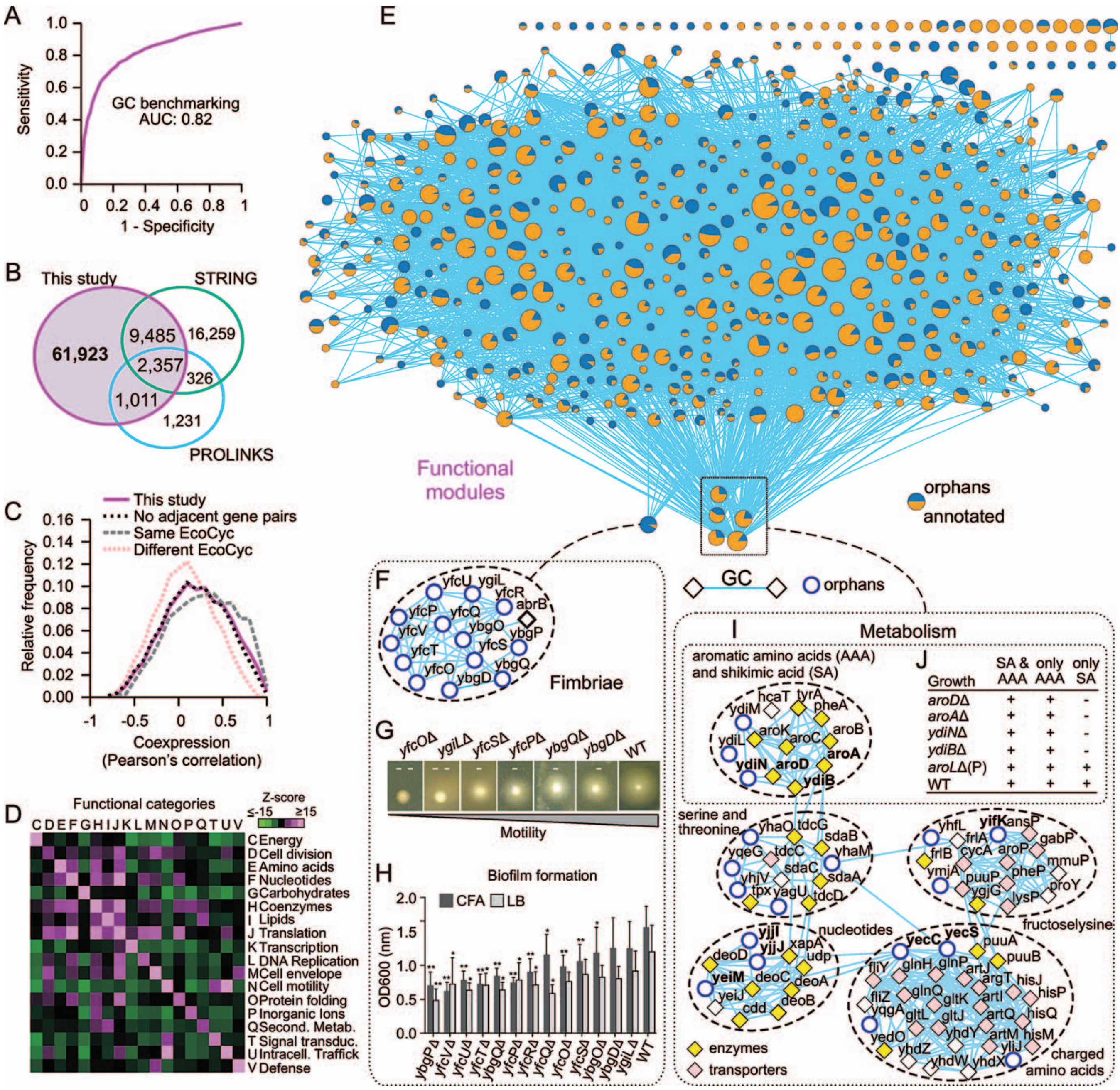
Despite the tendency of the orphans to exhibit more limited conservation notwithstanding the dependency of GC methods on homologs in multiple species (except for operon predictions based on intergenic distances [55]), our combined GC network implicated virtually all (1,367, or 96%) of the orphans in 23,365 pairwise functional interactions (Table S8). Moreover, relatively few (<18%) of our predicted interactions appear to have been reported previously (Figure 4B). Although we could not meaningfully compare our results to an alternate set of putative functional links generated recently [22] because of a lack of publicly accessible dataset scores, we found that less than 5% (3,368) of our predicted interactions are listed in the PROLINKS comparative genomics databank [62], whereas only approximately 16% (11,842, of which only 2,613 involve an orphan) were found in STRING (v. 7.1) at a more liberal 0.7 confidence threshold. More critically, greater than 85% of our predicted orphan interactions involve a functionally annotated *E. coli* protein, indicating a good potential to make functional inferences. The fact that PROLINKS has 1,657 predictions not attained by our integrative approach may reflect our use of a higher confidence threshold as well as differences in implementation of the GC measures and the identification of putative orthologs. For instance, whereas we used BLAST-BDBHs as criteria to detect orthologs between pairs of genomes, STRING uses COG-based definitions of orthology, whereas PROLINKS uses one-way BLAST hits (not necessarily orthologs). Conversely, most of the 16,585 predictions exclusive to the STRING database were compiled using text mining or alternate experimental criteria such as protein–protein interactions, whereas the highest numbers of predictions exclusive to our GC datasets come from operon rearrangements.

The reliability of our unified functional association network was independently corroborated based on the high correlations of expression among putatively interacting gene pairs (Figure 4C), which was comparable to that observed for components of the same curated EcoCyc pathway even after eliminating all pairs of genes belonging to an experimentally characterized operon or all adjacent gene pairs in *E. coli* (Figure 4C). We also observed a marked enrichment for interactions among proteins annotated to the same curated functional categories (Figure 4D), implicating by extension any associated orphans in these same processes.

Defining the Participation of Orphans as the Components of Functional Modules

Groups of functionally interacting genes form functional modules centered on a common process or biochemical pathway(s). To define orphan participation as components of such modules, we partitioned the high-confidence GC network using MCL (Protocol S4), generating a total of 507 putative functional modules consisting of two or more components (Figure 4E and Table S9). Examination of the functional homogeneity of these predicted modules (see Materials and Methods and Protocol S4) indicated, as for our putative multiprotein complexes, that they were highly enriched (p <0.0001) for concerted annotated biological processes (Figure S2D), again implicating the associated orphans in these same roles. Module membership followed a characteristic power law distribution (Figure S2E) with most modules having between two and 10 components, but the overall node connectivity did not (Figure S2F); further analysis is necessary to determine the significance of this divergent behavior.

Two hundred and eighty-nine (57%) of the modules had at least one of a total of 1,189 different orphans. One notable example is shown in Figure 4F. Diverse lines of experimental and bioinformatic evidence support the involvement of this putative module in the biogenesis and/or activity of fimbriae, appendages or pili that are shorter than the characteristic



J

Growth	SA & AAA	only AAA	only SA
*aroD*Δ	+	+	-
*aroA*Δ	+	+	-
*ydiN*Δ	+	+	-
*ydiB*Δ	+	+	-
*aroL*Δ(P)	+	+	+
WT	+	+	+

Figure 4. High-Confidence Genomic Context Associations and Putative Functional Modules

(A) Benchmarking of unified GC interactions in *E. coli* against positive and negative gold standards by ROC-curve analysis.

(B) Overlap of high-confidence functional interactions predicted in this study with two other public GC databases.

(C) Even after eliminating adjacent gene pairs to control for known and predicted *E. coli* operons, functionally linked genes have highly correlated patterns of mRNA expression comparable to components of the same curated EcoCyc pathways rather than different pathways.

(D) Functionally linked genes are enriched for annotations to the same COG functional categories.

(E) Graphical representation of putative *E. coli* functional modules (see Table S9 for listing); node size and colors are proportional to the number and fraction of orphan and annotated subunits, respectively, while lines represent interactions connecting modules.

(F) Putative fimbriae-related module.

(G) Defective motility of mutant strains deleted for orphans linked to fimbriae (from [F]); single dashes (-) indicate moderately impaired motility, while double dashes (–) represent strong repression. Other mutants displaying a normal phenotype comparable to the wild-type strain BW25113 (WT) are not shown.

(H) Defective biofilm formation by mutants deleted for fimbriae-related orphans (from [F]); significant differences (*t*-test) in cell adhesion (absorbance) between mutant and WT strains are denoted by asterisks (single asterisks [*], $p < 0.01$; double asterisks [**], $p < 0.0001$). Error bars indicate standard deviation of the mean. CFA, colonization factor antigen medium; LB, Luria Bertani medium.

(I) Metabolic modules mentioned in main text.

(J) Mutants auxotrophic for aromatic amino acids show defective growth on minimal media containing shikimic acid. A prototroph—*aroL*Δ mutant strain (P)—is shown for comparison.

doi:10.1371/journal.pbio.1000096.g004

flagellum of Gram-negative bacteria, which mediate cell adhesion, biofilm formation, motility, and host invasion [63,64]. For instance, 12 of the 13 orphan components possess sequence characteristics of bacterial adhesins and chaperone/Usher pili protein families [29,65]. Gene expression profiling studies [66,67] have previously established that most of these orphans are also coordinately induced during biofilm formation (Table S10). Perhaps most compellingly, we found that single-gene *E. coli* knockout mutants of six of the 13 orphans display markedly reduced swarming capabilities in semisolid agar (Figure 4G), while 11 out of 13 mutants were significantly impaired for biofilm formation in vitro as compared with a wild-type control (Figure 4H). Taken together, these observations strongly implicate this set of orphans in the formation and/or proper function of fimbriae.

Several other prominent modules are shown in Figure 4I. These comprise the orphans YdiN, YdiL, and YdiM predicted (based on operon rearrangements) to functionally interact with several members of the Aro operon known to participate in the metabolism of shikimate, a precursor of aromatic amino acids. Consistent with this, *ydiN*, *aroD*, and *ydiB* are reportedly overexpressed when *E. coli* is grown in media containing shikimate as the sole carbon source [68]. Moreover, we found that deletion of either *ydiN* or *ydiB* resulted in defective metabolism of shikimate causing phenotypic auxotrophy for aromatic amino acids (Figures 4J and S3) as is observed for mutants of known aromatic amino acid biosynthetic genes (e.g., *aroA* and *aroD*) [69].

Other functional modules include *frlA*/*frlB*, part of the Frl operon of *E. coli* responsible for the import and metabolism of the alternative carbon source fructoselysine, together with the orphan YifK, which has sequence characteristics of a transporter [38], implicating it in electrochemical potential-driven uptake of this sugar. Conversely, two orphans, YecC and YecS, had functional associations consistent with linkages to amino acid and nucleotide metabolism, four (YagU, YqeG, YhaO, and YhaM) were linked to a putative module involved in transport and metabolism of threonine and serine, whereas three others (YjjI, YeiM, and YjjJ) were found in a module enriched for factors involved in nucleotide transport and degradation of deoxyribonucleosides.

Taken as a whole, these results suggest discrete functional relationships for many previously unannotated proteins, implicating certain orphans within specific pathways.

Improved Functional Inference within an Integrated Network Framework

Examination of the extent of overlap between our physical and functional networks, both in terms of common binary interactions and shared components among the derived complexes (from PI) and modules (from GC), indicated that they are largely complementary (Table S11). Since a similar trend was also evident comparing other existing curated *E. coli* PI datasets (derived from either low-throughput or other high-throughput studies) with independent GC inferences (e.g., from STRING; Table S11), this presumably stems in part from the incomplete coverage obtained by these different approaches. Regardless, these observations imply that the union of PI and GC networks is necessary to capture the widest spectrum of biologically relevant interactions. Indeed, it has been shown previously that the combination of PI with functional genomic inferences, each statistically weighted according to dataset quality, can markedly improve both functional coverage and accuracy [59,61,70–72]. We therefore merged our experimental and predicted associations with the same method used to generate the unified GC network (Figure 2C; see Materials and Methods and Protocol S6 for details).

The resulting combined probabilistic network consisted of 80,370 high-confidence (probability ≥75%) putative pairwise interactions encompassing virtually the entire proteome of *E. coli*, including 2,769 (99%) annotated proteins and 1,375 (96%) functional orphans (Table S12). Graph analysis of this final integrated network (Protocol S7 and Table S13) indicated that the orphans tended to have a lower overall connectivity and betweenness centrality, measured as the number of shortest paths going through a given node, relative to annotated components, suggesting more peripheral positions in the integrated networks. However, the orphans also exhibited lower average closeness, defined as the average length of shortest paths between any two nodes, and had similar overall clustering coefficients, indicating that, in general, the orphans are functionally connected to, rather than isolated from, the annotated gene products. These observations implied that consideration of both the individual associations and overall placement of the orphans within the integrated interaction network would facilitate functional deduction.

We therefore devised a new network-based function prediction method (termed StepPLR; see Figure 2C and Materials and Methods) to exploit the global topological similarity among all the protein pairs and their corresponding functional annotations in the integrated network. Our method assigns functions to orphans based on the functional information from their first-order (direct) and second-order (indirect) annotated neighbors in the integrated network using penalized logistic regression models and a stepwise variable selection procedure to deduce optimal functional profiles (see Protocols S8 and S9 for details). We based our classifications on the discrete COG functional categories and on the hierarchical, multifunctional terms of the Gene Ontology (GO) [73,74] and MultiFun classification schemas [25]. To avoid potential sources of false predictions, we removed any proteins labeled with the evidence codes IPI (for inferred from protein interaction) and IGC (for inferred from genomic context method) when generating the GO reference set, as well as proteins in poorly characterized categories in COGs and MultiFun (Protocol S9 and Table S14).

As shown in Figure S4, StepPLR had better precision and recall prediction performance than several other widely used guilt-by-association procedures tested, such as majority-counting and chi-square–based methods (see Protocol S10 for details). Although the performance achieved for the different functional categories varied, our approach generated area-under-the-curve (AUC) values of 0.8 or higher for most of the COG (83%), GO (67%), and MultiFun (53%) categories (Table S15), and was relatively insensitive to the number of annotated proteins per function. Moreover, since our method exploited the correlation among the different categories, most orphans had multiple biologically consistent predicted functions (Table S16).

Functional Neighborhoods

As displayed graphically in Figure 5A, our prediction procedure ultimately linked many of the orphans to specific,

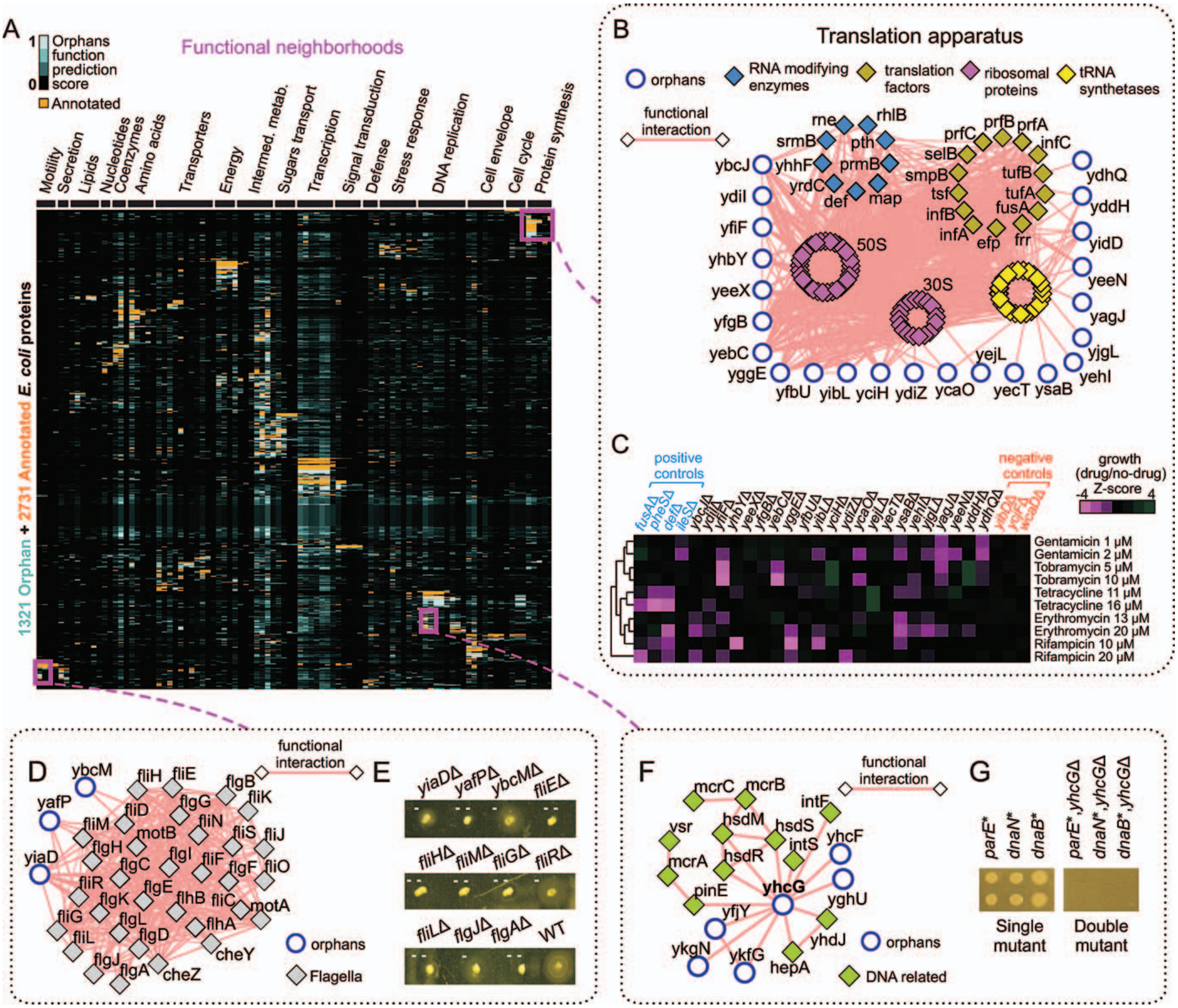


Figure 5. The Functional Neighborhoods of *E. coli*

(A) A "clustergram" displaying existing annotations (orange) and the orphan predicted functions (this study; blue) for all the protein-coding genes of *E. coli* (*y*-axis) and their associated biological processes (*x*-axis) (complete terms are provided in Figure S5 and Table S16). Proteins were clustered using MCL based on the paired similarity of the functional annotations and predictions in this matrix to delimit "functional neighborhoods" (see Table S17 for listing).

(B) Putative functional neighborhood showing high-confidence integrated functional interactions (combined PI and GC networks) of select orphans with the protein synthesis machinery. For clarity, individual names of ribosomal proteins and tRNA synthetases are not shown.

(C) Heat map showing the differential sensitivity of orphan deletion strains to antibiotics targeting protein synthesis relative to the colony size in the absence of drug (see Protocol S11 for details). Mutants deleted for annotated proteins from this neighborhood are shown as positive controls, whereas deletion mutants lacking genes not contained within this neighborhood are shown as negative controls.

(D) Neighborhood with three orphans putatively involved in flagellum assembly and motility.

(E) Deletions of the corresponding components reduce swarming capability; single dash (-), moderately impaired motility; double dash (--), strong repression.

(F) Subnetwork of orphans associated with DNA enzymes.

(G) Deletion of the orphan *yhcG* results in synthetic lethality when combined with hypomorphic alleles (as indicated by an asterisk [*]) of three essential DNA replication factors (*parE*, *dnaN*, and *dnaB*).

doi:10.1371/journal.pbio.1000096.g005

functionally related protein "neighborhoods". We again made use of the MCL algorithm to objectively delimit functionally highly homogeneous ($p < 0.0001$) protein groupings based on the profile similarity of annotations and predictions (see Protocol S4 for details and Table S17 for listing). One notable example is the protein translation machinery (Figure 5B), which has 23 associated orphans. To independently verify the functional relevance of these assignments, we examined the effects of deleting the corresponding genes in terms of conferring sensitivity to drugs that inhibit protein synthesis. Consistent with expectation of a direct role in protein synthesis, and similar to loss of bona fide annotated translation factors and tRNA synthetases, the mutant strains exhibited statistically significant ($p < 0.05$) differential sensitivity as compared to wild type and unrelated gene mutants to a variety of antibiotics that selectively block protein translation (Figure 5C and Table S18).

We also examined an alternate group of orphans (YafP, YiaD, and YbcM) associated with the flagellar biogenesis and motility apparatus (Figure 5D). Single-gene knockout mutants of annotated components in this neighborhood exhibit decreased motility in semisolid agar as compared to wild-type *E. coli* strains [13]. Consistent with our functional predictions, we likewise found that deletion of *yafP* ablated cell motility in vitro (Figure 5E), similar to mutants lacking core flagellum motor encoding genes (e.g. *fliH* and *fliM*), whereas loss of *yiaD* and *ybcM* reduced swarming (i.e., decreased halo formation) to an extent comparable to perturbation of other established flagellar components (e.g., *flgJ* and *fliR*). Additionally, a previous study [75] using phenotypic complementation analysis had suggested that a *ybcM* ortholog in *Yersinia enterocolitica* is likely an AraC-type regulatory protein involved in controlling bacterial motility. Taken together, these results suggest that, akin to several other recently discovered novel motility components [13,76], these orphans are required for the proper assembly and/or subsequent locomotion of the *E. coli* flagella, a fundamental bacterial structure.

Many other orphans were predicted to have roles in other conserved biological systems, such as DNA replication. For example, as shown in Figure 5F, we identified the orphan YhcG in association with DNA processing enzymes, including the restriction complexes HsdMRS and McrABC, the integrases IntF and IntS, and the recombinase PinE. YhcG has sequence characteristics of the PD-(D/E)XK superfamily of nucleases involved in DNA recombination and repair [77]. Consistent with these observations, we found that deletion of *yhcG* results in a synthetic-lethal phenotype (Figure 5G) when combined with hypomorphic alleles of the replicative primosome (*dnaB*), DNA polymerase III (*dnaN*), and DNA topoisomerase IV (*parE*), consistent with a direct role in DNA replication or the resolution of critical intermediates.

Novel Components of Bacterial Cell Envelope Biogenesis Pathways

Our functional predictions were particularly revealing about bacterial cell envelope biology, with implications for infectious disease and antibiotic susceptibility. Like other free-living microbes, *E. coli* is encased in a membranous cell envelope composed of proteins, lipids, and carbohydrates that serves as the interface to its environment and mammalian host, yet over a third of the approximately 1,000 predicted membrane-associated and periplasmic proteins of *E. coli* are presently functionally unannotated [38]. Figure 6A shows a set of eight orphans linked to a functional neighborhood along with 29 annotated proteins with established roles in the biogenesis of peptidoglycan, a major structural component of the bacterial cell wall that is a target of many antimicrobials. Consistent with our functional predictions, *E. coli* cells deleted for these same orphans exhibited differential sensitivity to various antibiotics that inhibit peptidoglycan assembly (Figure 6B). Moreover, the observed phenotypes were also characteristic of enzymes acting in the initial cytoplasmic and inner membrane stages of peptidoglycan biosynthesis, like *murA* and *mtgA*, rather than later periplasmic steps, like *pbpC*. This suggests orphans involvement in an early step in cell wall formation.

Figure 6C shows a second functionally homogeneous neighborhood composed of 14 orphans and 91 annotated proteins associated with pathways linked to the biosynthesis of lipopolysaccharide (LPS) and other core cell envelope components. Consistent with these assignments, subcellular protein localization studies based on green-fluorescent reporter fusions [78] have previously established that at least five of these orphans (YfbH, YfbJ, YfbW, YbjT, and YjdD) are physically tethered to the *E. coli* inner membrane, whereas signal peptides potentially mediating export across the inner membrane have been predicted for YafL and YadE [79]. Further, the interactions of annotated and orphan components within this neighborhood suggest specific relationships consistent with particular biological pathways. For instance, YfbJ, YfbH, and YfbW interact with Ugd and four other annotated proteins that participate in maturation of the lipid A anchor of LPS. Previous work [80] had suggested that an unknown transporter(s) shuttles lipid A intermediates between the cytoplasm and periplasm. Coincidentally, YfbJ and YfbW are paralogs that belong to a superfamily of multidrug efflux transporters [29], in which the C-terminus of YfbJ is cytoplasmic and that of YfbW periplasmic [78]. Since knockouts of the corresponding genes were recently reported to impair trafficking of lipid A precursors [81], YfbJ and YfbW appear to be the relevant transporters. Likewise, YadE and three other orphans (YjaH, YcaR, and YfhL) are predicted to function in Rfa-, Lpx-, and Kds-based pathways participating in KDO_2-lipid A (i.e., core LPS) biosynthesis. Consistent with these predictions, we found that deletion of either *yfbJ* or *yadE* resulted in aggravating synthetic genetic interactions when combined with mutant alleles of genes in parallel pathways (Figures 6C and 6D).

Other orphans in this neighborhood (e.g., YiiD, YbjT, and YbcH) are predicted to interact either with the Rff and Rfb pathways involved in the generation of other important cell envelope constituents, such as enterobacterial common antigen and O-antigen. Likewise, YibD and YafL interact with Wca and Wz proteins that participate in biosynthesis of colanic acid (M-antigen). Moreover, although they are not formally classified within this route, we detected functional interactions of RfaD with LpxD, KdsA, and KdsC involved in the synthesis of ADP-L-glycero-D-*manno*-heptose, which is used by many enzymes for LPS production.

Similar to mutants deficient for enzymes annotated to participate in the cell envelope biogenesis, we found that deletions of the orphans in this neighborhood significantly ($p < 0.05$; see Protocol S11) perturbed cell viability upon exposure to antibiotics that block formation of the bacterial cell envelope (Figure 6E), substantiating our functional predictions.

Functional Interactions of Orphans Extend beyond Proteobacteria

To investigate the evolutionary significance of the putative functional associations detected in *E. coli*, we examined the presence of orthologs of each of the interacting orphan and annotated protein pairs among currently available prokaryotic genomes (see Materials and Methods and Protocol S12

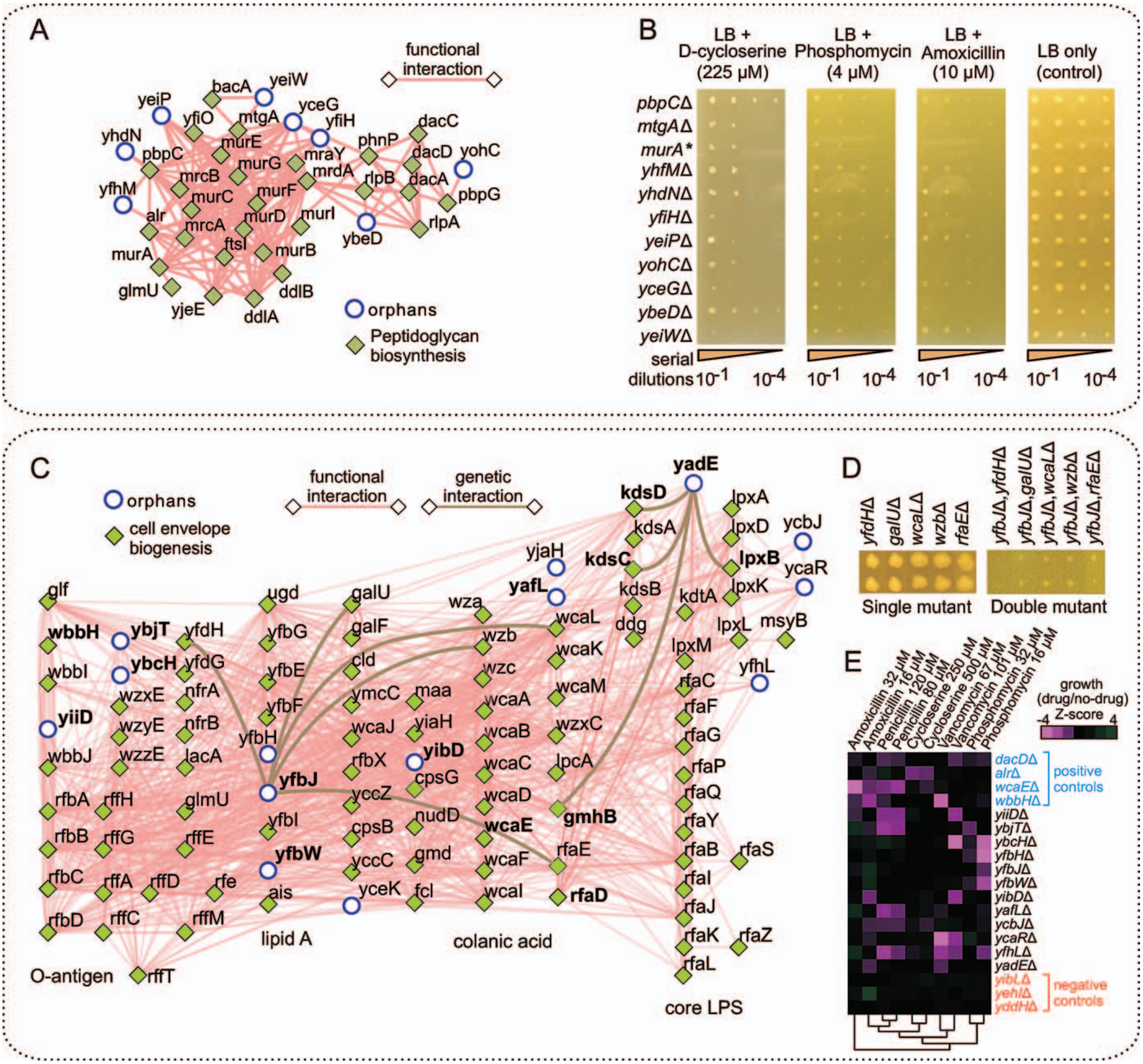

A
functional interaction
orphans
Peptidoglycan biosynthesis
yeiP yfiO bacA yeiW mtgA yceG yfiH yhdN pbpC murE murG mraY phnP dacC dacD yohC mrdA rlpB dacA mrcB murC murF yfhM pbpG alr murD murI mrcA rlpA ybeD ftsI murB murA glmU yjeE ddlB ddlA
B
LB + D-cycloserine (225 µM)
LB + Phosphomycin (4 µM)
LB + Amoxicillin (10 µM)
LB only (control)
pbpCΔ
mtgAΔ
murA*
yhfMΔ
yhdNΔ
yfiHΔ
yeiPΔ
yohCΔ
yceGΔ
ybeDΔ
yeiWΔ
serial dilutions
10-1 10-4
C
orphans
cell envelope biogenesis
functional interaction
genetic interaction
yadE kdsD lpxA kdsA lpxD ycbJ kdsC lpxB ycaR yjaH yafL kdsB kdtA lpxK ddg lpxL msyB lpxM yfhL
glf ugd galU wza wcaL wbbH ybjT yfdH yfbG galF wzb wcaK wbbI ybcH yfdG yfbE cld wzc wcaM wzxE nfrA ymcC maa wcaA wzxC yiiD wzyE nfrB yfbF wcaJ yiaH wcaB wbbJ wzzE lacA yfbH rfbX yibD wcaC lpcA rfbA rffH glmU yfbJ cpsG yccZ wcaD gmhB rfbB rffG rffE yfbI cpsB nudD wcaE rfbC rffA rffD rfe yfbW ais ycK gmd wcaF rfaE rfbD rffC rffM yceK fcl wcaI rfaD
rfaC rfaF rfaG rfaP rfaQ rfaY rfaB rfaS rfaI rfaJ rfaK rfaZ rfaL
O-antigen rffT lipid A colanic acid core LPS
D
yfdHΔ galUΔ wcaLΔ wzbΔ rfaEΔ
yfbJΔ,yfdHΔ yfbJΔ,galUΔ yfbJΔ,wcaLΔ yfbJΔ,wzbΔ yfbJΔ,rfaEΔ
Single mutant
Double mutant
E
Amoxicillin 32 µM
Amoxicillin 16 µM
Penicillin 120 µM
Penicillin 80 µM
Cycloserine 250 µM
Cycloserine 500 µM
Vancomycin 67 µM
Vancomycin 101 µM
Phosphomycin 32 µM
Phosphomycin 16 µM
growth (drug/no-drug) Z-score
-4 4
dacDΔ alrΔ wcaEΔ wbbHΔ
positive controls
yiiDΔ ybjTΔ ybcHΔ yfbHΔ yfbJΔ yfbWΔ yibDΔ yafLΔ ycbJΔ ycaRΔ yfhLΔ yadEΔ
yibLΔ yehIΔ yddHΔ
negative controls

Figure 6. Functional Neighborhoods Involved in Cell Envelope Biogenesis

(A) Functional neighborhood showing high-confidence integrated functional interactions (combined PI and GC networks) between components of the complete peptidoglycan biogenesis pathway and putatively functionally related orphans.

(B) Serial-dilution assay showing perturbed growth of *E. coli* strains deleted for putative cell wall–related components (from [A]) in the presence of antibiotic inhibitors of cell wall assembly; an asterisk (*) indicates a hypomorphic allele of the essential gene *murA*.

(C) Functional neighborhood of metabolic pathways involved in biogenesis of the bacterial cell envelope. Vertical groupings correspond to known pathways, whereas orphan components are tentatively positioned according to their interaction patterns; brown lines indicate aggravating genetic interactions recorded between the orphan *yfbJ* and *yadE* and annotated pathway components.

(D) Conjugation-based double-mutant growth assays (only test results for *yfbJ* are shown).

(E) Heat map showing the differential sensitivity of strains deleted for the orphans shown in (C) to antibiotics targeting the cell envelope. Mutants deleted of relevant annotated proteins are shown as positive controls, while mutants lacking genes not contained within this neighborhood are shown as negative controls.

doi:10.1371/journal.pbio.1000096.g006

for details). As might be expected from the more limited evolutionary scope of the orphans (cf. Figure 1G), functional interactions involving orphans were typically less broadly distributed than those of annotated proteins (Figure 7A), with the least-frequent distribution category consisting of interactions between the orphans themselves. Nevertheless, our analysis indicated extensive conservation of orphan associations across all sequenced forms of prokaryotic taxa. For instance, 5,553 putative interactions between orphan and annotated proteins, and 603 among orphan pairs alone, were predicted to be distributed as far as Archaea, again supporting the importance of orphans beyond that anticipated from

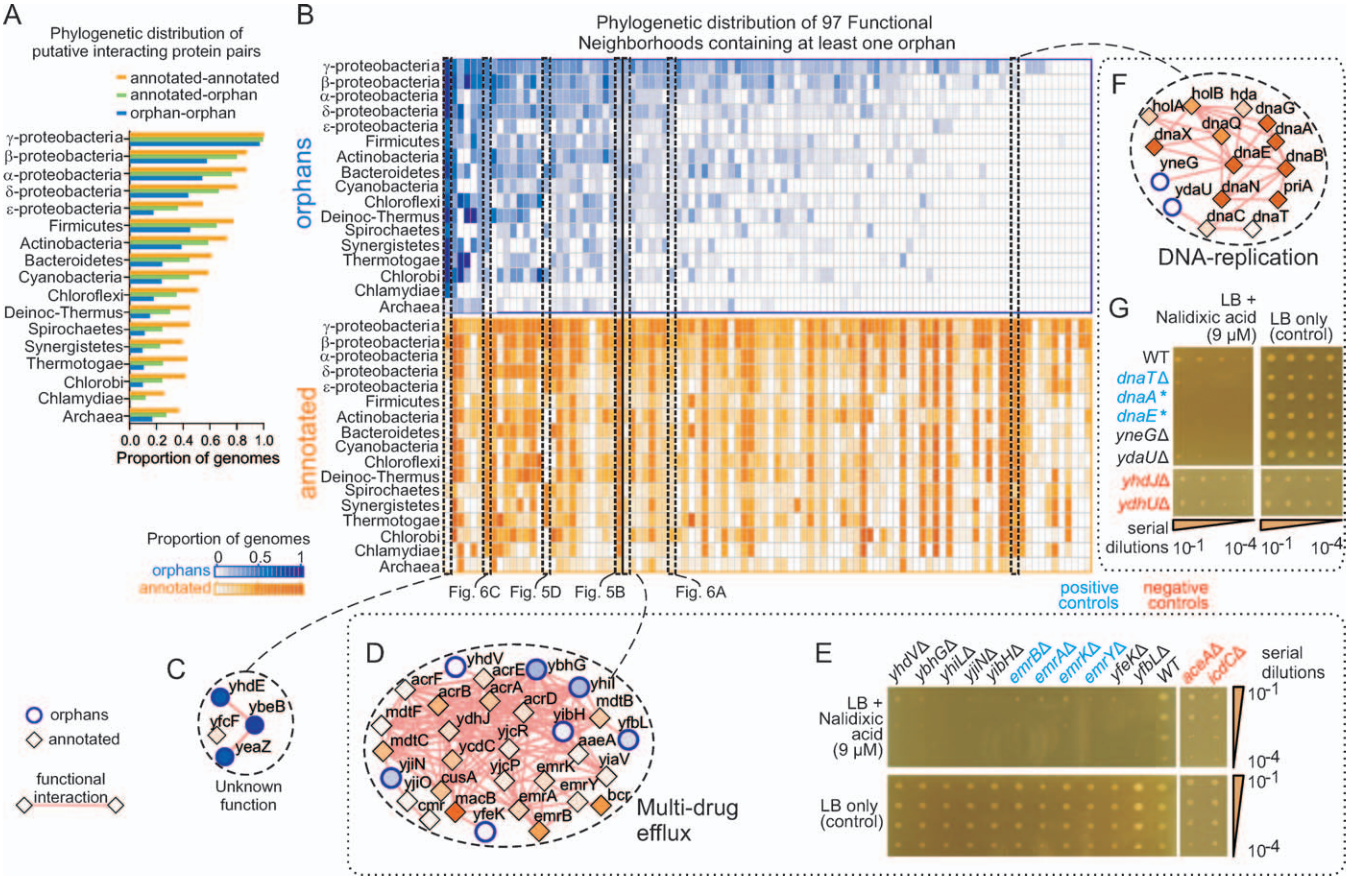


Figure 7. Evolutionary Conservation of Orphan Protein Function

(A) Evolutionary conservation of orphan and annotated *E. coli* protein interaction partners in the integrated PI-GC network based on the co-occurrence of putative orthologs across fully sequenced prokaryotic genomes.
(B) Phylogenetic distribution of the components of 97 functional neighborhoods with at least one orphan; the proportion of genomes showing conservation is indicated.
(C) Atypical neighborhood illustrating broader conservation of orphans than annotated components; node shading reflects average phylogenetic conservation.
(D) Representative neighborhood involved in drug efflux exhibiting similar phylogenetic distributions among its orphan and annotated components; analogous neighborhoods shown in preceding figures are indicated below [B].
(E) Serial-dilution assay showing the drug hypersensitivity of deletion mutants of the orphans listed in [D].
(F) DNA-replication neighborhood exhibiting a tendency of annotated components to be more widely distributed than the orphans.
(G) Serial-dilution assay showing the perturbed growth of strains deleted for the two orphan components shown in [F] in the presence of an antibiotic inhibitor of DNA-replication; an asterisk (*) indicates hypomorphic alleles. Mutants deleted of relevant annotated proteins are shown as positive controls, while mutants lacking genes not contained within this neighborhood are shown as negative controls. LB, Luria Bertani medium.
doi:10.1371/journal.pbio.1000096.g007

the biases evident in previous functional characterizations of *E. coli*. Moreover, although it might be expected that GC predictions involving gene-order conservation and phylogenetic profile similarity would be biased towards highly conserved genes, both the PI (independent of gene conservation) and functional interactions (with only operon rearrangement–based predictions less dependent on gene conservation) follow the same tendencies. Hence, if there was residual bias towards finding functional interactions for only the most highly conserved genes, it is not pronounced.

We next compared the phylogenetic distributions of the orphan and annotated components of the 97 predicted functional neighborhoods with at least one orphan (Figure 7B). Our results indicate that only four neighborhoods have orphans that are more widely distributed than annotated members. An example, consisting of three highly conserved orphans linked to metabolism and one poorly conserved annotated component is shown in Figure 7C. This pattern suggests the orphans are of primary functional significance. In contrast, most (57) other neighborhoods, including several described in previous sections (cf. Figures 5B, 5D, 6A, and 6C), generally have similar distribution profiles of orphan and annotated members (Pearson correlation >0.5), suggesting equal participation across diverse biological processes despite the incomplete previous classifications. For example, the average component distributions of one such representative neighborhood involved in drug efflux is shown in Figure 7D. Six of the seven orphans of this particular neighborhood are predicted to localize to the *E. coli* inner membrane [38]. We found that deletion of these orphans results in hypersensitivity to an otherwise exported drug, similar to that observed upon loss of the annotated components (Figure 7E), suggesting equal participation in the maintenance of cell homeostasis.

Conversely, annotated components exhibited a broader phylogenetic distribution in the remaining 36 neighborhoods. For example, two orphans (YneG and YdaU) linked to DNA replication (Figure 7F) were far more evolutionarily restricted than their annotated counterparts. Nevertheless, deletion of these same genes markedly reduced cell viability upon exposure to an inhibitor of DNA replication (Figure 7G) similar to hypomorphic strains of their broadly distributed replication partners DnaT, DnaA, and DnaE. Therefore, the sparser distribution of these orphans, and those of the other 35 neighborhoods, may reflect roles as critical accessory factors in their respective biological systems, perhaps to fine tune cellular responses to particular environmental adaptations and selective pressures (e.g., exposure to antibiotics).

Discussion

Defining the precise biological roles and relationships of bacterial gene products in an often dynamically changing physiological context is a challenging proposition. Historically, systematic assessments of protein function in bacteria have tended to rely on molecular inferences based on sequence alignments and domain architectures, whereas experimental characterization has traditionally been driven by specific scientific interests rather than with the aim of providing the broader community with unbiased collections of functionally related proteins and phenotypes. Since the biological role of a protein is not necessarily reflected in its primary sequence, the elucidation of molecular interaction networks can provide an alternate perspective even in the absence of detailed phenotypic data [16,71]. Here, we have opted to view a model microbial cell mechanistically as a series of modular molecular interaction networks that underlie the major biochemical processes that mediate cell homeostasis and proliferation, wherein the functional attributes of particular gene products are reflected in their overall patterns of associations.

To this end, we have generated an extensive compendium of physical and functional linkages covering almost the entire protein-coding complement of *E. coli*. This led to the elucidation of hundreds of putative soluble multiprotein complexes and functional modules encompassing virtually all the many gene products currently lacking public annotations. Although existing integrative probabilistic interaction databases such as STRING [61] and EcID [82] provide valuable additional binary interactions that are potentially useful for protein function prediction or as complementary evidence to those reported in this study, our machine learning strategy goes beyond describing binary interactions by explicitly describing the most probable biological functions of the orphans. Of particular noteworthiness, our functional predictions and phylogenetic projections associate a sizeable fraction of the functional orphans with core bacterial processes, suggesting they may have previously eluded detection in part due to prior analytical biases.

Since the various methods used in this study to discover different types of molecular relationships also have their own intrinsic biases, complementary information was obtained through data integration. The limited overlap between the high-confidence physical and functional interaction networks presumably stems in part from the incomplete coverage typically achieved by high-throughput experiments and their methodological differences [13,83]. For example, certain orphans were difficult to evaluate by GC methods due to a lack of apparent orthologs at medium-to-high evolutionary distances, which hinders comparative genomic inferences. Likewise, although we performed large-scale tandem affinity tagging and purification under near-native physiological conditions to generate highly purified preparations of stable, endogenous multiprotein complexes, we did not achieve complete coverage of the proteome. We did not attempt to purify a large number of membrane-associated proteins, which require specialized solubilization procedures, whereas the soluble proteins that we failed to tag or detect by mass spectrometry were presumably either of very low abundance or not expressed in our growth conditions.

Comparison of our physical interaction network with analogous public datasets produced for other model species, such as worm, fly, yeast, and even the bacterium *Helicobacter pylori*, revealed very limited ($<1\%$) overlap. These observations are congruent with recent findings by Rajagopala and colleagues [13] showing that only a third (49) of the 173 experimentally derived PI in the cell motility network of the spirochete *Treponema pallidum* predicted to occur in the ε-proteobacteria *Campylobacter jejuni* on the basis of orthology could subsequently be confirmed by targeted two-hybrid testing. The limited overlap between proteomic datasets presumably reflects a combination of incomplete coverage by various experimental assays, methodological differences and evolutionary divergence.

The observation that the intersection of functional genomics inferences with low-throughput curated physical interaction data is somewhat higher (cf. Table S11) might be explained by two nonmutually exclusive ways: first, protein–protein interactions reported in the literature based on traditional biochemical methods might be biased towards the most evolutionarily conserved multiprotein complexes, which tend to be enriched for essential components with broadly distributed phylogenetic profiles that are more easily and accurately predicted by GC methods; second, the relatively high sensitivity of the two complementary forms of protein mass spectrometry used in this study may have resulted in the detection of lower-abundance orphan proteins that have previously not been studied in depth.

The last point is consistent with the notion that different proteomic methods capture different PI types [83]. Hence, alternative proteomic methods, such as two-hybrid screens [13,84–86] or in vivo protein-fragment complementation assays [87], may be better suited for detecting certain PI currently underrepresented in our dataset. In a similar vein, additional functional relationships will undoubtedly be uncovered by different experimental and computational procedures, such as high-throughput comparative analysis of mutant cellular phenotypes [2], genome-wide genetic interaction screens [88,89], and automated text mining [90,91].

The topological properties inherent to biological networks (e.g., their hierarchical organization and degree distributions) combined with incomplete interactome coverage make establishing definitive functional groupings difficult [92]. Our approach was to take into account both the correlations among functional categories and the overall topological structure of the integrated network to generate a more balanced probabilistic model. Whereas alternative methods may provide enhanced interpretations of the organizational properties of the PI and GC networks, the functional enrichment and experimental validations established here suggest that our network-based computational inferences provide a reasonable perspective for exploring bacterial protein function. Similar strategies have resulted in powerful predictors of protein function in Eukaryotes [49,72,93–95]. The potential tradeoff is that additional error or uncertainty may have occasionally been introduced by assuming functional similarity among more loosely connected proteins. Moreover, the probabilities associated with particular functional terms may not be directly comparable. Functional orphans associated with very well-characterized biological processes are more likely to be correctly assigned by computational methods [72], whereas those associated with relatively poorly studied pathways will tend to remain obscure. Nonetheless, they can be grouped together on the basis of specific PI, GC, or even other functional associations (cf. Figure 7C) and hence serve as functional groupings rather than isolated entities.

In general, the high-confidence functional relationships we inferred for *E. coli* could be validated by independent experimental tests, and can be extrapolated to other bacterial species, including pathogens. Over 35% of the orphans find orthologs as far away as Archaea, and hence are likely associated with the same basic housekeeping processes we predict for *E. coli*, such as formation of the cell wall and protein synthesis. For instance, we have established putative roles in sugar and lipid metabolic pathways for several dozen evolutionarily conserved orphans that appear to be critical for proper biogenesis of the bacterial cell envelope, and hence may represent novel targets for antibiotic development. Conversely, our systematic comparisons also revealed some unique aspects of the orphans in the evolutionary history of *E. coli*, such as the potential fimbriae factors that appear to be restricted to Enterobacteriaceae. One interpretation is that orphans with limited phylogenetic distributions contribute to fine tuning of adaptive physiological responses upon changing environmental conditions, as previously suggested for peripheral metabolic genes acquired by horizontal transfer [96]. Conversely, the fact that the interactions of orphans with annotated proteins show a higher proportion of conservation across taxa implies that conserved biological systems are still to be discovered, and whose member contributions could extend across evolutionary domains. The physical and functional associations reported here are therefore presented as a Web-accessible public resource called "eNet" (http://ecoli.med.utoronto.ca; see Protocol S13 for details) to facilitate exploration of the fundamental molecular biology of bacteria in general and for hypothesis-driven studies of unique aspects pertaining to *E. coli* more specifically.

Materials and Methods

PI network generation. Large-scale SPA tagging and purifications were performed essentially as previously described [4,32]. Briefly, a DNA cassette encoding the SPA-tag and a selectable marker flanked by gene-specific targeting sequences was amplified by PCR using primers with homology to a selected locus. The cassette was then transformed and integrated using homologous recombination in the lysogenic *E. coli* strain DY330 (W3110 background), which harbors the highly efficient λ-phage–encoded homologous recombination enzymes *exo*, *bet*, and *gam* under the control of the temperature-sensitive CI857 repressor (the "Red" system), to create a C-terminal fusion with the protein of interest. Strains in which the PCR product was integrated were subjected to antibiotic selection, and tagged protein expression was confirmed by western blotting. Tagging primer sequences are available upon request.

Two complementary mass spectrometry techniques (gel-based MALDI peptide mass fingerprinting and gel-free LCMS shotgun sequencing) were used to detect physically interacting proteins. Details about the large-scale strain culture, protein extraction and purification, and protein identification procedures are provided in Protocol S3. Scoring of tentative PI from the LCMS and MALDI assays was conducted using a logistic regression model using reference PI obtained by low-throughput experiments curated in the DIP, BIND, and IntAct databases [35–37] as a positive training set. Our negative training set consisted of pairs of proteins in which one component was experimentally determined or predicted with high confidence to be cytoplasmic and the other residing in the outer membrane or the periplasm [38]; inner membrane proteins were discarded from this negative dataset since they are in physical proximity (and hence could potentially physically interact) to cytoplasmic and periplasmic proteins. Our logistic regression procedure also took into account the degree of consistency of copurifying protein pairs, balancing the tradeoff between "spoke" and "matrix" representation models of interactions within copurified groups of proteins to decrease the false discovery rate. We then combined the scores derived from LCMS and MALDI into a a single PI network using a previously established procedure for integrating probabilistic networks [61], which assumes the reliabilities of associations generated by these methods are independent (see Protocol S6 for details). To facilitate independent critical evaluation, all our processed interaction data is available through the Web site in HUPO-PSI molecular interaction reporting format (standard level 2.5).

GC network generation. The four GC methods used to predict functional interactions among *E. coli* proteins were based on: (1) functional linkages among genes which fuse to form a single open reading frame in at least one other genome, i.e., gene fusion [48]; (2) the mutual information of the coordinated presence or absence of pairs of genes across a set of 440 nonredundant genomes, i.e., phylogenetic profiles [51,97]; and the natural chromosomal association of bacterial genes in operons as detected by two alternative methods, namely (3) the tendency of genes forming operons to show small intergenic distances [98,99], and (4) the conservation of gene order, in which a confidence value for each pair of adjacent genes in the same strand was used as indicator that those genes likely form an operon, as compared with the conservation of adjacent genes in opposite strands [53]. For the last two methods, subsequent operon rearrangements were detected by genomic mapping of orthologs across 440 nonredundant bacterial genomes [55].

For all four GC methods, we used the BLAST-BDBHs as an operational definition of orthology (see Protocol S5 for details). To avoid circularity, the prediction scores of the four GC methods were benchmarked separately using proteins belonging to the same metabolic pathway according to EcoCyc [11] as positive reference set, and proteins in different pathways as negatives (Protocol S5). A single, unified high-confidence functional association network was then constructed by integrating the interaction predictions generated by the four GC methods using the same scoring model [61] used to integrate the MALDI and LCMS data (Protocol S6).

Clustering of networks. Protein clusters were generated from three different networks using MCL [40] (Figure 2): (1) the PI network (generating protein complexes); (2) the unified GC network (generating functional modules); and (3) the function prediction/annotation profiles derived from the integration of PI and GC networks (generating functional neighborhoods). The core idea of MCL is to simulate random walks (i.e., flux) among the proteins (nodes) within each network to delimit regions with high flux, taking into account the connectivity and weight of interaction edges. In this work, edge weights correspond to the likelihood of pairwise protein interactions in each network. In each case, the global MCL inflation parameter, which tunes the granularity of the delimited clusters, was optimized by balancing the mass fraction of clusters and efficiency of partitions (see Protocol S4 for details). The resulting clusters were individually assessed for functional homogeneity in terms of COG annotations as described previously [100] (Protocol S4).

Network-based function prediction and benchmarking. Our algorithm (StepPLR) for assigning biological functions is essentially a network topology–based method in which the functions of the orphans are predicted based on the functions of their associated annotated proteins in the immediate (direct) and adjacent (indirect) network vicinity (see Protocol S9 for details). Briefly, a single network integrating the high-confidence PI and GC probabilistic networks was first created using the same scoring model [61] used to integrate the PI data and the four GC networks. The weighted topological overlap [101] between each pair of protein nodes in the integrated network was then calculated to determine the correlated functional profiles based on a penalized logistic regression model (see Protocol S8 for details). Finally, a stepwise variable selection procedure to optimize function profiles in the final logistic regression was used (see Protocol S9 for details). Only functional categories with at least 15 annotated *E. coli* proteins were used in our integrated functional association network (see Table S14): 18 COG classes, corresponding to bacterial protein functions; 19 biological classes from MultiFun, in which the proteins can have multiple annotations based on different classification criteria; and 51 biological process classes in GO. Other guilt-by-association representative methods (e.g., majority-counting and chi-square–based) were also evaluated (results shown in Figure S4A). Expanded descriptions of benchmarking and other computational procedures of our function prediction algorithm are provided in Protocols S9 and S10.

Experimental validation of functional predictions. Orphans were selected for experimental validation of functional predictions based on the following criteria: (1) the orphan was predicted to perform a function for which a suitable phenotypic assay was previously reported (e.g., an antibiotic targeting the associated function was available); (2) the orphan was clearly grouped with select annotated genes, allowing the inclusion of positive as well as negative controls; and (3) the orphan had high (>0.8 confidence) function prediction score(s). Antibiotic susceptibility assays were performed by pinning orphan and annotated gene knockout mutants [2] onto solid media plates in the presence or absence of antibiotics, and then imaging and comparing colony sizes. Details of the antibiotic sensitivity, translation, and auxotrophy assays are provided in Protocol S11. Motility assays were performed with overnight *E. coli* strain cultures pinned onto rectangular Petri dishes (Singer) containing semisolid swarming agar (LB medium with 0.25% agar). The swarming phenotype was classified visually based on the cell spreading-halo diameter observed after approximately 8 h incubation at 32 °C. Biofilm formation assays were conducted essentially as described in [102], with replicate data normalized relative to wild-type controls (Protocol S11). Epistatic genetic interactions between pairs of gene mutants in *E. coli* were identified using a newly developed conjugation-based screening method [88]. Briefly, a drug resistance–marked query gene deletion in a high-frequency recombination donor strain was crossed into either single-gene deletion knockout mutants from the Keio strain collection [2] or select essential gene hypomorphs to generate double mutants. After double drug selection, synthetic lethal or sick phenotypes were scored visually according to measured colony sizes (Protocol S11).

Supporting Information

Figure S1. Influence of Gene Expression and Subcellular Localization on PI Detection

Found at doi:10.1371/journal.pbio.1000096.sg001 (37 KB PDF).

Figure S2. Functional Homogeneity, Connectivity, and Cluster Size in PI Complexes and GC Modules

Found at doi:10.1371/journal.pbio.1000096.sg002 (38 KB PDF).

Figure S3. Auxotrophy of *ydiB*Δ and *ydiN*Δ for Shikimic- and Aromatic Amino Acids

Found at doi:10.1371/journal.pbio.1000096.sg003 (620 KB PDF).

Figure S4. Precision and Recall Benchmark Analysis of Function Prediction Algorithms

Found at doi:10.1371/journal.pbio.1000096.sg004 (29 KB PDF).

Figure S5. Clustered Annotation Terms and Functional Neighborhoods

Found at doi:10.1371/journal.pbio.1000096.sg005 (348 KB PDF).

Protocol S1. Gene Product Attributes

Found at doi:10.1371/journal.pbio.1000096.sd001 (25 KB DOC).

Protocol S2. Metagenomic Analysis

Found at doi:10.1371/journal.pbio.1000096.sd002 (52 KB DOC).

Protocol S3. Proteomic Analysis

Found at doi:10.1371/journal.pbio.1000096.sd003 (85 KB DOC).

Protocol S4. Clustering to Define Protein Complexes, Functional Modules, and Neighborhoods

Found at doi:10.1371/journal.pbio.1000096.sd004 (34 KB DOC).

Protocol S5. Prediction of Functional Interactions by Genomic Context

Found at doi:10.1371/journal.pbio.1000096.sd005 (72 KB DOC).

Protocol S6. Global Integration of Different Data Sources for Function Prediction

Found at doi:10.1371/journal.pbio.1000096.sd006 (30 KB DOC).

Protocol S7. Analysis of Topological Network Properties

Found at doi:10.1371/journal.pbio.1000096.sd007 (28 KB DOC).

Protocol S8. Calculating Node Similarity in the Integrated Functional Association Network

Found at doi:10.1371/journal.pbio.1000096.sd008 (34 KB DOC).

Protocol S9. Network-Based Protein Function Prediction

Found at doi:10.1371/journal.pbio.1000096.sd009 (154 KB DOC).

Protocol S10. Comparison of Our New Algorithm (StepPLR) with Established Methods

Found at doi:10.1371/journal.pbio.1000096.sd010 (33 KB DOC).

Protocol S11. Experimental Validations of Functional Predictions

Found at doi:10.1371/journal.pbio.1000096.sd011 (47 KB DOC).

Protocol S12. Analysis of Functional Interactions of Orphans Extend beyond Proteobacteria

Found at doi:10.1371/journal.pbio.1000096.sd012 (22 KB DOC).

Protocol S13. Public Web Server Details

Found at doi:10.1371/journal.pbio.1000096.sd013 (38 KB DOC).

Table S1. *E. coli* K-12 (W3110) Gene Annotations and Properties

Found at doi:10.1371/journal.pbio.1000096.st001 (1.73 MB XLS).

Table S2. Genomic and Metagenomic Conservation of *E. coli* K-12 (W3110) Genes

Found at doi:10.1371/journal.pbio.1000096.st002 (2.17 MB XLS).

Table S3. Performance Comparison of Different Methods in PI Analysis

Found at doi:10.1371/journal.pbio.1000096.st003 (8 KB PDF).

Table S4. Reciprocal PI

Found at doi:10.1371/journal.pbio.1000096.st004 (75 KB XLS).

Table S5. Promiscuous "Hub" Proteins Filtered from the PI Dataset

Found at doi:10.1371/journal.pbio.1000096.st005 (7 KB PDF).

Table S6. Integrated PI Network Data

Found at doi:10.1371/journal.pbio.1000096.st006 (2.15 MB XLS).

Table S7. Predicted Protein Complexes

Found at doi:10.1371/journal.pbio.1000096.st007 (40 KB TXT).

Table S8. Integrated Functional Interaction Data Generated by Genomic Context

Found at doi:10.1371/journal.pbio.1000096.st008 (4.08 MB TXT).

Table S9. Predicted Functional Modules

Found at doi:10.1371/journal.pbio.1000096.st009 (81 KB TXT).

Table S10. Association of the Orphan Genes in Fimbrial Module to Biofilm Gene Expression Studies

Found at doi:10.1371/journal.pbio.1000096.st010 (30 KB XLS).

Table S11. Comparison of PI versus GC Interactions

Found at doi:10.1371/journal.pbio.1000096.st011 (8 KB PDF).

Table S12. Integrated PI and GC Network

Found at doi:10.1371/journal.pbio.1000096.st012 (3.49 MB TXT).

Table S13. Topological PI, GC, and Integrated Network Properties

Found at doi:10.1371/journal.pbio.1000096.st013 (24 KB XLS).

Table S14. Gold Standards of Functional Categories Used for Function Prediction

Found at doi:10.1371/journal.pbio.1000096.st014 (1.15 MB XLS).

Table S15. Function Prediction Performance Measured by AUC Scores

Found at doi:10.1371/journal.pbio.1000096.st015 (35 KB XLS).

Table S16. Orphan Gene Properties

Found at doi:10.1371/journal.pbio.1000096.st016 (2.25 MB XLS).

Table S17. Neighborhood Members

Found at doi:10.1371/journal.pbio.1000096.st017 (78 KB TXT).

Table S18. Relative Fitness and Functional Enrichment Analyses of Orphan and Annotated Genes Using Drug Screens

Found at doi:10.1371/journal.pbio.1000096.st018 (151 KB XLS).

Acknowledgments

We thank John Parkinson, Quaid Morris, and Radhakrishnan Mahadevan for helpful suggestions; and Ziqi Zhang and other members of the Emili, Greenblatt, and Moreno-Hagelsieb laboratories for technical assistance and critical input. GM-H acknowledges having used computer resources from the SHARCNET (Shared Hierarchical Academic Research Computing Network) consortium.

Author contributions. GM-H and AE conceived and designed the experiments. PH, SCJ, MB, JJDM, GB, WY, OP, XG, PW, SC, CC, AN-A, NKK, GM, MA, NN, VE, AG, AP, and GM-H performed the experiments. PH, SCJ, MB, JJDM, GB, SP, and GM-H analyzed the data. JFG provided critical suggestions and data interpretations. PH, SCJ, MB, JJDM, GB, JFG, GM-H, and AE wrote the paper.

Funding. This work was supported with grants from the Canadian Institutes of Health Research (MOP-77639), the Ontario Research and Development Challenge Fund, Genome Canada, and the Ontario Genomics Institute to JFG and AE, and from the Natural Sciences and Engineering Research Council of Canada to AG and GM-H.

Competing interests. The authors have declared that no competing interests exist.

References

1. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, et al. (2006) Large-scale identification of protein-protein interaction of Escherichia coli K-12. Genome Res 16: 686–691.
2. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006.0008.
3. Barrett CL, Herring CD, Reed JL, Palsson BO (2005) The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. Proc Natl Acad Sci U S A 102: 19103–19108.
4. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, et al. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature 433: 531–537.
5. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5: e8. doi:10.1371/journal.pbio.0050008
6. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3: 121.
7. Joyce AR, Reed JL, White A, Edwards R, Osterman A, et al. (2006) Experimental and computational assessment of conditionally essential genes in Escherichia coli. J Bacteriol 188: 8259–8271.
8. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, et al. (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. Nucleic Acids Res 34: 1–9.
9. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, et al. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res 36: D120–124.
10. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
11. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, et al. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res 33: D334–337.
12. Hawkins T, Kihara D (2007) Function prediction of uncharacterized proteins. J Bioinform Comput Biol 5: 1–30.
13. Rajagopala SV, Titz B, Goll J, Parrish JR, Wohlbold K, et al. (2007) The protein network of bacterial motility. Mol Syst Biol 3: 128.
14. Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T (2008) Functional maps of protein complexes from quantitative genetic interaction data. PLoS Comput Biol 4: e1000065. doi:10.1371/journal.pcbi.1000065
15. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, et al. (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature 436: 861–865.
16. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, et al. (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nat Genet 40: 181–188.
17. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, et al. (2005) Discovery of biological networks from diverse functional genomic data. Genome Biol 6: R114.
18. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, et al. (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J Biol 5: 11.
19. Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparison. Nat Biotechnol 24: 427–433.
20. Campillos M, von Mering C, Jensen LJ, Bork P (2006) Identification and analysis of evolutionarily cohesive functional modules in protein networks. Genome Res 16: 374–382.
21. Slonim N, Elemento O, Tavazoie S (2006) Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. Mol Syst Biol 2: 2006 0005.
22. Yellaboina S, Goyal K, Mande SC (2007) Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data. Genome Res 17: 527–535.
23. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115–119.
24. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501–504.
25. Serres MH, Goswami S, Riley M (2004) GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. Nucleic Acids Res 32: D300–302.
26. Rudd KE (1998) Linkage map of Escherichia coli K-12, edition 10: the physical map. Microbiol Mol Biol Rev 62: 985–1019.
27. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.
28. Selinger DW, Saxena RM, Cheung KJ, Church GM, Rosenow C (2003) Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. Genome Res 13: 216–223.
29. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J (2004) The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res 32: D235–239.
30. Yao Z, Ruzzo WL (2006) A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics 7 Suppl 1: S11.
31. Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3: e43. doi:10.1371/journal.pcbi.0030043
32. Zeghouf M, Li J, Butland G, Borkowska A, Canadien V, et al. (2004) Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. J Proteome Res 3: 463–468.
33. Zhang B, Park BH, Karpinets T, Samatova NF (2008) From pull-down data to protein interaction networks and complexes with biological relevance. Bioinformatics 24: 979–986.
34. Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T (2006) A direct comparison of protein interaction confidence assignment schemes. BMC Bioinformatics 7: 360.
35. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, et al. (2000) DIP: the database of interacting proteins. Nucleic Acids Res 28: 289–291.
36. Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31: 248–250.
37. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, et al. (2007) IntAct–open source resource for molecular interaction data. Nucleic Acids Res 35: D561–565.
38. Diaz-Mejia JJ, Babu M, Emili A (2009) Computational and experimental approaches to chart the Escherichia coli cell-envelope-associated proteome and interactome. FEMS Microbiol Rev 33: 66–97.
39. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643.

40. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.
41. Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.
42. King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. Bioinformatics 20: 3013–3020.
43. Sabina J, Dover N, Templeton LJ, Smulski DR, Soll D, et al. (2003) Interfering with different steps of protein synthesis explored by transcriptional profiling of Escherichia coli K-12. J Bacteriol 185: 6158–6170.
44. Barkan A, Klipcan L, Ostersetzer O, Kawamura T, Asakura Y, et al. (2007) The CRM domain: an RNA binding module derived from an ancient ribosome-associated protein. RNA 13: 55–64.
45. Deana A, Celesnik H, Belasco JG (2008) The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. Nature 451: 355–358.
46. Jiang M, Datta K, Walker A, Strahler J, Bagamasbad P, et al. (2006) The Escherichia coli GTPase CgtAE is involved in late steps of large ribosome assembly. J Bacteriol 188: 6757–6770.
47. Jiang M, Sullivan SM, Walker AK, Strahler JR, Andrews PC, et al. (2007) Identification of novel Escherichia coli ribosome-associated proteins using isobaric tags and multidimensional protein identification techniques. J Bacteriol 189: 3434–3444.
48. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86–90.
49. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285: 751–753.
50. Gaasterland T, Ragan MA (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microb Comp Genomics 3: 199–217.
51. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
52. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324–328.
53. Janga SC, Moreno-Hagelsieb G (2004) Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res 32: 5392–5397.
54. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96: 2896–2901.
55. Janga SC, Collado-Vides J, Moreno-Hagelsieb G (2005) Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. Nucleic Acids Res 33: 2521–2530.
56. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, et al. (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30: 2212–2223.
57. Snel B, Bork P, Huynen MA (2002) The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci U S A 99: 5890–5895.
58. Chua HN, Sung WK, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 22: 1623–1630.
59. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306: 1555–1558.
60. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21 Suppl 1: i302–310.
61. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33: D433–437.
62. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, et al. (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol 5: R35.
63. Fronzes R, Remaut H, Waksman G (2008) Architectures and biogenesis of non-flagellar protein appendages in Gram-negative bacteria. EMBO J 27: 2271–2280.
64. Hahn E, Wild P, Hermanns U, Sebbel P, Glockshuber R, et al. (2002) Exploring the 3D molecular architecture of Escherichia coli type 1 pili. J Mol Biol 323: 845–857.
65. Nuccio SP, Baumler AJ (2007) Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. Microbiol Mol Biol Rev 71: 551–575.
66. Domka J, Lee J, Bansal T, Wood TK (2007) Temporal gene-expression in Escherichia coli K-12 biofilms. Environ Microbiol 9: 332–346.
67. Domka J, Lee J, Wood TK (2006) YliH (BssR) and YceP (BssS) regulate Escherichia coli K-12 biofilm formation by influencing cell signaling. Appl Environ Microbiol 72: 2449–2459.
68. Johansson L, Liden G (2006) Transcriptome analysis of a shikimic acid producing strain of Escherichia coli W3110 grown under carbon- and phosphate-limited conditions. J Biotechnol 126: 528–545.
69. Lobner-Olesen A, Marinus MG (1992) Identification of the gene (aroK) encoding shikimic acid kinase I of Escherichia coli. J Bacteriol 174: 525–529.
70. Beyer A, Bandyopadhyay S, Ideker T (2007) Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet 8: 699–710.
71. Ideker T, Sharan R (2008) Protein networks in disease. Genome Res 18: 644–652.
72. Myers CL, Troyanskaya OG (2007) Context-sensitive data integration and prediction of biological networks. Bioinformatics 23: 2322–2330.
73. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
74. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, et al. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32: D262–266.
75. Bresolin G, Neuhaus K, Scherer S, Fuchs TM (2006) Transcriptional analysis of long-term adaptation of Yersinia enterocolitica to low-temperature growth. J Bacteriol 188: 2945–2958.
76. Girgis HS, Liu Y, Ryu WS, Tavazoie S (2007) A comprehensive genetic characterization of bacterial motility. PLoS Genet 3: 1644–1660. doi:10.1371/journal.pgen.0030154
77. Kosinski J, Feder M, Bujnicki JM (2005) The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. BMC Bioinformatics 6: 172.
78. Daley DO, Rapp M, Granseth E, Melen K, Drew D, et al. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. Science 308: 1321–1323.
79. Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338: 1027–1036.
80. Breazeale SD, Ribeiro AA, McClerren AL, Raetz CR (2005) A formyltransferase required for polymyxin resistance in Escherichia coli and the modification of lipid A with 4-Amino-4-deoxy-L-arabinose. Identification and function of UDP-4-deoxy-4-formamido-L-arabinose. J Biol Chem 280: 14154–14167.
81. Yan A, Guan Z, Raetz CR (2007) An undecaprenyl phosphate-aminoarabinose flippase required for polymyxin resistance in Escherichia coli. J Biol Chem 282: 36077–36089.
82. Andres-Leon EA, Ezkurdia I, Garcia B, Valencia A, Juan D (2009) EcID. A database for the inference of functional interactions in E. coli. Nucleic Acids Res 37: D629–D635.
83. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322: 104–110.
84. Parrish JR, Yu J, Liu G, Hines JA, Chan JE, et al. (2007) A proteome-wide protein interaction map for Campylobacter jejuni. Genome Biol 8: R130.
85. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, et al. (2001) The protein-protein interaction map of Helicobacter pylori. Nature 409: 211–215.
86. Titz B, Rajagopala SV, Goll J, Hauser R, McKevitt MT, et al. (2008) The binary protein interactome of Treponema pallidum–the syphilis spirochete. PLoS ONE 3: e2292. doi:10.1371/journal.pone.0002292
87. Tarassov K, Messier V, Landry CR, Radinovic S, Molina MM, et al. (2008) An in vivo map of the yeast protein interactome. Science 320: 1465–1470.
88. Butland G, Babu M, Diaz-Mejia JJ, Bohdana F, Phanse S, et al. (2008) eSGA: E. coli synthetic genetic array analysis. Nat Methods 5: 789–795.
89. Typas A, Nichols RJ, Siegele DA, Shales M, Collins SR, et al. (2008) High-throughput, quantitative analyses of genetic interactions in E. coli. Nat Methods 5: 781.
90. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36: 664.
91. Rzhetsky A, Seringhaus M, Gerstein M (2008) Seeking a new biology through text mining. Cell 134: 9–13.
92. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3: 88.
93. McDermott J, Bumgarner R, Samudrala R (2005) Functional annotation from predicted protein interaction networks. Bioinformatics 21: 3217–3226.
94. Murali TM, Wu CJ, Kasif S (2006) The art of gene function prediction. Nat Biotechnol 24: 1474–1475.
95. Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. Nat Biotechnol 18: 1257–1261.
96. Pal C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37: 1372–1375.
97. Moreno-Hagelsieb G, Janga SC (2008) Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. Proteins 70: 344–352.
98. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci U S A 97: 6652–6657.
99. Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18 Suppl 1: S329–336.
100. Loganantharaj R, Cheepala S, Clifford J (2006) Metric for measuring the

effectiveness of clustering of DNA microarray expression. BMC Bioinformatics 7 Suppl 2: S5.

101. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4: Article17.
102. O'Toole GA, Pratt LA, Watnick PI, Newman DK, Weaver VB, et al. (1999) Genetic approaches to study of biofilms. *Methods Enzymol* 310: 91–109.
103. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, et al. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res 36: D866–870.
104. Vlasblom J, Wu S, Pu S, Superina M, Liu G, et al. (2006) GenePro: a Cytoscape plug-in for advanced visualization and analysis of interaction networks. Bioinformatics 22: 2178–2179.