The binary protein-protein interaction landscape of *Escherichia coli*

Seesandra V Rajagopala1, Patricia Sikorski1,10, Ashwani Kumar2,10, Roberto Mosca3,10, James Vlasblom2, Roland Arnold4, Jonathan Franca-Koh1, Suman B Pakala1, Sadhna Phanse2, Arnaud Ceol3, Roman Häuser5, Gabriella Siszler5, Stefan Wuchty6,9, Andrew Emili4, Mohan Babu2, Patrick Aloy3,7, Rembert Pieper1 & Peter Uetz8

Efforts to map the *Escherichia coli* interactome have identified several hundred macromolecular complexes, but direct binary protein-protein interactions (PPIs) have not been surveyed on a large scale. Here we performed yeast two-hybrid screens of 3,305 baits against 3,606 preys (~70% of the *E. coli* proteome) in duplicate to generate a map of 2,234 interactions, which approximately doubles the number of known binary PPIs in *E. coli*. Integration of binary PPI and genetic-interaction data revealed functional dependencies among components involved in cellular processes, including envelope integrity, flagellum assembly and protein quality control. Many of the binary interactions that we could map in multiprotein complexes were informative regarding internal topology of complexes and indicated that interactions in complexes are substantially more conserved than those interactions connecting different complexes. This resource will be useful for inferring bacterial gene function and provides a draft reference of the basic physical wiring network of this evolutionarily important model microbe.

Mapping PPIs is essential to understand biological function. High-throughput tandem affinity purification followed by mass spectrometry (AP-MS) has been used to identify PPIs on a large scale in *E. coli*1–3, but this method cannot be used to distinguish direct interactions when several proteins are purified together. Various high-throughput biochemical approaches have been proposed to examine binary PPIs, including the yeast two-hybrid (Y2H) method, which has been applied in several eukaryotes4,5, prokaryotes6 and viruses7. Here we explored the global architecture of the binary interactome in the gram-negative bacterium *E. coli* by carrying out proteome-scale, high-throughput Y2H screening. We detected network associations for 1,269 *E. coli* proteins and more than half of the interactions are new. We derived a comprehensive interactome of *E. coli* by integrating literature-curated binary PPIs with Y2H interactions identified in this study and analyzed the network structure and function of this *E. coli* interactome.

**RESULTS**

**Testing the Y2H system for *E. coli* proteins**

We constructed *E. coli* Y2H libraries by transferring 3,971 Gateway-compatible *E. coli* entry clones8 into three different Y2H vectors (pGADT7g, pGBK7Tg and pGBKCatg; Fig. 1a and Supplementary Fig. 1a). After the removal of autoactivating baits9, our libraries contained 3,305 baits (in pGBK7Tg as fusions to a DNA-binding domain) and 3,606 preys (in pGADT7g as fusions to an activation domain) (Fig. 1a).

To evaluate the efficacy of the Y2H system for *E. coli* proteins, we first benchmarked our system with a reference set of likely (positive) and unlikely (random) PPIs. We compiled the positive reference set (PRS) from manually curated databases10,11. Approximately 1,941 binary PPIs have been documented for *E. coli*, most of which have been reported as purified dimers or binary complexes in AP-MS studies. Previous studies have shown that interactions supported by multiple publications are more reliable than those supported by a single publication12, so we selected 303 PPIs that had been described in multiple publications or had been characterized by two independent methods.

Our pGBK7Tg-pGADT7g library contains both bait and prey clones for 212 PPIs from the selected 303 PPIs (Supplementary Table 1a). From these 212 PPIs, we randomly selected 94 PPIs as a PRS (Supplementary Table 1b) and tested them using the pGBK7Tg-pGADT7g Y2H vector systems (Fig. 1b and Online Methods). We also tested 100 PRS interactions using the pGBKCatg-pGADT7g vector system (Supplementary Table 1c and Online Methods). Approximately 29% (27 of 94) of the PRS interactions were detected by the pGBK7Tg-pGADT7g system, a sensitivity that is comparable to yeast and human Y2H array screening12,13. We assayed 500 randomly chosen protein pairs comprising the random reference set and detected an interaction for only four pairs, a specificity of ~99% (Fig. 1c and Supplementary Table 1d). However, this number cannot be directly extrapolated to the whole data set, given that most false positives arise through a small

Received 13 July 2013; accepted 16 January 2014; published online 23 February 2014; doi:10.1038/nbt.2831
set of promiscuous baits and preys. We removed such nonspecific interactions from our Y2H data set (Online Methods).

High-throughput Y2H screening

We conducted the Y2H screening with pools of three baits and selected positive yeast colonies on plates containing synthetic medium (Supplementary Fig. 1c and Online Methods). We identified the interacting preys by yeast colony PCR, followed by DNA sequencing. We retested the resulting interactions in quadruplicate for each bait-prey combination using fresh yeast bait and prey cultures from archival stocks and eliminated interactions that were either nonreproducible or autoactivated (Supplementary Fig. 1d)\(^9,14\). By this approach, we systematically screened 3,305 baits against 3,606 preys, thus covering ~70% of the \textit{E. coli} interactome space (Fig. 1d and Supplementary Fig. 2).

When screening millions of protein pairs, the search space needs to be sampled multiple times to detect all the interactions that are detectable by the assay used\(^5,15\). Therefore, to estimate the sampling sensitivity of our Y2H system, we performed eight independent sampling screens using 92 baits against the entire prey library (3,606 proteins), which resulted in saturation in the number of detectable interactions (Fig. 1e). Based on this analysis, we estimated a mean sampling sensitivity of 69% (±2.5%, s.e.m.) for a combination of two screens; at least fivefold sampling would be required to obtain 90% saturation (Fig. 1e). We sampled the entire search space twice independently, which resulted in 2,234 high-quality PPIs among 1,269 proteins (Fig. 1e). Based on this analysis, we estimated a mean sampling sensitivity of 69% (±2.5%, s.e.m.) for a combination of two screens; at least fivefold sampling would be required to obtain 90% saturation (Fig. 1e). We sampled the entire search space twice independently, which resulted in 2,234 high-quality PPIs among 1,269 proteins (Fig. 1e).

Of the 2,234 PPIs reported in this study, 239 are supported by manually curated PPIs from the literature\(^10,11\), and another 165 are supported by AP-MS studies\(^1-3\) (Fig. 2d). The limited overlap with the literature data is mainly due to low Y2H assay sensitivity (i.e., many false negatives) rather than low specificity (i.e., false positives; Fig. 1c), which is consistent with the low overlap observed for the yeast interactome\(^5\).

We have shown previously that the Y2H system can be used effectively to analyze direct interactions among proteins in a complex\(^7\). An assessment of our Y2H interactions against structures of \textit{E. coli} protein complexes from the Protein Data Bank\(^8\) (Supplementary Fig. 3a–d and Supplementary Table 4) revealed that >90% (101 of 110) of interactions among complex subunits corresponded to direct PPIs (Supplementary Fig. 3 and Supplementary Table 4)\(^1-3\). We then tested 227 putative \textit{E. coli} protein complexes containing three or more protein subunits\(^5\) by pair-wise Y2H assays of the subunits. To enhance assay sensitivity, we conducted the Y2H screening using both N-terminal and C-terminal fusion bait proteins\(^11\). By this method, we identified 458 PPIs between the subunits of the complexes (Fig. 2e, Supplementary Fig. 4 and Online Methods).

High-quality binary interaction network

To obtain a more comprehensive binary \textit{E. coli} interactome, we combined the high-quality Y2H data set from this study with binary interactions that we manually curated from the literature\(^10,11\). This data set, named ‘combined-binary’ data set, contains 3,946 binary interactions among 2,048 proteins (Fig. 3 and Supplementary Table 5). Based on our measurements of precision, completeness, assay sensitivity and sampling sensitivity, we estimated that the total size of the \textit{E. coli} interactome is on the order of ~10,000 PPIs (Supplementary Results).

The combined-binary data set described in this study comprises ~39% of this estimate of the \textit{E. coli} binary interactome. We estimate that approximately half of the interactions in the combined-binary interaction data set are likely to be between components of multimeric complexes (i.e., complexes inferred from AP-MS data\(^3\) and complexes from the EcoCyc database) and that the remaining half are transient binary interactions (Supplementary Results).

As is the case for other biological networks\(^19\), interactions in both the combined-binary data set and the combined AP-MS data set\(^2,3\) are
Figure 2  Quality assessment and comparison of Y2H interactions with data from the literature. (a) Immunoblots of 6 of the 114 randomly selected PPIs identified by Y2H assay that were tested by co-immunoprecipitation. Plasmids encoding interacting pairs and negative controls (‘vector’) were co-transformed into E. coli, and protein binding was detected by immunoblotting (Online Methods). Red boxes indicate each pair of interactions. (b) LUMIER assay\(^1\) of the 114 randomly selected interacting protein pairs mentioned in a. Interactions were scored as positive when they resulted in a luminescence intensity ratio (LIR) > 3 and \(P < 0.05\). Lines indicate LIR and \(P\)-value thresholds: (log(LIR) > 0.477; log(\(P\)) < −1.30). (c) Number of PPIs validated by coimmunoprecipitation (Co-IP), LUMIER and both assays. (d) Overlap between interactions detected in this large-scale Y2H study and in the literature. Manually curated ‘literature-binary’ PPIs were compiled from the microbial protein interaction database (MPIDB; http://jcvi.org/mpidb/about.php) which consists of 1,941 manually curated binary PPIs. The combined AP-MS PPIs were predicted from large-scale AP-MS studies\(^1\) and consists of 20,425 PPIs. (e) Pipeline used to detect direct interactions in complexes by Y2H matrix screening.

As in yeast\(^2\), we found that physically interacting proteins from the combined-binary data set and from the combined AP-MS data set were more likely to have highly correlated genetic interaction (GI) profiles than randomly selected protein pairs (\(P \leq 0.05\), Student’s \(t\)-test; Fig. 4c). Physically interacting protein pairs in both data sets also showed positively correlated coexpression profiles (\(P \leq 0.05\)) and phenotypic profiles (\(P \leq 0.05\)) (Fig. 4d,e). Moreover, the average semantic similarity\(^3\) of Gene Ontology (GO) annotations in the combined-binary PPI network was comparable to that of the combined AP-MS network, suggesting that the binary PPIs are as functionally coherent as the AP-MS–derived PPIs (Fig. 4f).

Figure 3  Structural analysis of the E. coli binary interactome. Nodes represent proteins and are colored according to the availability of structural data. Edges represent PPIs and are colored according to their source (literature, our Y2H experiment or both). The subnetworks (enlarged parts) are based on a previous AP-MS study\(^2\) and for which Y2H and literature binary interactions provided the topology for a subcomplex of at least three components. Protein structural data are overlaid. For example, protein complex 42 (bottom left) contains seven proteins; the Y2H binary interactions suggest a topology for five subunits in the complex, and all the subunits have either a complete structure or a model. For two interactions between three different components we have structures of the binary subcomplexes. For complexes 72, 100 and 103 the Y2H binary interactions identified in our screens suggest an almost complete topology. Complexes are numbered according to ref. 3.
AP-MS–derived information about complexes does not provide information about the internal topology of multiprotein assemblies\(^2,3\). We therefore assessed whether Y2H experiments could be used to detect direct physical interactions in complexes identified in large-scale AP-MS experiments. We compiled a list of 227 *E. coli* protein complexes that have three or more components as identified in a large-scale AP-MS screen\(^3\). Next, we identified interactions between subunits of these complexes in our Y2H data set and among literature-curated binary interactions (Supplementary Table 6 and Online Methods). We mapped 745 binary interactions in 203 complexes (Supplementary Fig. 6), of which 319 (43%) interactions were identified in this study and 426 (57%) were obtained from the literature (108 of which were recapitulated in our Y2H study). We could deduce a putative complete internal topology for 15 multiprotein complexes; for these 15 complexes the binary interactions connected all components of the complex, thereby providing a hypothesis on connectivity in the complex. For another 45 complexes we determined the putative internal topology of a subcomplex with at least three subunits, and for 46 complexes we determined the internal connections between pairs of

![Figure 4](image1)  
**Figure 4** Comparison of the properties of binary and AP-MS interaction networks. (a) The degree of a node in a network (degree distribution) involving essential *E. coli* protein pairs from the combined-binary network (data from this study and literature binary interactions\(^10\)) and combined AP-MS network\(^2,3\). (b) Frequency distribution involving protein pairs from the combined-binary and combined AP-MS PPI networks at different path lengths. Data shown in (a) and (b) were sampled by considering the same number of interactions among the same number of nodes in the two data sets. (c) Distribution of the Pearson correlation coefficient (PCC) between the GI profiles for gene pairs encoding interacting proteins derived from combined-binary network or combined–AP-MS network versus random gene pairs. (d, e) Distribution of coexpression profiles (d) and condition-dependent phenotypic correlation profiles (e) with corresponding interacting proteins, shown as in (c). P values (c–e; AP-MS versus random (blue) and combined-binary versus random (brown)) were computed using Student’s t-test. (f) Average semantic similarity of the combined binary and combined AP-MS PPI networks by GO category.

![Figure 5](image2)  
**Figure 5** Integrative analyses on the PPI and GI data. (a–d) Examples of subnetworks showing the physical and genetic connectivity among the components of various bioprocesses. Subnetworks containing Y2H physical interactions (gray edges) among 1,269 proteins are derived using a Markov clustering approach. The GIs (red edges for negative interactions and green for positive interactions) from the published large scale eSGA surveys\(^24,29,31\) were overlaid on the PPI network. Subnetworks with positive and/or negative interactions are highlighted with shaded ovals: secretion components (a; cluster 5); flagellum or motility components (b; cluster 9); subunits of ATP-dependent protease complexes (c; cluster 14); and ycfM and pepG (d; cluster 8). For details on GIs in each of the subnetworks and cluster identifiers, see Supplementary Tables 7 and 8. Large nodes in each subnetwork represent genetic associations with the indicated protein subunits; small nodes indicate an absence of GIs.
subunits (Supplementary Fig. 6). For an additional 97 complexes we could only map homomeric interactions (that is, interactions between multiple copies of the same protein). We found in the literature a structure (or a homology model) for 1,097 proteins and 202 interactions (Supplementary Fig. 7) in the complexes. For several complexes, we found a subcomplex of at least three components for which complete topology was revealed by experimental binary interaction data and structural data were available for all its subunits (Fig. 3 and Supplementary Fig. 8).

Integration of PPI and GI networks

Identification of physical interactions among proteins does not necessarily imply that all interacting partners belong to the same functional pathway. By contrast, genetic screens can elucidate functional relationships between genes. Analogous to the yeast synthetic genetic array approach, we have previously developed a high-throughput synthetic genetic array screening technology to map GI networks in E. coli, termed eSGA, to elucidate pathway-level relationships. Integration of these GI data with the PPI networks should provide information that is not attainable from either study alone. Using a Markov clustering algorithm, we identified physically related groupings in the binary interaction network and overlaid them with currently available large-scale E. coli GI data sets (Supplementary Table 7) for the corresponding protein interactions, to determine the functional relationships of individual proteins and multiprotein complexes in various pathways and processes (Fig. 5, and Supplementary Tables 7 and 8).

This integrative analysis captured interactions that had been missed by either one of the methods. For example, in the secretary pathway the inner membrane protein SecA is peripherally associated with the multisubunit translocation apparatus SecYEG; Y2H captured the physical interaction between SecA and SecY, but eSGA did not identify this interaction (Fig. 5a). However, interactions between other Sec components missed by Y2H were detected as GIs by eSGA. Also, for components of the flagellum (for example, fliN-cheR and fliA-fliC; Fig. 5b), physical interactions among the proteins are overlaid by a negative phenotype in the GI network, consistent with their joint participation in flagellum assembly and motility. Another example involves a subnetwork of physically connected subunits of the ATP-dependent protease complex (ClpAX) known for their cooperative function in protein quality control. These had a positive GI phenotype (Fig. 5c), suggesting that they operate in the same pathway. Similarly, the Y2H interaction between the recently characterized outer membrane lipoprotein YcfM (also known as LpoB) and penicillin-binding protein PbpG showed a strong positive interaction (Fig. 5d), consistent with a recently proposed regulatory role of YcfM in peptidoglycan synthesis.

Conservation of binary PPIs

To investigate the evolutionary significance of PPIs detected in E. coli with respect to other microbes, we examined the presence of PPI orthologs in other bacteria (Fig. 6a) by using the combined-binary interactions of E. coli, and binary interactions reported for other bacteria in a literature-curated database and from large-scale bacterial interactome studies (Online Methods). Previous investigations have shown that interactions identified in yeast are conserved in worms and that interactions identified in metazoans can predict homologous connections in yeast. We found that interacting protein pairs from our study are highly conserved in bacteria that are closely related to E. coli but are less conserved with increasing distance from E. coli in the phylogenetic tree. However, we observed the opposite trend for interactions between proteins encoded by essential E. coli genes (Fig. 6b). We also found that interactions between the subunits of protein complexes are more conserved than interactions between proteins belonging to different complexes (Fig. 6b).

DISCUSSION

Our study substantially expands the available resources for protein-complex information and the connections of protein complexes to pathways in E. coli. Y2H and AP-MS studies provide largely complementary information about the interactome and both are essential to obtaining complete protein networks in prokaryotic biology. Our data set confirmed some known binary interactions and contains many new PPIs within and between complexes. Binary PPI maps also provide information about the internal topology of multimeric protein complexes,
which should allow them to be structurally modeled, after the
stoichiometry of each component is known. We illustrated this using
structures of two well-characterized E. coli complexes: RecBCD and
the F1 portion of the ATP synthase complex containing three copies
of each of its subunits (alpha, beta and delta, AtpD; Fig. 3).

Comparison of the binary interactome to complexes found through
AP-MS and in the EcoCyc database (Supplementary Tables 9 and 10)
suggests that the majority of PPIs are not between members of
complexes. We suspect that many protein complexes remain unde-
tected and may form only under certain growth conditions; the Y2H
assay may detect PPIs irrespective of the growth condition in E. coli as
it is carried out in yeast. Furthermore, the conditions used for affinity
purification may preclude the detection of less stable interactions that
are detectable by Y2H. Our data suggest that there are distinct subsets
of interactions found in complexes or as more transient binary
interactions. Based on the assessment of the total number of interactions
in each group, we estimate that the combined total is on the order of
9,200–11,400 interactions (Supplementary Results). To our knowl-
edge, this study represents the largest experimentally derived catalog
to date of the E. coli binary interaction network and thus is a draft
reference of the basic physical wiring network of an evolutionarily
well-conserved model microbe.

METHODS

Methods and any associated references are available in the online
version of the paper.

Accession codes. International Molecular Exchange: IM-22059.

Note: Any Supplementary Information and Source Data files are available in the
online version of the paper.

ACKNOWLEDGMENTS

S.V.R. and P.U. were supported by US National Institutes of Health grant
GM079710. P.U., R.H. and P.A. were supported by the Seventh Research
Framework Programme of the European Union (AntiPathoGN; EU grant
HEALTH-F3-2009-223101). M.B. was supported by the Discovery Grant from the
Natural Sciences and Engineering Research Council (DG-20234). We thank X. De
Bolle (University of Namur, Belgium) for providing pRH016 and pRH018 vectors.

AUTHOR CONTRIBUTIONS

The project was conceived by S.V.R. and P.U. and directed by S.V.R. and R.P.; P.S.,
J.F.K., R.H., G.S. and S.V.R. performed the experiments: S.V.R., A.K., R.M., S.B.P.,
J.V., R.A., S.P., A.C., S.W. and M.B. performed computational analysis; and
A.E. provided support for computational analysis tools. S.V.R., P.A. and M.B. wrote
the manuscript, with input from P.U.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/
reprints/index.html.

1. Artiguanan, M. et al. Large-scale identification of protein-protein interaction of
Escherichia coli K-12. Genome Res. 16, 686–691 (2006).
2. Butland, G. et al. Interaction network containing conserved and essential protein
complexes in Escherichia coli. Nature 433, 531–537 (2005).
3. Hu, P. et al. Global functional atlas of Escherichia coli encompassing previously
uncharacterized proteins. PLoS Biol. 7, e96 (2009).
4. Rual, J.F. et al. Towards a proteome-scale map of the human protein-protein
interaction network. Nature 437, 1173–1178 (2005).
5. Yu, H. et al. High-quality binary protein interaction map of the yeast interactome
network. Science 322, 104–110 (2008).
6. Titz, B. et al. The binary protein interactome of Treponema pallidum—the phyllost
spirochete. PLoS ONE 3, e2292 (2008).
7. Uetz, P. et al. Herpesviral protein networks and their interaction with the human
proteome. Science 311, 239–242 (2006).
8. Rajagopala, S.V. et al. The Escherichia coli K-12 ORFeome: a resource for
comparative molecular microbiology. BMC Genomics 11, 470 (2010).
9. Rajagopala, S.V. & Uetz, P. Analysis of protein-protein interactions using high-
throughput yeast two-hybrid screens. Methods Mol. Biol. 781, 1–29 (2011).
10. Goll, J. et al. MPIDB: the microbial protein interaction database. Bioinformatics
24, 1743–1744 (2008).
11. Rajagopala, S.V. et al. MI-LIT: a literature-curated dataset of microbial binary
protein–protein interactions. Bioinformatics 24, 2622–2627 (2008).
12. Braun, P. et al. An experimentally derived confidence score for binary protein-protein
interactions. Nat. Methods 6, 91–97 (2009).
13. Chen, Y.C., Rajagopala, S.V., Spellberger, T. & Uetz, P. Exhaustive benchmarking of
the yeast two-hybrid system. Nat. Methods 7, 667–668 (2010).
14. Vidalain, P.O., Boxem, M., Ge, H., Li, S. & Vidal, M. Increasing specificity in high-
throughput yeast two-hybrid experiments. Methods 32, 363–370 (2004).
15. Schuster, S. et al. An empirical framework for binary interactome mapping.
Nat. Methods 6, 83–90 (2009).
16. Barrios-Rodiles, M. et al. High-throughput mapping of a dynamic signaling
network in mammalian cells. Science 307, 1621–1625 (2005).
17. Uetz, P. et al. A comprehensive analysis of protein interaction networks.
Nature 403, 623–627 (2000).
18. Barabasi, A.L. Scale-free networks: a decade and beyond. Science 325, 412–413
(2009).
19. Collins, S.R. et al. Functional dissection of protein complexes involved in yeast
chromosome biology using a genetic interaction map. Nature 446, 806–810
(2007).
20. Wang, J., Du, Z., Payattakool, R., Yu, P.S. & Chen, C.J. A new method to measure
the semantic similarity of GO terms. Bioinformatics 23, 1274–1281 (2007).
21. Aloy, P. et al. Structure-based assembly of protein complexes in yeast. Science
303, 2026–2029 (2004).
22. Lasker, K. et al. Integrative structure modeling of macromolecular assemblies from
proteomics data. Mol. Cell. Proteomics 9, 1689–1702 (2010).
23. Babu, M. et al. Genetic interaction maps in Escherichia coli reveal functional
crosstalk among cell envelope biogenesis pathways. PLoS Genet. 7, e1002377
(2011).
24. Babu, M. et al. Genetic interaction maps in Escherichia coli reveal functional
crosstalk among cell envelope biogenesis pathways. PLoS Genet. 7, e1002377
(2011).
25. Bandopadhyay, S., Kelley, R., Krogan, N.J. & Ideker, T. Functional maps of protein
complexes from quantitative genetic interaction data. PLoS Comput. Biol. 4,
e1000665 (2008).
26. Beltrao, P., Cagney, G. & Krogan, N.J. Quantitative genetic interactions reveal
biological modularity. Cell 141, 739–745 (2010).
27. Boone, C., Bussey, H. & Andrews, B.J. Exploring genetic interactions and networks
with yeast. Nat. Rev. Genet. 8, 437–449 (2007).
28. Costanzo, M. et al. The genetic landscape of a cell. Science 327, 425–431
(2010).
29. Butland, G. et al. esga: E. coli synthetic genetic array analysis. Nat. Methods 5,
789–795 (2008).
30. Brohee, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein
interaction networks. BMC Bioinformatics 7, 488 (2006).
31. Babu, M. Quantitative genome-wide genetic interaction screens reveal global epistatic
relationships of protein complexes in Escherichia coli. PLoS Genet. (in the press).
32. Oliver, D.B. & Beckwith, J. P. Analysis of a pleiotropically defective in the export
of secreted proteins. Cell 25, 765–772 (1981).
33. Rajagopala, S.V. et al. The protein network of bacterial motility. Mol. Syst. Biol. 3,
128 (2007).
34. Costanzo, M., Mu, W., Seshijiohi, A.W., Zhou, J. & Shakhnovich, E.I. Protein quality
control acts on folding intermediates to shape the effects of mutations on organismal
fitness. Mol. Cell 49, 133–144 (2013).
35. Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B. & Boone, C. Systematic
mapping of genetic interaction networks. Annu. Rev. Genet. 43, 601–625 (2009).
36. Tysias, A. et al. Regulation of peptidoglycan synthesis by outer-membrane proteins.
Cell 143, 1097–1109 (2010).
37. Matthews, L.R. et al. Identification of potential interaction networks using sequence-
based searches for conserved protein-protein interactions or “interologs”. Genome
Res. 11, 2120–2126 (2001).
38. Sharan, R. et al. Conserved patterns of protein interaction in multiple species. Proc.
Natl. Acad. Sci. USA 102, 1974–1979 (2005).
39. Singleton, M.R., Dillingham, M.S., Gaudier, M., Kowalczykowski, S.C. & Wigley,
D.B. Crystal structure of RecBCD enzyme reveals a machine for processing DNA
breaks. Nature 432, 187–193 (2004).
40. Cingolani, G. & Duncan, T.M. Structure of the ATP synthase catalytic complex (F1)
from Escherichia coli in an autoinhibited conformation. Nat. Struct. Mol. Biol. 18,
701–707 (2011).
ONLINE METHODS

Y2H library construction. All entry clones (i.e., 3,971 open reading frames (ORFs)) from the E. coli ORFeome collection were transferred into activation domain (AD) and DNA-binding domain (DB) Y2H expression vectors (namely, pGADT7g, pGBGT7g and pGBKCatg) using Gateway recombination cloning (Life Technologies). After bacterial transformation, plasmid DNA was isolated using Millipore plasmid preparation kit (LSPK096234) and all AD and DB clones were transformed into Y2H strains MATa Y187 (AD, prey clones) and MATα AH109 (DB, bait clones), respectively (Supplementary Fig. 1a). All prey yeast clones were grown independently in a deep-well plate and then combined into a single pool as the prey library (Supplementary Fig. 1b). The pGBKCatg vector was constructed by adding the start codon ATG to pGBKcg vector at position 760 bp (which is after the attR1 sites); this modification enabled expression of entry clones lacking a start codon as C-terminal fusions (Supplementary Note).

Autoactivation test. All bait constructs were screened for autoactivation essentially as previously described6. Briefly, the baits were mated with empty prey vector, and the resulting diploids were grown on synthetic medium lacking leucine, tryptophan and histidine, which contained different concentrations (1 mM, 5 mM and 10 mM) of 3-amino-1,2,4-triazole (3-AT) to identify any autoactivators (Supplementary Fig. 1b). All autoactivating baits were removed from Y2H screening. A total of 3,305 non-autoactivating baits were subsequently screened with 3,606 preys (Supplementary Fig. 2).

High-throughput Y2H screening. A yeast strain expressing a single protein or a pool of three proteins with known identity as a DB fusion was mated with a single pool prey library (Supplementary Fig. 1c). After mating, the colonies are transferred to the Y2H selective medium. The interacting preys were identified by yeast-colony PCR, and the resulting positive PCR products were sequenced (Supplementary Fig. 1c)6.

Mating procedure. The fresh yeast culture of baits with known identity and prey library were mixed in a 1:1 ratio. We used four OD (optical density at 600 nm) of bait (4 ml of OD = 1) and four OD of prey (4 ml of OD = 1) cultures in a 15-ml Falcon tube. For each of the bait proteins, we included one negative control (i.e., bait and empty prey vector). The bait and prey mixture was centrifuged for 2 min at 3,000 r.p.m. at room temperature. After centrifugation, the supernatant was discarded, and the cell pellet was resuspended in 200 µl YPDA (yeast peptone dextrose adenine) liquid medium. The yeast cells suspended in YPDA were then plated on YPDA agar plate (60 mm × 15 mm), and air-dried. The plates were incubated at 30 °C for 6 h or overnight at room temperature. After incubation, the cells were collected by washing the plate with 2 ml of sterile water. The cells were then spun down to remove the supernatant. The cell pellets were washed with 2 ml of sterile water, and resuspended in 2 ml of selective liquid medium (i.e., medium lacking tryptophan, leucine and histidine). About 500 µl of the suspension was plated on the selective agar plates (i.e., medium lacking tryptophan, leucine and histidine + 5 mM 3-AT). To measure the number of diploids, an aliquot of 1:100 dilution of mated cells was plated on selective agar plates (i.e., medium lacking tryptophan and leucine). The screening depth was estimated to be 0.1 million to 1 million diploids, which is at least 20× the number of prey library size.

Interaction selection. The section medium plates were incubated at 30 °C in a humidity-controlled incubator for 4–6 d, until the colonies were ~1 mm in diameter. Those colonies that showed growth on selection plates but not on control plates were designated as two-hybrid ‘positive’ yeast colonies. The positive yeast colonies were picked either manually or using robotics and subjected to yeast-colony PCR essentially as described6. The identity of the prey was determined using end-read sequencing of the PCR products. If we used a pool of three baits with known identity, the bait identity was resolved by pair-wise Y2H retest screening.

Pair-wise Y2H test (phenotyping II). A major consideration when using the Y2H system is the number of false positives. The major sources for false positives are promiscuous (‘sticky’) preys and nonreproducible signals that arise through autoactivation. The promiscuous preys bind to many different baits and probably interact with the baits nonspecifically. From the raw Y2H data, we filtered nonspecific interactions by a simple approach based on the number of times a certain protein is found as a prey. Preys found more than 60 times (arbitrary threshold) are likely to be nonspecific interactors and thus were excluded from the data set in further analyses. All the Y2H interactions identified by the high-throughput screening were subjected to pair-wise Y2H test as quadruplicates with activation tests conducted side by side (Supplementary Fig. 1d). The nonreproducible interactions and those that showed signal in the autoactivation test were removed from the interaction list (Supplementary Fig. 1d).

Mapping protein-complex topology. To identify the direct interactions in protein-complex subunits, we selected the E. coli protein complexes from AP-MS studies2,3 and subjected all the protein complexes to systematic Y2H matrix screening. We used the N- and C-terminal fusion baits and N-terminal prey clones (Supplementary Fig. 4a). The Y2H matrix screening was carried out as quadruplicate to ensure statistical reproducibility of interactions. The screening was performed independently for N- and C-terminal fusion bait vectors (Supplementary Fig. 4b). In brief, the yeast strains expressing a single protein as bait fusion were mated to individual preys of all other subunits in the complex. After mating, the colonies were transferred to the interaction selection medium plates, and the interacting protein pairs were identified by the resulting positive yeast colony. The positive interactions showed clear colony growth on the selection-medium plate, whereas no growth was usually seen for the corresponding negative control, i.e., the bait mated with the empty prey vector strain (Supplementary Fig. 4b).

Validating Y2H interactions. To validate the quality of the Y2H interactions, we randomly selected 114 PPIs (~5%) from the 2,234 PPIs identified in this study. These PPIs were subjected to two independent methods, namely, communoprecipitation in E. coli and LUMIER in mammalian cells16.

Validating interacting protein pairs using communoprecipitation. The E. coli ORFeome entry clone plasmids corresponding to selected interacting protein pairs were selected from the ORFeome plasmid library9. The Entry clones of both bait and prey proteins were cloned into Gateway compatible HaloTag vector pFN18Ag and c-myc tag vector pRH018 (ref. 41). The Halo-tag vector pFN18Ag contains the N-terminal HaloTag (Promega) coding sequence upstream of the Gateway recombination site. The pFN18Ag HaloTag vector was constructed by inserting Gateway cassette into pFN18Ag Halo (Supplementary Note). The pRH018 vector contains sequence encoding the Myc epitope (c-myc) downstream of the Gateway recombination site. Both vectors express the recombinant proteins from the lac promoter.

The interacting pairs and negative control (as pFN18Ag vector plus sequence encoding interacting c-myc fusion protein) plasmid DNA were cotransformed into E. coli BL21 (DE3) Magic cells (Invitrogen). The transformed cells were selected by plating on LB agar plates containing ampicillin (100 µg/ml) and chloramphenicol (35 µg/ml). The transformants were cultured overnight in 2 ml of SOB medium with ampicillin and chloramphenicol; the next day the culture was diluted to OD (optical density at 600 nm) 0.8–0.9 using SOB medium with ampicillin and chloramphenicol. The proteins were induced with IPTG (1 µg/ml) for 5–6 h at 30 °C.

After induction, the cell pellet was dissolved in 250 µl of lysis buffer (50 mM Tris pH 7.4, 100 mM NaCl, 0.5% Triton X-100, protease inhibitors cocktail, 1 mM EDTA and Lysonase (Novagen)) and incubated at room temperature for 20 min, followed by sonication on an ice bath. Lysates were cleared by centrifugation at 14,000 g (4 °C), and the supernatant was collected into a fresh tube. The protein expression and solubility of Halo fusion and c-myc fusion proteins were tested using the TMR ligand or rabbit polyclonal anti-HaloTag (G92851, G9281, Promega) and anti-myc antibodies (Sigma-Aldrich).

The supernatant was used for the purification of protein complexes using HaloLink resin (G9192, Promega). Purified complexes and control lysate samples were separated on Nu-PAGE acrylamide gels (Invitrogen). After transfer of proteins from Nu-PAGE to PVDF membrane, the myc-tagged proteins were detected using standard immunoblotting techniques. The horseradish peroxidase–conjugated secondary anti-rabbit antibody produced in goat (Sigma-Aldrich, A0545) was used for chemiluminescence detection.
Validating interacting protein pairs by LUMIER assay. The *E. coli* ORFeome entry clone plasmids corresponding to selected interacting protein pairs were cloned into Gateway-compatible LUMIER expression vectors, pT-REx-DEST30-ntPra (produces protein A fusions) and pcDNA3-RLuc (produces Renilla luciferase fusions) using an LR reaction (Invitrogen). The expression clones were validated by bacterial colony PCR, and the plasmid DNA of the recombinant proteins were propagated and isolated from *E. coli* (DH5-alpha).

The interacting protein pair plasmids were then cotransfected into HEK (human embryonic kidney) 293T cells, which was followed by LUMIER pull-down assays.

For the negative control, the cells were cotransfected with pT-REx-DEST30 plasmid containing a dimer of protein A and luciferase-tagged recombinant protein.

For each interaction sample, four values were measured in triplicate: the luciferase present in 10% of the sample before washing (‘input’), the luciferase activity present on the beads after washing (‘bound’) and the same values for the negative controls (‘input-nc’ and ‘bound-nc’).

First, the ratio (LR) of the measured luciferase activity after (Lu_bound) and before washing (Lu_input) was calculated for each individual test position:

\[
LR = \frac{Lu_{\text{bound}}}{Lu_{\text{input}}}
\]

The LR values from the triplicates and the corresponding individual negative control measurements were used to calculate a P value by a t-test.

Second, a mean value (\(\bar{LR}\)) was calculated from the LR values for samples and the corresponding negative controls and finally given a luminescence intensity ratio (LIR) between the sample and negative control:

\[
LIR = \frac{\bar{LR}_{\text{sample}}}{\bar{LR}_{\text{nc}}}
\]

We scored an interaction as positive for LIR > 3 with \(P \leq 0.05\). This threshold was chosen based on the published studies. These two values were plotted against each other on the graph in logarithmic scale i.e., (log(LIR) > 0.477, log(P) < −1.30; Fig. 2b).

Validating binary interaction with structural data. First, we compiled a list of structures of *E. coli* protein complexes deposited in the Protein Data Bank (PDB) by taking all the biological units corresponding to PDB entries containing structures of *E. coli* proteins. For those entries with more than one annotated biological unit, we compiled all the available units with more than one protein. We then took every biological unit and analyzed the contacts between pairs of chains. Two chains were classified as in ‘physical direct interaction’ if they had at least five residue-residue contacts among: (i) covalent interactions (disulfide bridges), defined as two sulfur atoms of a pair of cysteines at a distance ≤ 2.56 Å (two times the covalent radius of sulfur plus 0.5 Å); (ii) hydrogen bonds, defined as all atom pairs N-O and O-N at a distance ≤ 3.5 Å; (iii) salt bridges, defined as all atom pairs N-O and O-N at a distance ≤ 5.5 Å; and (iv) van der Waals interactions, defined as all pairs of carbon atoms at a distance ≤ 5.0 Å. We built several data sets of interactions to be evaluated against direct physical interactions inside complexes (Supplementary Results).

Data set of literature binary interactions for the internal topology of complexes. The data set of literature binary interactions used in the reconstruction of the internal topology of complexes was obtained by merging the literature binary data set described in validating binary interactions with structural data (Supplementary Table 11), with all the binary interactions detected by X-ray crystallography and NMR spectroscopy (Supplementary Table 12). We compiled these data set by measuring the ‘socio-affinity’ index of all protein pairs in the complexes detected in refs. 2, 3 as described in ref. 42, and selecting all the interactions with an index greater than 10.

Structural characterization of the complex components and internal interactions. Experimental structures and homology models for the complexes were obtained by submitting the set of binary complex interactions to the Interactome3D webserver (http://interactome3d.irbbarcelona.org/).

Compilation of data sets for assessing the properties of Y2H and AP-MS data sets. The biological attributes of Y2H and AP-MS networks were analyzed by extracting the following data sets: (i) coexpressed genes from microarray-based mRNA transcript profiles downloaded from the M3D database were compiled essentially as previously described; (ii) chemical phenotypic data were obtained from a recent study, in which a library of *E. coli* strains either deleted for nonessential components or hypomorphic (i.e., partial loss of gene function) alleles of essential genes was screened against hundreds of conditions to study gene essentiality; (iii) high-confidence genetic interaction data sets downloaded from the recently published large-scale epistatic studies and (iv) GO terms for *E. coli* downloaded from the UniProt database. Shortest path length between all possible gene pairs in both combined Y2H and AP-MS networks was computed using a Python-based library, NetworkX.

Conserved protein-protein interactions (interologs). We collected the known binary PPIs reported for all bacterial species from the literature-curated database MPIDB and from the large-scale bacterial interactome studies of *Treponema pallidum, Campylobacter jejuni, Helicobacter pylori, Mycobacterium tuberculosis, Rickettsia sibirica, Streptococcus pneumoniae, Synecocystis and Mesorhizobium loti*. Orthologs have been identified between all proteins in these interactions, by using the stand alone megablast program with a set cut-off of \(E < 10^{-10}\). This resulted in 693 interologs (Supplementary Table 13).

Similarly, we analyzed conservation of *E. coli* protein complex1 across 20 bacterial species. The conservation of a protein complex was calculated as the percentage of the orthologous proteins present in each species (Supplementary Fig. 9).

To determine the significance of the differences of conserved interactions that appear within and between complexes, we first calculated the ratio of the corresponding percentages in each organism. In particular, we determined the observed ratio \(R = P_a/P_b\) where \(P_a\) is the percentage of conserved protein interactions within complexes, and \(P_b\) is the percentage of conserved interactions between complexes. As null model we randomly sampled orthologous sets of proteins 10,000 times and determined the corresponding expected ratios of percentages of conserved interactions within and between complexes, \(R = P_{a,W}/P_{b,W}\). Finally, we calculated empirical \(P\) values as the ratio \(P_a = (N(R_a > R))/10,000\), where \(N\) is the number of times when the expected, random ratio \(R\) exceeded the observed ratio \(R\).

Construction of phylogenetic tree. To analyze a phylogenetic kinship of 20 bacterial species, we collected the corresponding complete proteomes of *S. enterica, V. cholera, P. aeruginosa, Acinetobacter, H. influenza, B. mellitensis, C. crescentus, S. coelicolor, B. anthracis, B. subtilis, Y. pestis, N. meningitidis, M. tuberculosis, S. aureus, F. tularensis, L. lactis, H. pylori, S. pneumoniae, R. prowazekii, and M. genitalium* from the US National Center for Biotechnology Information database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). We compared the proteomes of these bacteria using the CVTree program and visualized the corresponding phylogenetic tree by the T-REX program. This approach accounted for all protein sequences of an organism’s proteome and counted the number of overlapping k-tuples to form a raw compositional vector with 20k components. Random background frequencies were subtracted by predicting the number of k-tuples from \((k−1)\)-mers and \((k−2)\)-mers through a simple Markovian model. By putting these ‘normalized’ frequencies in a fixed order a normalized 20k dimensional composition vector for each organism was obtained. Finally, a correlation between two species was determined by calculating the projection of one normalized vector on the other. By calculating a normalized distance between organisms, a distance matrix that allowed the construction of a phylogenetic tree was obtained.

Determination of orthologous sequences. By using all-versus-all BLASTP searches with the InParanoid script for proteins from the sets of two species, sequence pairs with mutually best scores were selected as central orthologous pairs. Proteins from both species that showed such an elevated degree of homology were clustered around these central pairs, forming orthologous groups. The quality of the clustering was further assessed by a standard bootstrap procedure. We only considered the central orthologous sequence pair with a confidence level of 100% as the real orthologous relationship.

NATURE BIOTECHNOLOGY

doi:10.1038/nbt.2831
41. Hallez, R., Letesson, J.J., Vandenhaute, J. & De Bolle, X. Gateway-based destination vectors for functional analyses of bacterial ORFeomes: application to the Min system in Brucella abortus. Appl. Environ. Microbiol. 73, 1375–1379 (2007).
42. Gavin, A.C. et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631–636 (2006).
43. Faith, J.J. et al. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res. 36, D866–D870 (2008).
44. Nichols, R.J. et al. Phenotypic landscape of a bacterial cell. Cell 144, 143–156 (2011).
45. Hagberg, A.A. et al. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference 11–15 (2008).
46. Xu, Z. & Hao, B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res. 37, W174–W178 (2009).
47. Alix, B., Boubacar, D.A. & Vladimir, M.T.-R.E.X. a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res. 40, W573–W579 (2012).
48. Remm, M., Storm, C.E. & Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314, 1041–1052 (2001).