Network Context and Selection in the Evolution to Enzyme Specificity

Hojung Nam,[1]* Nathan E. Lewis,[1,3]*‡ Joshua A. Lerman,[2] Dae-Hee Lee,[1]† Roger L. Chang,[2] Donghyuk Kim,[1] Bernhard O. Palsson[1]‡

Enzymes are thought to have evolved highly specific catalytic activities from promiscuous ancestral proteins. By analyzing a genome-scale model of *Escherichia coli* metabolism, we found that 37% of its enzymes act on a variety of substrates and catalyze 65% of the known metabolic reactions. However, it is not apparent why these generalist enzymes remain. Here, we show that there are marked differences between generalist enzymes and specialist enzymes, known to catalyze a single chemical reaction on one particular substrate in vivo. Specialist enzymes (i) are frequently essential, (ii) maintain higher metabolic flux, and (iii) require more regulation of enzyme activity to control metabolic flux in dynamic environments than do generalist enzymes. Furthermore, these properties are conserved in Archaea and Eukarya. Thus, the metabolic network context and environmental conditions influence enzyme evolution toward high specificity.

Ancestral enzymes are proposed to have exhibited broad substrate specificity and low catalytic efficiency (*1*). Through mutation, duplication, and horizontal gene transfer, gene families diversified and promiscuous enzymes apparently were refined to exhibit specific and more efficient catalytic abilities (*2*, *3*). Thus, today's metabolic enzymes are commonly assumed to be "specialists," having evolved to catalyze one reaction on a unique primary substrate in an organism. However, some enzymes are "generalists" that promiscuously catalyze reactions on a variety of substrates in vivo (*2*) or exhibit multifunctionality by catalyzing multiple classes of reactions, often at different active sites (*4*). Thus, a fundamental question arises: Why do some enzymes evolve to become specialists, whereas others retain generalist characteristics? By analyzing enzyme functions and properties in experimental data and in silico metabolic network models, we show that the in vivo biochemical network context in which an enzyme resides may influence the evolution of enzyme specificity.

How many metabolic enzymes are generalists? To answer this question, we used a comprehensive reconstruction of the *Escherichia coli* K-12 MG1655 metabolic network, which accounts for the metabolic functions of 1260 gene products (28% of the predicted and experimentally validated open reading frames in *E. coli*) (*5*), which contribute to 1081 enzyme complexes analyzed in this study. In the reconstruction, we define a reaction as a unique set of substrates that are chemically transformed into a unique set of products. With this definition, we classified 677 enzymes as specialists because they catalyze one unique reaction and 404 as generalists because they catalyze multiple reactions. Thus, we estimate that 37% of metabolic enzymes in *E. coli* are generalists, most of which exhibit substrate promiscuity (fig. S1A). Furthermore, specialist and generalist enzymes catalyze 454 and 859 metabolic reactions, respectively, distributed across many metabolic subsystems (Fig. 1, A and B). Thus, contrary to the textbook view of enzymes as "specific catalysts," generalist enzymes have a prominent role in *E. coli*, catalyzing at least 65% of the nonspontaneous metabolic reactions.

We performed several network-wide analyses to provide additional support for our estimates and the classification. First, we found that almost all genes in the network have been well characterized and studied in more than 61,727 published studies (fig. S1D). Second, we found no correlation between our classification and knowledge depth, i.e., neither specialist nor generalist enzymes had been studied in more depth (fig. S1E). Third, our generalist enzymes did not likely include many latent promiscuous reactions measured in vitro that likely do not occur in vivo, because 85% of the generalist enzymes reactions (GERxns) were active in silico in common growth conditions. This is the same percentage seen for specialist enzyme reactions (SERxns) (fig. S2). Fourth, because enzyme classification may vary with further study, we tested the sensitivity of the results presented in this work. We found the results to be qualitatively robust with improvements in the metabolic network from the discovery of new enzymes, variations in enzyme classification, and the exclusion of promiscuous enzymes or multifunctional enzymes from the generalist class (fig. S3). Although transporter reactions were not included in the groups of SERxns or

[1] Department of Bioengineering, University of California San Diego, La Jolla, CA 92093–0412, USA. [2] Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093–0412, USA. [3] Wyss Institute for Biologically Inspired Engineering and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

*These authors contributed equally to this work.

†Present address: Systems and Synthetic Biology Research Center, Korea Research Institute of Bioscience and Biotechnology, 125 Gwahak-ro, Yuseong-gu, Daejeon 305-806, Korea.

‡To whom correspondence should be addressed. E-mail: nlewis@genetics.med.harvard.edu (N.E.L.); palsson@ucsd.edu (B.O.P.)

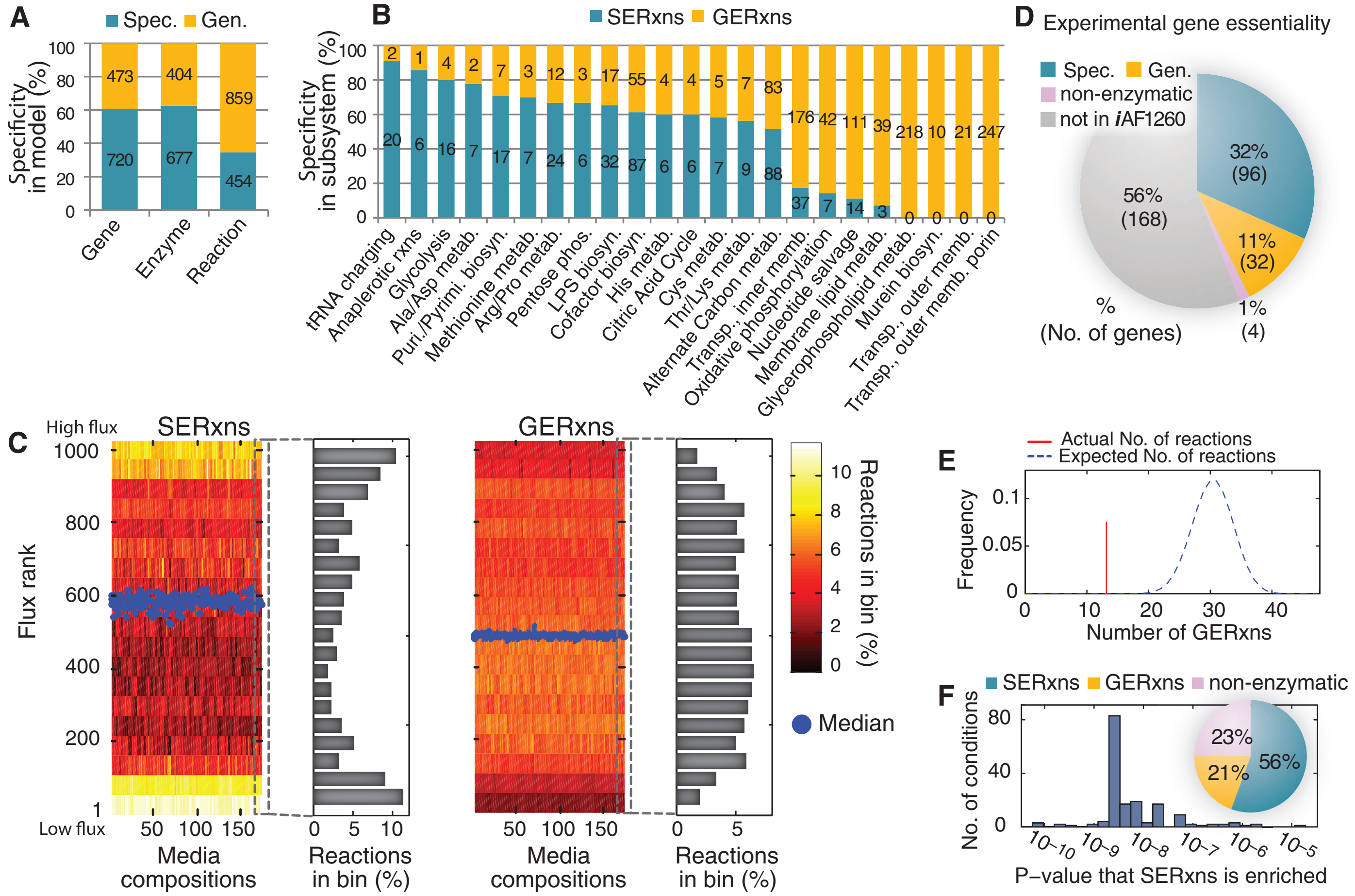

Fig. 1. (**A**) Specialist and generalist genes and proteins and their associated reactions were enumerated in *E. coli* metabolism. (**B**) Several metabolic subsystems were enriched in specialist enzyme reactions (SERxns) or generalist enzyme reactions (GERxns) in *E. coli* (hypergeometric $P \leq 0.05$). (**C**) Reaction flux magnitudes were rank-ordered and binned in histograms for each unique media condition. A heat map was used to visualize histograms for all 174 media conditions (columns) with each row representing bins spanning the given flux rank ranges (*y* axis). Color intensity shows histogram bin height, corresponding to the percentage of reactions in the bin. Example histograms (shown on the right) provide for one representative condition. SERxns tend to have a higher flux, but low-flux SERxns are enriched in enzymes that synthesize low-abundance essential cell components, such as cofactors and prosthetic groups (fig. S4C). (**D**) Genes for specialist enzymes are more frequently essential in vivo. (**E**) In silico, few essential GERxns were identified for growth on glucose minimal medium. (**F**) For all 174 simulated growth conditions, SERxns are significantly enriched among in silico–predicted reactions essential for growth, representing 56% of the essential reactions (inset).

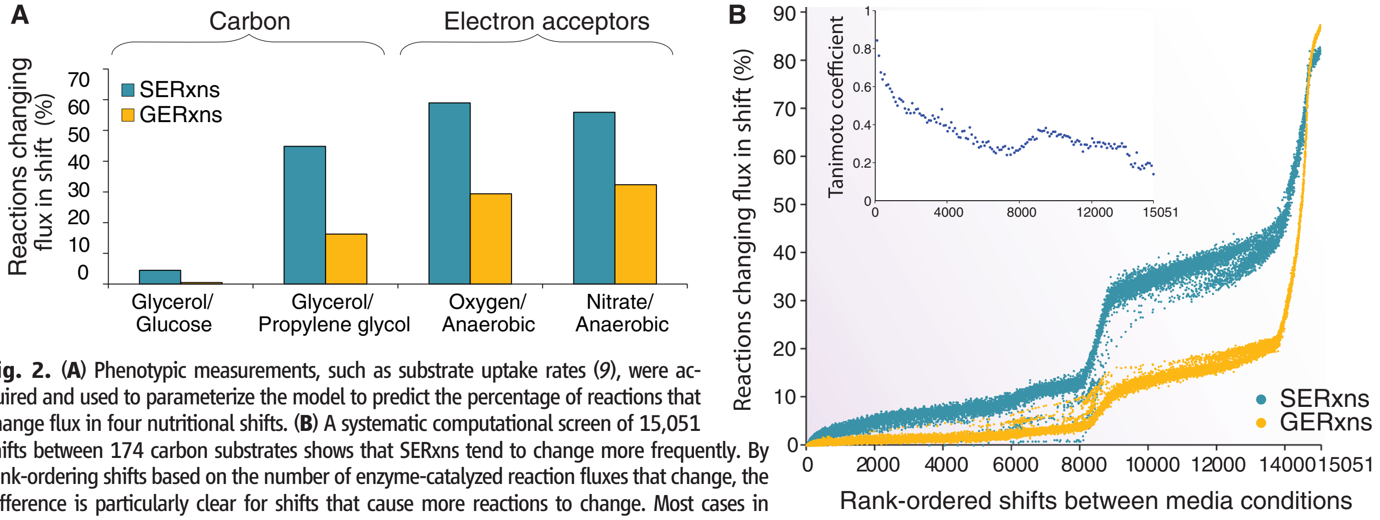

Fig. 2. (**A**) Phenotypic measurements, such as substrate uptake rates (*9*), were acquired and used to parameterize the model to predict the percentage of reactions that change flux in four nutritional shifts. (**B**) A systematic computational screen of 15,051 shifts between 174 carbon substrates shows that SERxns tend to change more frequently. By rank-ordering shifts based on the number of enzyme-catalyzed reaction fluxes that change, the difference is particularly clear for shifts that cause more reactions to change. Most cases in which there is only a weak difference involve shifts between two similar primary carbon susbstrates, as measured by their Tanimoto coefficients (inset; Tanimoto coefficients are averaged across sets of 100 shifts).

GERxns, their inclusion would not qualitatively change the results in this work (fig. S3). Thus, the classification and results from our subsequent analysis are robust.

Why are so many generalist enzymes evolutionarily retained, whereas others became specialists? Demands for higher metabolic flux may provide an evolutionary selective pressure to enhance an enzyme's catalytic rate and reduce the required enzyme concentration. However, catalytic improvements for one substrate of a generalist enzyme can suppress other catalytic activities (*6*). To determine if specialists maintain higher flux, we estimated the steady-state metabolic flux rates (*7*) for all *E. coli* enzymes using a genome-scale metabolic network model. We employed a Markov chain Monte Carlo sampling method (*8*) to simulate flux on 174 media conditions with different nutrient compositions (*9*). For each growth condition, the median flux for each reaction was rank-ordered to determine the relative flux among reactions.

Across all simulated growth conditions, SERxns maintained higher flux than GERxns (Fig. 1C and fig. S4). Gene duplications may have been fixed in the population when specialization occurred to increase activity of high-flux enzymes. Higher activity would permit lower enzyme concentrations, thereby offsetting the cost of duplication (*10*). Consistent with this reasoning, k_{cat} values are significantly higher for high-flux specialist enzymes than for all other enzymes (fig. S5C, Wilcoxon $P = 2.8 \times 10^{-7}$).

Although flux level may contribute to enzyme specialization, gene essentiality may also contribute. High substrate affinity for essential enzymes could mitigate substrate competition in the synthesis of necessary biomass components, irrespective of flux level. Consistent with this hypothesis, we found that essential enzymes have lower K_m values and therefore higher substrate affinity (fig. S5F, Wilcoxon $P = 1.1 \times 10^{-11}$). Furthermore, specialist enzymes are enriched among experimentally determined essential genes (*11*) (hypergeometric $P = 8.65 \times 10^{-5}$, Fig. 1D). In silico simulation also demonstrated that cell growth rarely directly depends on flux through generalist enzymes (Fig. 1E), whereas many SERxns were essential for growth across all 174 tested media conditions (Fig. 1F and fig. S6).

Gene essentiality (*12*, *13*) and reaction fluxes often vary (*8*, *14*, *15*) because natural environments are dynamic and nutrient concentrations fluctuate in the microbial microenvironment (*16*). The need to regulate reaction flux in dynamic environments could induce gene duplication and enzyme specialization to simplify the combinatorial complexity of regulating multiple reactions on a single enzyme (e.g., see serine hydroxymethyltransferase in fig. S7). To test this hypothesis, we identified enzymes that will require more metabolic regulation in dynamic environments by simulating changes in carbon source and electron acceptors for *E. coli*. For each substrate shift, the model predicted whether reaction flux should increase or decrease, and these predictions were consistent with measured differential gene expression (fig. S8) (*17*).

Across all shifts in growth media, there was a considerable difference in the percentages of active SERxns and GERxns that significantly changed their flux between growth conditions (Fig. 2A). SERxn fluxes were often more than twice as likely to change than GERxn fluxes. Thus, flux through SERxns is considerably more sensitive to environmental change, whereas GERxn fluxes vary less. To examine if this is a general property, we simulated 15,051 pairwise environmental shifts. In 96% of these shifts, SERxns changed more frequently than GERxns (Fig. 2B). This difference was strongest for environmental shifts that cause more than 8% of the reactions to change flux (fig. S9). Because SERxns are subject to greater flux changes in nutritionally dynamic environments, it seems that duplication may have occurred to allow more focused regulation of fluxes. This duplication would be reinforced as the enzymes enhance their catalytic specificity.

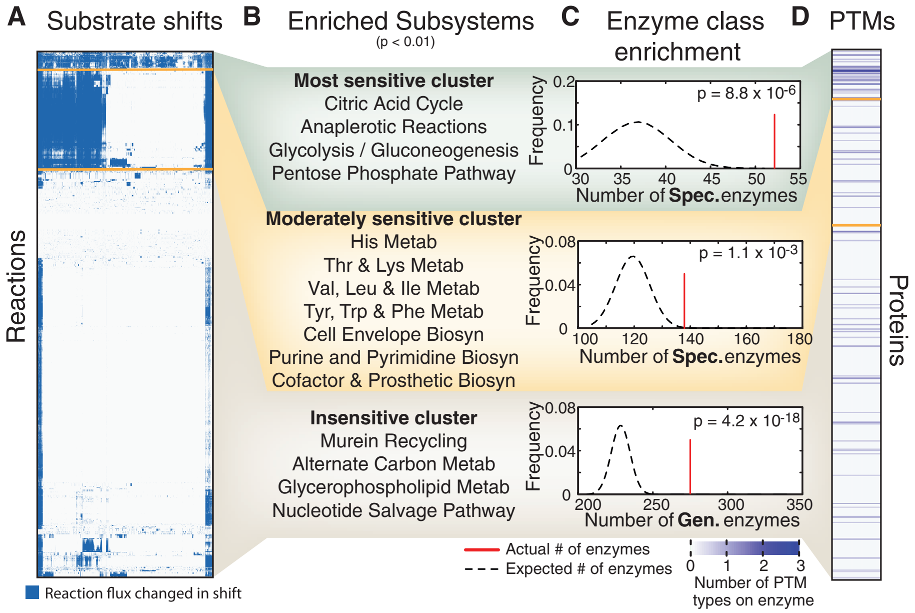

Fig. 3. (**A**) Clustering reactions that change (blue) or do not change (white) across 15,051 different media shifts (*x* axis) yields three distinct groups, (**B**) which are each enriched in unique metabolic subsystems. (**C**) Specialist enzymes are enriched in more sensitive clusters, whereas generalist enzymes are enriched in the cluster with few flux changes. (**D**) The number of PTMs (acetylation, phosphorylation, and/or succinylation) on enzymes increases with sensitivity of clusters.

In dynamic environments, metabolic flux can be regulated through metabolite-protein interactions or posttranslational modifications (PTMs) (*18*, *19*). We quantified the association of metabolic regulation with enzyme specificity, using a few hundred metabolite-mediated regulatory interactions obtained from the EcoCyc database and enzyme PTMs from mass spectrometry studies in *E. coli* (*9*). Allosteric, uncompetitive, and noncompetitive regulatory interactions were enriched among specialists (hypergeometric $P = 9 \times 10^{-4}$), as were PTMs (hypergeometric $P = 5 \times 10^{-3}$). Metabolic regulation was less prevalent among generalists, consistent with the decreased need to change flux through their reactions in dynamic environments. Moreover, fluxes for reactions catalyzed by the same generalist often covary, thereby reducing requirements for more complex regulation (fig. S10).

To further assess the association of specificity with regulation, we quantified how frequently each reaction changed flux across all simulated 15,051 media shifts. K-means clustering elucidated three dominant reaction clusters (Fig. 3A). Two clusters show frequent changes in flux, and these were enriched in specialists, particularly those associated with central and amino acid metabolism (Fig. 3B). The reaction cluster with few changes in flux was significantly enriched in generalists (Fig. 3C). PTMs and small-molecule–mediated allosteric regulation were enriched within the cluster that experienced the most change in flux (hypergeometric $P = 5 \times 10^{-3}$), but depleted from the cluster dominated by generalists (hypergeometric $P = 3 \times 10^{-3}$; Fig. 3D and fig. S11). Thus, enzymes that exhibit more extensive metabolic regulation tend to have evolved increased enzyme specificity.

The aforementioned properties show how enzyme specificity correlates with holistic functions of the *E. coli* metabolic network. However, these properties should be conserved if they influence selection of enzyme specificity in protein evolution. Thus, we examined their conservation using genome-scale metabolic models of microbes from the other domains of life, including the archeon *Methanosarcina barkeri* (*20*) and the eukaryotes *Saccharomyces cerevisiae* (*21*) and *Chlamydomonas reinhardtii* (*22*). Similar to *E. coli*, the three organisms contain numerous generalist enzymes. Common growth conditions were simulated for each organism to estimate metabolic flux. In each organism, specialist enzymes maintained a higher flux on average than generalist enzymes. Moreover, when environmental shifts were simulated for each organism, generalist enzymes were less likely to change flux between growth conditions (fig. S12). Even as microbes diversified, high flux and a need for focused regulation in varying environments remained as general features of specialist enzymes.

It is generally believed that highly promiscuous ancestral enzymes eventually evolved to become specific and highly efficient (*1*). However, many current enzymes are only moderately efficient (*23*), and there are numerous generalists. Thus, evolution has not converged to a point where metabolic enzymes are all specialists. Our results suggest that this convergence has been hindered in part by the lower essentiality, smaller flux, and reduced regulatory requirements of generalist enzymes, including those that are multifunctional and those exhibiting substrate promiscuity (figs. S3B and S4C). The specialization of these enzymes may not provide adequate fitness advantages to offset the cost of gene duplication and maintenance (*10*) that accompanies the separation of catalytic functions into several specialists. In addition, these selective pressures may not influence some classes of enzymes if their generalist activities are desirable, such as in the degradation and clearance of diverse toxins (*24*) or the synthesis of structural lipids or glycoconjugates. However, our results suggest that many metabolic enzymes will specialize when an environmental change elicits a fitness challenge that causes a generalist to contribute to the high-flux (*8*) or essential biomass-producing core (*25*) of metabolism, or if new environmental fluctuations require more focused regulation of flux. Preliminary analysis suggests that potential examples of this divergence include serine hydroxymethyltransferase and its isozyme LtaE (fig. S7) or pyruvate formate lyase and TdcE (see supplementary materials).

Our results demonstrate that the metabolic network, as a whole, supports organismal survival and influences cell physiology in a given environment. By analyzing the functions of its pathways and using biomolecular networks to integrate many disparate data types into a coherent whole, we show that systems biology allows the elucidation of selection pressures that are not apparent at the level of a single enzyme (*26*–*29*).

References and Notes

1. R. A. Jensen, *Annu. Rev. Microbiol.* **30**, 409 (1976).
2. O. Khersonsky, D. S. Tawfik, *Annu. Rev. Biochem.* **79**, 471 (2010).
3. H. Innan, F. Kondrashov, *Nat. Rev. Genet.* **11**, 97 (2010).
4. O. Khersonsky, S. Malitsky, I. Rogachev, D. S. Tawfik, *Biochemistry* **50**, 2683 (2011).
5. A. M. Feist *et al.*, *Mol. Syst. Biol.* **3**, 121 (2007).
6. A. Aharoni *et al.*, *Nat. Genet.* **37**, 73 (2005).
7. N. E. Lewis, H. Nagarajan, B. O. Palsson, *Nat. Rev. Microbiol.* **10**, 291 (2012).
8. E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, A. L. Barabási, *Nature* **427**, 839 (2004).
9. Materials and methods are available as supplementary materials on *Science* Online.
10. A. Wagner, *J. Exp. Zool. B Mol. Dev. Evol.* **308B**, 322 (2007).
11. T. Baba *et al.*, *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
12. B. Papp, C. Pál, L. D. Hurst, *Nature* **429**, 661 (2004).
13. D. Deutscher, I. Meilijson, M. Kupiec, E. Ruppin, *Nat. Genet.* **38**, 993 (2006).
14. S. Bordel, R. Agren, J. Nielsen, *PLOS Comput. Biol.* **6**, e1000859 (2010).
15. R. Schuetz, L. Kuepfer, U. Sauer, *Mol. Syst. Biol.* **3**, 119 (2007).
16. E. Gur, D. Biran, E. Z. Ron, *Nat. Rev. Microbiol.* **9**, 839 (2011).
17. N. E. Lewis, B. K. Cho, E. M. Knight, B. O. Palsson, *J. Bacteriol.* **191**, 3437 (2009).
18. Z. Zhang *et al.*, *Nat. Chem. Biol.* **7**, 58 (2011).
19. L. Gerosa, U. Sauer, *Curr. Opin. Biotechnol.* **22**, 566 (2011).
20. A. M. Feist, J. C. Scholten, B. O. Palsson, F. J. Brockman, T. Ideker, *Mol. Syst. Biol.* **2**, 2006 0004 (2006).
21. M. L. Mo, B. O. Palsson, M. J. Herrgård, *BMC Syst. Biol.* **3**, 37 (2009).
22. R. L. Chang *et al.*, *Mol. Syst. Biol.* **7**, 518 (2011).
23. A. Bar-Even *et al.*, *Biochemistry* **50**, 4402 (2011).
24. M. Morar, G. D. Wright, *Annu. Rev. Genet.* **44**, 25 (2010).
25. E. Almaas, Z. N. Oltvai, A. L. Barabási, *PLOS Comput. Biol.* **1**, e68 (2005).
26. S. D. Copley, *J. Biol. Chem.* **287**, 3 (2012).
27. B. Papp, R. A. Notebaart, C. Pál, *Nat. Rev. Genet.* **12**, 591 (2011).
28. H. Nam, T. M. Conrad, N. E. Lewis, *Curr. Opin. Biotechnol.* **22**, 595 (2011).
29. P. Carbonell, G. Lecointre, J. L. Faulon, *J. Biol. Chem.* **286**, 43994 (2011).

Acknowledgments: We thank D. Zielinski for insightful discussion on this work. This work was supported by NIH, NSF, and U.S. Department of Energy grants 2R01GM057089-13, NSF GK-12 742551, DE-SC0004917, and DE-FG02-09ER25917. Data are available at the NCBI Gene Expression Omnibus (GEO) database (GSE34631).

Supplementary Materials

www.sciencemag.org/cgi/content/full/337/6098/1101/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S17
Tables S1 and S2
References (*30*–*78*)
Database S1

21 November 2011; accepted 29 June 2012
10.1126/science.1216861