Title

Mutant phenotypes for thousands of bacterial genes of unknown function

Journal

Nature, 557(7706)

ISSN

0028-0836

Authors

Price, Morgan N

Wetmore, Kelly M

Waters, R Jordan

et al.

Publication Date

2018-05-24

DOI

10.1038/s41586-018-0124-0

Peer reviewed

Mutant Phenotypes for Thousands of Bacterial Genes of Unknown Function

Morgan N. Price[1], Kelly M. Wetmore[1], R. Jordan Waters[2], Mark Callaghan[1], Jayashree Ray[1], Hualan Liu[1], Jennifer V. Kuehl[1], Ryan A. Melnyk[1], Jacob S. Lamson[1], Yumi Suh[1], Hans K. Carlson[1], Zuelma Esquivel[1], Harini Sadeeshkumar[1], Romy Chakraborty[3], Grant M. Zane[4], Benjamin E. Rubin[5], Judy D. Wall[4], Axel Visel[2,6], James Bristow[2], Matthew J. Blow[2,*], Adam P. Arkin[1,7,*], Adam M. Deutschbauer[1,8,*]

[1]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory
[2]Joint Genome Institute, Lawrence Berkeley National Laboratory
[3]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory
[4]Department of Biochemistry, University of Missouri
[5]Division of Biological Sciences, University of California, San Diego
[6]School of Natural Sciences, University of California, Merced
[7]Department of Bioengineering, University of California, Berkeley
[8]Department of Plant and Microbial Biology, University of California, Berkeley
[*]To whom correspondence should be addressed:
MJB (MJBlow@lbl.gov)
APA (APArkin@lbl.gov)
AMD (AMDeutschbauer@lbl.gov)

Website for interactive analysis of mutant fitness data:
http://fit.genomics.lbl.gov/

Website with supplementary information and bulk data downloads:
http://genomics.lbl.gov/supplemental/bigfit/

Summary

One third of all protein-coding genes from bacterial genomes cannot be annotated with a function. To investigate these genes' functions, here we collected genome-wide mutant fitness data from 32 diverse bacteria across dozens of growth conditions each. We identified mutant phenotypes for 11,779 protein-coding genes that had not been annotated with a specific function. Many genes could be associated with a specific condition because the gene affected fitness only in that condition, or with another gene in the same bacterium because they had similar mutant phenotypes. 2,316 of these poorly-annotated genes had associations that are of high confidence because they are conserved in other bacteria. By combining these conserved associations with comparative genomics, we identified putative DNA repair proteins and we proposed specific functions for poorly-annotated enzymes and transporters and for uncharacterized protein families. Our study demonstrates the scalability of microbial genetics and its utility for improving gene annotations.

Background

Thousands of bacterial genomes have been sequenced, revealing the predicted amino acid sequences of millions of proteins. Only a small proportion of these proteins have been studied experimentally and most proteins' functions are predicted via their similarity to experimentally characterized proteins. However, about one third of bacterial proteins are not similar enough to any characterized protein to be annotated by this approach [1]. Furthermore, these predictions are often incorrect as homologous proteins may have different substrate specificities [2]. This sequence-to-function gap represents a growing challenge for microbiology, because new bacterial genomes are being sequenced at an ever-increasing rate, while experimental protein characterization continues to be relatively slow [1].

One approach for investigating an unknown protein's function is to assess the consequences of a loss-of-function mutation of the corresponding gene under multiple conditions [3-6]. The mutant phenotypes can be combined with comparative genomics to provide evidence-based annotations for a fraction of the proteins [3,4]. Transposon mutagenesis followed by sequencing (TnSeq) measures mutant phenotypes genome-wide from a single experiment in which tens of thousands of different mutants are growth together [7,8]. Coupling TnSeq with random DNA barcoding of each mutant (RB-TnSeq) makes it easier to measure phenotypes across many conditions [9]. Here, we use RB-TnSeq to address the sequence-to-function gap by systematically exploring the mutant phenotypes of thousands of genes from each of 32 bacteria under multiple experimental conditions (**Fig. 1a**).

Results

Mutant fitness compendia for 32 bacteria

To perform a systematic assessment of mutant phenotypes across a diverse set of bacteria, we studied 32 genetically tractable bacteria representing six different bacterial divisions and 23 different genera. In addition to 30 aerobic heterotrophs, we also studied a strictly anaerobic sulfate-reducing bacterium (*Desulfovibrio vulgaris*) and a strictly photosynthetic cyanobacterium

(*Synechococcus elongatus*) (**Fig. 1b**). We generated a randomly barcoded transposon mutant library in each of the 32 bacteria, ten of which were previously described [9-12]. For each mutant population, we used TnSeq to generate genome-wide maps of transposon insertion locations. Genes that have very few or no transposon insertions are likely to be essential for viability, or nearly so, in the conditions that were used to select the mutants (**Supplementary Note 1)**. We identified 289 to 614 likely-essential protein-coding genes per bacterium (**Fig. 1b; Supplementary Table 1**).

To identify conditions for mutant fitness profiling, we tested the growth of the 30 aerobic heterotrophic bacteria in a range of conditions, including the utilization of 94 different carbon sources and 45 different nitrogen sources, and their inhibition by 34 to 55 stress compounds including antibiotics and metals (**Supplementary Tables 2-4**). In the typical mutant fitness experiment, we grew a pool of mutants for 4-8 generations and used DNA barcode sequencing [13] to compare the abundance of the mutants before and after growth (**Fig. 1a**). We defined gene fitness to be the $\log_2$ change in abundance of mutants in that gene during the experiment (**Fig. 1a**). For example, a gene fitness value of -2 means that the strains with transposon mutant insertions in that gene dropped to 25% of their initial relative abundance by the end of the experiment, while a fitness value of 0 means that their relative abundance was unchanged. Genes with insufficient coverage were excluded from our dataset ("no data" in **Fig. 1b**) and only protein-coding genes were considered. Including all replicates, we conducted a total of 4,870 genome-wide fitness experiments that met our criteria for biological and internal consistency [9], representing 26 to 129 different experimental conditions for each bacterium (**Fig. 1b, Supplementary Table 5**).

To illustrate the consistency of our data with known protein functions, we examined fitness data for the three most common classes of experiments: carbon utilization, nitrogen utilization, and stress. For the utilization of D-fructose or 4-hydroxybenzoate as the sole source of carbon in *Cupriavidus basilensis*, the fitness data identifies expected proteins for the catabolism of each substrate (**Fig. 1c**). Similarly, the fitness data identified the key enzymes and transporters required for the utilization of D-alanine or cytosine as the sole nitrogen source in *Azospirillum brasilense* (**Extended Data Fig. 1a**). Lastly, in *Shewanella loihica*, orthologs of the CzcCBA heavy metal efflux pump [14] and the zinc responsive regulator ZntR were important for fitness in the presence of an elevated concentration of zinc (**Extended Data Fig. 1b**). For each condition, we also identified proteins that were previously not known to be involved in the respective processes, including an efflux pump important for 4-hydroxybenzoate utilization by *C. basilensis* (**Fig. 1c**, **Supplementary Table 6**).

We computed a *t*-like test statistic for each gene fitness value [9] and identified a statistically significant mutant phenotype in at least one condition for 30% of the genes for which we collected fitness data (**Fig. 1b**). 18% of all genes with fitness measurements were significantly detrimental to fitness (fitness > 0) in at least one condition (**Extended Data Fig. 2**), which is consistent with previous reports that many genes are detrimental in some conditions [3,15]. Genes annotated with a TIGRFAM [16] functional role (TIGR role) were particularly likely to have statistically significant phenotypes, with more than half of those with fitness data (52%) showing a significant phenotype (**Fig. 1d**). In contrast, genes with vague annotations that are not specific (i.e. "transporter") or with functionally-uninformative annotations (i.e. "hypothetical protein" or "membrane protein") were less likely to have phenotypes (28% or 20%, respectively).

Nevertheless, our assays identified phenotypes for 11,779 genes that are not annotated with a detailed function (**Fig. 1d**), including 4,135 genes that encode proteins that do not belong to any characterized family in either Pfam or TIGRFAMs [16,17].

Conserved functional associations

To gain insight into the biological function of individual proteins from the mutant fitness data, we used two strategies: (1) Identification of "specific" phenotypes that are observed only under one or a small number of conditions in a given bacterium; (2) "Cofitness" patterns, where multiple genes in a bacterium show similar fitness profiles across all conditions. Furthermore, we identified conserved specific phenotypes and conserved cofitness by comparing the data from 32 bacteria, and we tested the reliability of these conserved associations for understanding protein function.

To assign specific phenotypes, we identified genes that had |fitness| > 1 and $|t|$ > 5 in an experiment but had little phenotype in most other experiments. For example, the fluoride efflux protein CrcB [18] is specifically important for fitness under elevated fluoride stress in 8 bacteria, but not for fitness in any of the hundreds of other experimental conditions that we tested (**Fig. 2a** shows data from 5 bacteria). Among all genes with a significant phenotype under any condition, 33% have a specific phenotype. We considered a specific phenotype to be conserved if a similar protein in another bacterium (a putative ortholog) had a similar phenotype.

Since catabolic processes in *E. coli* are well understood, we compared the specific-important phenotypes (fitness < 0) on carbon and nitrogen sources to the EcoCyc database [19]. We found that many of these gene-condition associations resulted from the gene's direct involvement in the uptake or catabolism of the compound or in the regulation of those processes (**Supplementary Table 7**). If the specific phenotype was conserved, then the association was much more likely to be direct (**Fig. 2b**; $P < 10^{-4}$, Fisher's exact test).

We identified specific phenotypes and conserved specific phenotypes for genes of all annotation classes (**Fig. 2c**). In particular, specific phenotypes linked 3,927 genes with vague or hypothetical annotations to over 100 conditions, including 82 carbon sources, 43 nitrogen sources, and 54 stresses.

Our second strategy for gaining insight into a protein's function was based on the observation that genes with related functions often have similar fitness patterns across multiple conditions, which we term "cofitness". Cofitness is the Pearson (linear) correlation of all of the fitness values for a pair of genes in the same bacterium, and cofitness across dozens of conditions has already been shown to be useful for understanding protein function [3-5,20]. For example, Npr and PtsP of the nitrogen phosphotransferase system (PTS) in *E. coli* were cofit (**Fig. 2d**), probably because PtsP phosphorylates and activates Npr [21]. We defined conserved cofitness as a pair of orthologous genes that had high cofitness in more than one bacterium, regardless of the conditions. For example, the orthologs of Npr and PtsP (SO1332 and SO3965) also have high cofitness in *S. oneidensis*, but they have phenotypes in different conditions than in *E. coli* (**Fig. 2d**). Phenotypic variability among orthologous genes has been reported before [22].

To test how accurately cofitness or conserved cofitness linked together functionally-related genes, we determined for each query gene whether its functional role (TIGR subrole [16]) could be accurately predicted by its cofit genes' roles. We found that high-scoring cofitness in a single bacterium lead to predictions of TIGR subroles that were mostly correct, but the accuracy decayed rapidly as the cofitness score decreased (**Fig. 2e**). In contrast, for conserved cofitness, the decay was much slower (**Fig. 2e**). Furthermore, conserved cofitness was significantly more accurate for a given number of predictions: for example, the top 2,000 predictions from cofitness (r > 0.86 for gene pairs from one bacterium) had 63% agreement in TIGR subroles, while the top 2,000 predictions from conserved cofitness (r > 0.66 for gene pairs from both bacteria) had 74% agreement ($P < 10^{-12}$, Fisher's exact test). Conserved cofitness may be more predictive because it filters out cases of spurious cofitness between functionally-unrelated genes in one bacterium. Using thresholds of r > 0.8 for cofitness or r > 0.6 for conserved cofitness, we identified at least one association from cofitness or conserved cofitness for 15% of the genes with fitness data and for 44% of genes with statistically significant phenotypes. We identified cofitness associations for all types of proteins (**Fig. 2f**), including for 4,773 vaguely-annotated or hypothetical proteins.

Overall, we identified a functional association (either a specific phenotype or high cofitness) for 25,276 genes, of which 13,192 (52%) had conserved functional associations. Among the genes with conserved associations, 10,699 (81%) had conserved associations across genera and 7,811 (59%) had conservation across divisions. For 2,316 genes with hypothetical or otherwise vague annotations, we identified conserved and hence high-confidence associations (**Supplementary Table 8**).

Genetic overviews of cellular processes

Genome-wide mutant fitness profiling of diverse bacteria provides a broad genetic overview of each biological condition studied. For example, cisplatin reacts with DNA to form crosslinks that block DNA replication, so we expected that genes encoding DNA repair proteins would be specifically important for growth during cisplatin stress [3]. Indeed, of the 67 protein families that were specifically important for resisting cisplatin in more than one bacterium, 33 are known to be involved in DNA repair including the UvrABC nucleotide excision repair complex (**Fig. 3a**, **Supplementary Table 9**). Three of these proteins were recently shown to be involved in DNA repair: RadD (YejH) [23], MmcB (DUF1052) [24], and FAN1-like VRR-NUC domain protein [25], and individual mutants in these genes were sensitive to cisplatin (**Extended Data Figs. 3, 4).** Seven of the other characterized families that had conserved sensitivity to cisplatin are involved in cell division or chromosome segregation (**Fig. 3a**), probably because DNA damage can inhibit cell division and lead to filamentous cells [26]. Among the remaining 27 families, we predicted that 8 poorly-understood families are novel DNA repair families because of their domain content or because of regulation by the DNA damage response regulator LexA [27-29], including the nuclease EndA (**Supplemental Note 2**). An *endA* deletion strain is sensitive to cisplatin (**Extended Data Fig. 4**) and the catalytic residue of EndA's nuclease domain is important for cisplatin resistance (**Extended Data Fig. 5)**.

We obtained similar genetic overviews for many metabolic processes. For example, we examined D-xylose catabolism, which we assayed as the sole carbon source in 12 bacteria. We found that XylAB was important for xylose utilization in *E. coli* and in 8 other bacteria, confirming

its central and conserved role (**Fig. 3b**). In contrast, the well-characterized *E. coli* XylR regulator and XylF transporter are not required in each of the other 8 bacteria: two Pseudomonads use alternative transport proteins for D-xylose while *Phaeobacter inhibens* and *Sinorhizobium meliloti* require a LacI-like regulator for D-xylose utilization, as previously predicted for *P. inhibens* (PGA1_c13990) [9,30]. Three bacteria use an oxidative pathway for D-xylose utilization [31-33] instead of the XylAB pathway (**Supplementary Table 10)**.

More accurate gene annotations

Large-scale mutant fitness data can also be used to improve our understanding of proteins annotated with a general biochemical function but lacking substrate specificity. To illustrate this, we used the mutant fitness data to systematically re-annotate the substrate specificities of 101 permease subunits of ABC transporters that had strong and specific-important phenotypes (fitness < -2) during the utilization of diverse carbon or nitrogen sources (**Fig. 3c**; **Supplementary Table 11**). Using the fitness data, we predicted substrates for all of these proteins (**Supplementary Note 3**). For 24 of the 50 ABC transport proteins that had already been annotated with a substrate (48%), the annotation was incorrect or did not include all of the substrates. Our data also provided more specific annotations: for example, Ac3H11_2942 and Ac3H11_2943 from *Acidovorax* sp. 3H11 were annotated as transporting "various polyols", whereas our data shows they are important for utilizing the polyol D-sorbitol but not the polyol D-mannitol. Overall, we improved the annotations for 75 of 101 transport proteins (**Fig. 3c**).

Next we examined transport proteins with specific-important phenotypes in carbon or nitrogen source experiments and catabolic enzymes with specific-important phenotypes in carbon experiments and identified instances where the mutant fitness data led to a new annotation. In total, we re-annotated 456 proteins that were annotated vaguely or incorrectly in either KEGG [34] or SEED [35]: 238 transport proteins (this includes 68 of the ABC transport proteins described above) and 218 catabolic proteins (**Supplementary Table 12**). Of these 456 proteins, 287 (63%) were not annotated correctly by either KEGG or SEED. Most of the re-annotated proteins are homologous to characterized enzymes but were too distant for the correct substrate to be identified computationally. We also identified a number of proteins that we could link to novel enzymatic reactions (**Supplementary Note 4**, **Supplementary Figs. 1-3**). For example, we identified the putative *Pseudomonas* gene encoding glucosaminate ammonia-lyase, a known biochemical activity with no known gene [36].

Insights for uncharacterized families

We identified a conserved functional association for 335 genes that encode representatives of 87 different domains of unknown function (DUFs) [17] (**Supplementary Table 13**). We examined the phenotypes of 77 of these DUFs and we propose broad functional annotations for 13 of them and specific molecular functions for an additional 8 (**Supplementary Note 5**). For example, proteins containing UPF0126 are specifically important for glycine utilization in 11 bacteria (**Fig. 4a** shows the data from 5 bacteria). Since UPF0126 is predicted to be a membrane protein, we propose that it is a glycine transporter. Individual mutants of three members of this family had reduced growth on glycine (**Extended Data Fig. 6**) and PGA1_c00920 partially rescues the glycine growth defect of an *E. coli* strain that lacks the

glycine transporter CycA [37] (**Extended Data Fig. 7**). Second, we found that genes encoding members of the UPF0060 family of predicted transmembrane proteins have a conserved specific phenotype under thallium stress in four bacteria (**Fig. 4b**). Consequently, we propose that UPF0060-containing proteins may function as a thallium-specific efflux pump. In support of this hypothesis, the expression of UPF0060 proteins confers thallium resistance in *E. coli* (**Extended Data Fig. 8**). Third, we found that in three bacteria, genes encoding DUF2849-containing proteins are cofit with an adjacently encoded sulfite reductase, CysI (**Fig. 4c**). Sulfite reductase is important for sulfate assimilation, which is the only source of sulfur in our defined media. Thus, DUF2849 is also presumably involved in the same process. The three bacteria containing DUF2849 lack CysJ, which typically provides the electron source for CysI. Since other bacterial genomes that contain DUF2849 also contain *cysI* but not *cysJ*, we propose that DUF2849 is an alternate electron source for sulfite reductase. Lastly, we found that the uncharacterized proteins YeaH and YcgB and the poorly characterized protein kinase YeaG [38,39] had high cofitness across seven different bacteria, but with varying phenotypes across bacteria (**Supplementary Figs. 4, 5**). Given the protein kinase activity of YeaG, we propose that these three proteins act together in a conserved signaling pathway that is required for distinct cellular functions in different bacteria.

Discussion

We identified mutant phenotypes for over ten thousand poorly-annotated genes from 32 bacteria. By manually combining these functional associations with comparative sequence analysis, we proposed specific functions for transporter proteins, catabolic enzymes, and domains of unknown function (DUFs), and we identified putative novel DNA repair families. Most of these predictions require additional experimental validation. To facilitate further analyses of mutant phenotypes and protein sequences, we developed the Fitness Browser (http://fit.genomics.lbl.gov).

A major challenge in extending our results to all bacterial proteins is their incredible diversity. We identified functional associations for potential orthologs of just 12% of all bacterial proteins that lack detailed annotations (**Supplementary Note 6, Extended Data Fig. 9**). Improving this coverage will require a larger effort to generate mutants in more diverse bacteria: our study included representatives of only 6 of the ~40 divisions of bacteria that have been cultivated so far. In summary, our study demonstrates the scale with which bacterial fitness data can be collected and the utility of these data to provide insights into the functions of many proteins.

Figure legends

Figure 1. High-throughput genetics for 32 bacteria. (a) Our approach for measuring gene fitness. (b) For each bacterium, we show the types of conditions that we studied and how many genes had statistically significant mutant phenotypes (using a *t*-like test statistic [9] and FDR < 0.05). (c) Gene fitness during the utilization of two carbon sources by *Cupriavidus basilensis*. See **Supplementary Table 6** for details on the highlighted genes. The 4-hydroxybenzoate data is the average from two biological replicates. (d) We classified the genes from all 32 bacteria by how informative their annotations were, and for each class, we show how many genes have each type of phenotype.

Figure 2. Identification of conserved phenotypes. (a) An example of a conserved and specific phenotype. Each point shows the fitness of *crcB* in an experiment, with fluoride stress experiments highlighted in red. Values less than -4 are shown at -4. The y-axis is random. (b) The fraction of *E. coli* genes with a specific phenotype in defined media whose encoded proteins are 'directly' involved in the uptake or catabolism of the compound or regulation thereof. For this analysis, we examined all 61 *E. coli* genes with conserved specific-important phenotypes and a random sample of 40 of the other genes with specific-important phenotypes. The confidence interval is from the binomial test. (c) How many genes of each type had a conserved specific phenotype or a specific phenotype. (d) Comparison of fitness values for *ptsP* and *npr* from *E. coli* (n = 162 independent experiments) and *S. oneidensis* (n = 176 independent experiments). The experiments are color coded by type. *r* is the linear correlation. (e) Using TIGR subroles to test the accuracy of the gene-gene associations. For each query gene with a TIGR subrole, we asked how often the most-cofit gene with a TIGR subrole had the same subrole as the query gene, across varying levels of cofitness or conserved cofitness for that most-cofit gene. The confidence interval is from the binomial test. (f) How many genes of each type had at least one association from conserved cofitness ($r > 0.6$ in both bacteria) or else from cofitness ($r > 0.8$).

Figure 3. Genetic overviews for a condition or a class of proteins. (a) Overview of conserved specific phenotypes in cisplatin stress across 28 bacteria. The data is the average for all successful cisplatin experiments for each bacterium, at up to 5 different concentrations. Each row shows an ortholog group formed by greedy clustering of orthologs (bidirectional best BLAST hits with 80% coverage). Some families are split into multiple ortholog groups and are marked with brackets (i.e. DinG). Some genes have more pleiotropic phenotypes in some (but not all) bacteria. (b) Overview of specific phenotypes for the utilization of D-xylose as a carbon source in 12 bacteria. Putative orthologs are included in the heatmap even if they are not important for D-xylose utilization. In the oxidative pathway, gene names for *xylBCDX* are from Stephens et al. [33] and should not be confused with *E. coli xylB*, which is not related. The data is the average of 1 to 3 replicate experiments for each bacterium. The color scale is the same as panel (a). (c) Summary of annotation improvements for ABC transporter proteins based on an analysis of specific phenotypes.

Figure 4. Conserved functional associations for genes encoding uncharacterized protein families. (a, b) Conserved specific phenotypes for proteins of unknown function. Each point represents the fitness of the protein in an individual experiment. Values under -4 are shown at -4. The y-axis is random. (c) Heatmap of fitness data for *cysI* (sulfite reductase) and DUF2849 in three bacteria.

References main text

1. Chang, Y.-C. *et al.* COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. *Nucleic Acids Res.* **44,** D330–5 (2016).
2. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* **5,** e1000605 (2009).
3. Deutschbauer, A. *et al.* Towards an informative mutant phenotype for every bacterial

gene. *J. Bacteriol.* **196,** 3643–3655 (2014).
4. Deutschbauer, A. *et al.* Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. *PLoS Genet.* **7,** e1002385 (2011).
5. Nichols, R. J. *et al.* Phenotypic landscape of a bacterial cell. *Cell* **144,** 143–156 (2011).
6. Price, M. N. *et al.* The genetic basis of energy conservation in the sulfate-reducing bacterium Desulfovibrio alaskensis G20. *Front Microbiol* **5,** 577 (2014).
7. Langridge, G. C. *et al.* Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res* **19,** 2308–2316 (2009).
8. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* **6,** 767–772 (2009).
9. Wetmore, K. M. *et al.* Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio* **6,** e00306–15 (2015).
10. Liu, H. *et al.* Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria. *mSystems* **3,** e00143–17 (2018).
11. Rubin, B. E. *et al.* The essential gene set of a photosynthetic organism. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E6634–43 (2015).
12. Melnyk, R. A. *et al.* Novel mechanism for scavenging of hypochlorite involving a periplasmic methionine-rich Peptide and methionine sulfoxide reductase. *MBio* **6,** e00233–15 (2015).
13. Smith, A. M. *et al.* Quantitative phenotyping via deep barcode sequencing. *Genome Res* **19,** 1836–1842 (2009).
14. Rensing, C., Pribyl, T. & Nies, D. H. New functions for the three subunits of the CzcCBA cation-proton antiporter. *J. Bacteriol.* **179,** 6871–6879 (1997).
15. Hottes, A. K. *et al.* Bacterial Adaptation through Loss of Function. *PLoS Genet.* **9,** e1003617 (2013).
16. Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* **41,** D387–95 (2013).
17. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–30 (2014).
18. Baker, J. L. *et al.* Widespread genetic switches and toxicity resistance proteins for fluoride. *Science* **335,** 233–235 (2012).
19. Keseler, I. M. *et al.* EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **41,** D605–12 (2013).
20. Hillenmeyer, M. E. *et al.* Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome Biol.* **11,** R30 (2010).
21. Rabus, R., Reizer, J., Paulsen, I. & Saier, M. H. Enzyme I(Ntr) from Escherichia coli. A novel enzyme of the phosphoenolpyruvate-dependent phosphotransferase system exhibiting strict specificity for its phosphoryl acceptor, NPr. *J Biol Chem* **274,** 26185–26191 (1999).
22. van Opijnen, T., Dedrick, S. & Bento, J. Strain Dependent Genetic Networks for Antibiotic-Sensitivity in a Bacterial Pathogen with a Large Pan-Genome. *PLoS Pathog.* **12,** e1005869 (2016).
23. Chen, S. H., Byrne, R. T., Wood, E. A. & Cox, M. M. Escherichia coli radD (yejH) gene: a novel function involved in radiation resistance and double-strand break repair. *Mol Microbiol* **95,** 754–768 (2015).

24. Lopes-Kulishev, C. O. *et al.* Functional characterization of two SOS-regulated genes involved in mitomycin C resistance in Caulobacter crescentus. *DNA Repair (Amst.)* **33,** 78–89 (2015).
25. Gwon, G. H. *et al.* Crystal structure of a Fanconi anemia-associated nuclease homolog bound to 5' flap DNA: basis of interstrand cross-link repair by FAN1. *Genes Dev.* **28,** 2276–2290 (2014).
26. Justice, S. S., Hunstad, D. A., Cegelski, L. & Hultgren, S. J. Morphological plasticity as a bacterial survival strategy. *Nature Publishing Group* **6,** 162–168 (2008).
27. da Rocha, R. P., Paquola, A. C. de M., Marques, M. D. V., Menck, C. F. M. & Galhardo, R. S. Characterization of the SOS regulon of Caulobacter crescentus. *J. Bacteriol.* **190,** 1209–1218 (2008).
28. Abella, M., Campoy, S., Erill, I., Rojo, F. & Barbé, J. Cohabitation of two different lexA regulons in Pseudomonas putida. *J. Bacteriol.* **189,** 8855–8862 (2007).
29. Cirz, R. T., O apos Neill, B. M., Hammond, J. A., Head, S. R. & Romesberg, F. E. Defining the Pseudomonas aeruginosa SOS response and its role in the global response to the antibiotic ciprofloxacin. *J. Bacteriol.* **188,** 7101–7110 (2006).
30. Wiegmann, K. *et al.* Carbohydrate catabolism in Phaeobacter inhibens DSM 17395, a member of the marine roseobacter clade. *Appl Environ Microbiol* **80,** 4725–4737 (2014).
31. Brouns, S. J. J. *et al.* Identification of the missing links in prokaryotic pentose oxidation pathways: evidence for enzyme recruitment. *J Biol Chem* **281,** 27378–27388 (2006).
32. Johnsen, U. *et al.* D-xylose degradation pathway in the halophilic archaeon Haloferax volcanii. *J Biol Chem* **284,** 27290–27303 (2009).
33. Stephens, C. *et al.* Genetic analysis of a novel pathway for D-xylose metabolism in Caulobacter crescentus. *J. Bacteriol.* **189,** 2181–2185 (2007).
34. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–62 (2016).
35. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42,** D206–14 (2014).
36. Iwamoto, R. & Imanaga, Y. Direct evidence of the Entner-Doudoroff pathway operating in the metabolism of D-glucosamine in bacteria. *J. Biochem.* **109,** 66–69 (1991).
37. Ghrist, A. C. & Stauffer, G. V. The Escherichia coli glycine transport system and its role in the regulation of the glycine cleavage enzyme system. *Microbiology (Reading, Engl.)* **141 (Pt 1),** 133–140 (1995).
38. Figueira, R. *et al.* Adaptation to sustained nitrogen starvation by Escherichia coli requires the eukaryote-like serine/threonine kinase YeaG. *Sci Rep* **5,** 17524 (2015).
39. Tagourti, J., Landoulsi, A. & Richarme, G. Cloning, expression, purification and characterization of the stress kinase YeaG from Escherichia coli. *Protein Expr. Purif.* **59,** 79–85 (2008).

Acknowledgements

We thank Victoria Lo, Wenjun Shao, and Keith Keller for technical assistance with the Fitness Browser web site. Sequencing was performed at: the Vincent J. Coates Genomics Sequencing Laboratory (University of California at Berkeley), supported by NIH S10 Instrumentation Grants S10RR029668, S10RR027303, and OD018174; at the DOE Joint Genome Institute; at the

College of Biological Sciences UCDNA Sequencing Facility (UC Davis); and at the Institute for Genomics Sciences (University of Maryland).

Studies of novel isolates were conducted by ENIGMA and were supported by the Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy, under contract DE-AC02-05CH11231. The other data collection was supported by Laboratory Directed Research and Development (LDRD) funding from Berkeley Lab, provided by the Director, Office of Science, of the U.S. Department of Energy under contract DE-AC02-05CH11231 and a Community Science Project from the Joint Genome Institute to M.J.B., J.B., A.P.A., and A.M.D. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

Author contributions

AMD, APA, MNP, MJB, and JB conceived the project. AMD, APA, MJB, and JB supervised the project. AMD led the experimental work. AMD, KMW, RJW, RAM, MC, JR, JVK, HL, HKC, JSL, YS, ZE, and HS collected data. RC isolated bacteria. MNP and AMD analyzed the fitness data. RAM, RJW, and MNP assembled genomes. BER provided resources and advice on *S. elongatus* experiments. GMZ and JDW generated gene deletion mutants in *Pseudomonas stutzeri* RCH2. AV edited the manuscript and provided advice. MNP, MJB, and AMD wrote the paper.

Author information

The authors declare no competing financial interests.

Methods

Bacteria for high-throughput genetics. We attempted to make transposon mutant libraries in over 100 bacteria as part of this study, including representatives of Proteobacteria, Bacteroidetes, Firmicutes, Actinobacteria, and Planctomycetes; we present data from the 32 bacteria that we successfully applied RB-TnSeq to (**Supplementary Table 14)**. Eight bacteria were isolated from groundwater collected from different monitoring wells at the Oak Ridge National Laboratory Field Research Center (FRC; http://www.esd.ornl.gov/orifrc/), and five have not been described previously: *Acidovorax* sp. GW101-3H11, *Pseudomonas fluorescens* FW300-N1B4, *P. fluorescens* FW300-N2E3, *P. fluorescens* FW300-N2C3, and *P. fluorescens* GW456-L13. *Acidovorax* sp. GW101-3H11 was isolated as a single colony on a Luria-Bertani (LB) agar plate grown at 30°C using an inoculum from FRC well GW101. *P. fluorescens* FW300-N1B4, *P. fluorescens* FW300-N2E3, and *P. fluorescens* FW300-N2C3 were all isolated at 30°C under anaerobic denitrifying conditions with acetate, propionate, and butyrate as the carbon source, respectively, using inoculum from FRC well FW300. *Pseudomonas fluorescens* GW456-L13 was isolated from FRC well FW456 under anaerobic incubations on a LB agar plate. We previously described the isolation of *Pseudomonas fluorescens* FW300-N2E2 [40], *Cupriavidus basilensis* 4G11 [41], and *Pedobacter* sp. GW460-11-11-14-LB5 [10]. The eight FRC isolates have been submitted to the Leibniz Institute DSMZ (German Collection of Microorganisms and Cell Cultures GmbH).

Our study included 7 strains of *Pseudomonas* and 4 strains of *Shewanella* because of their genetic tractability. The different strains in these genera have large differences in gene content. For example, the two most closely-related strains we studied were *P. fluorescens* FW300-N2E2 and FW-N2C3, and the typical pair of orthologous proteins from these strains has 97% amino acid identity. Nevertheless, each genome contains over 1,000 predicted protein-coding genes that do not have orthologs in the other strain.

Strain construction for single gene studies. The oligonucleotides and gBlocks used in this study are listed in **Supplementary Table 15**, the plasmids and details on their construction in **Supplementary Table 16**, and the single strains for follow-up studies in **Supplementary Table 17**. We constructed kanamycin-marked gene deletions in *Pseudomonas stutzeri* RCH2 using a previously described double homologous recombination strategy [42]. We used a similar strategy to construct kanamycin-marked deletions in *Dinoroseobacter shibae* and *Phaeobacter inhibens*, except we delivered the deletion constructs by conjugation from *E. coli*. We constructed markerless gene deletions of SO4008 and SO1319 from *Shewanella oneidensis* MR-1 and Pf6N2E2_4800 and Pf6N2E2_4801 from *Pseudomonas fluorescens* FW300-N2E2 using *sacB* counter-selection. For genetic complementation experiments, we cloned the genes into the broad range vector pBBR1-MCS5 [43]. For the thallium resistance experiments, we cloned members of UPF0060 into plasmid pFAB2286. For *E. coli* BW25113, we used single-gene deletions from the Keio collection [44]. For *S. oneidensis* MR-1, we also used transposon mutants that had been individually sequenced [4]. Unless indicated otherwise, we tested the growth phenotypes of these individual strains in 96-well microplates at 30°C using the standard rich or minimal defined growth media for each bacterium. These growth assays were performed in a Tecan microplate reader (either Sunrise or Infinite F200) with absorbance readings (OD_{600}) every 15 minutes.

Media and standard culturing conditions. A full list of the media used in this study and their components are given in **Supplementary Table 18**. We routinely cultured *Acidovorax* sp. GW101-3H11, *Azospirillum brasilense* Sp245, *Burkholderia phytofirmans* PsJN, *Dyella japonica* UNC79MFTsu3.2, *Escherichia coli* BW25113, *Herbaspirillum seropedicae* SmR1, *Klebsiella michiganensis* M5a1, all Pseudomonads and Shewanellae, *Sinorhizobium meliloti* 1021, and *Sphingomonas koreensis* DSMZ 15582 in LB. *Caulobacter crescentus* NA1000 was typically cultured in PYE media. *Cupriavidus basilensis* 4G11 and *Pedobacter* sp. GW460-11-11-14-LB5 were typically cultured in R2A media. *Dechlorosoma suillum* PS was cultured in ALP media [12]. *Desulfovibrio vulgaris* Miyazaki F was grown anaerobically in lactate-sulfate (MOLS4) media, as previously described [45,46]. We used marine broth (Difco 2216) for standard culturing of *Dinoroseobacter shibae* DFL-12, *Echinicola vietnamensis*, *Kangiella aquimarina* SW-154T, *Marinobacter adhaerens* HP15, *Phaeobacter inhibens* BS107, and *Pontibacter actiniarum*. *Synechococcus elongatus* PCC 7942 was normally cultured in BG-11 media with either 7,000 or 9,250 lux. All bacteria were typically cultured at 30°C except *E. coli* BW25113 and *Shewanella amazonensis* SB2B, which were cultured at 37°C, and *P. inhibens* BS107, which was grown at 25°C. The *E. coli* conjugation strain WM3064 was cultured in LB at 37°C with diaminopimelic acid (DAP) added to a final concentration of 300 μM.

High-throughput growth assays of wild-type bacteria. To assess the phenotypic capabilities of 30 aerobic heterotrophic bacteria and to identify conditions suitable for mutant fitness profiling, we monitored the growth of the wild-type bacterium in a 96-well microplate assay. These prescreen growth assays were performed in a Tecan microplate reader (either Sunrise or Infinite F200) with absorbance readings (OD_{600}) every 15 minutes. All 96-well microplate growth assays contained 150 μL culture volume per well at a starting OD_{600} of 0.02. We used the grofit package in R [47] to analyze all growth curve data in this study. For carbon and nitrogen source utilization, we tested 94 and 45 possible substrates, respectively, in a defined medium (**Supplementary Tables 2 and 3**). We classified a bacterium as positive for usage of a particular substrate if (1) the maximum OD_{600} on the substrate was at least 1.5 greater than the average of the water controls and the integral under the curve (spline.integral) was 10% greater than the average of the water controls or (2) a successful genome-wide fitness assay was collected on the substrate, as described below. We included the second criterion because our automated scoring of the wild-type growth curves was conservative and did not include all conditions used for genome-wide fitness assays.

Additionally, for each of the 30 heterotrophic bacteria, we determined the inhibitory concentrations for 34 to 55 diverse stress compounds including antibiotics, biocides, metals, furans, aldehydes, and oxyanions. For each compound, we grew the wild-type bacterium across a 1,000-fold range of inhibitor concentrations in a rich media. We used the spline.integral parameter of grofit to fit dose-response curves and calculate the half-maximum inhibitory concentration (IC_{50}) values for each compound (**Supplementary Table 4)**. For *Desulfovibrio vulgaris* Miyazaki F and *Synechococcus elongatus* PCC 7942, we did not perform these growth prescreen assays, rather, we just performed the mutant fitness assays across a broad range of inhibitor concentrations.

Genome sequencing. We sequenced *Acidovorax sp.* GW101-3H11 and five Pseudomonads by using a combination of Illumina and Pacific Biosciences. For Illumina-first assembly, we used scythe (https://github.com/vsbuffalo/scythe) and sickle (https://github.com/najoshi/sickle) to trim

and clean Illumina reads, we assembled with SPAdes 3.0 [48], we performed hybrid Illumina/PacBio assembly on SMRTportal using AHA [49], we used BridgeMapper on SMRTportal to fix misassembled contigs, we mapped Illumina reads to the new assembly with bowtie 2 [50], and we used pilon to correct local errors [51]. For *Acidovorax sp.* GW101-3H11, we instead used A5 [52] to assemble the Illumina reads and we used AHA to join contigs together. For PacBio-first assembly, we used HGAP3 on SMRTportal, we used circlator to find additional joins [53], and we again used bowtie 2 and pilon to correct local errors. See **Supplementary Table 19** for a summary of these genome assemblies and their accession numbers. In addition, *Sphingomonas koreensis* DSMZ 15582 was sequenced for this project by the Joint Genome Institute, using Pacific Biosciences.

Constructing pools of randomly barcoded transposon mutants. The transposon mutant libraries for ten bacteria were described previously [9-12]. The other 22 bacteria were mutagenized with randomly barcoded plasmids containing a *mariner* or *Tn*5 transposon, a *pir*-dependent conditional origin of replication, and a kanamycin resistance marker, using the vectors that we described previously [9]. The plasmids were delivered by conjugation with *E. coli* WM3064, which is a diaminopimelate auxotroph and is *pir*$^+$. The conditions for mutagenizing each organism are described in **Supplementary Table 20**. Generally, we conjugated mid-log phase grown WM3064 donor (either *mariner* donor plasmid library APA752 or *Tn*5 donor plasmid library APA766) and recipient cells on 0.45 µM nitrocellulose filters (Millipore) overlaid on rich media agar plates supplemented with DAP. We used the rich medium preferred by the recipient (**Supplementary Table 20**). After conjugation, filters were resuspended in recipient rich media and plated on recipient rich media agar plate supplemented with kanamycin. After growth, we scraped together kanamycin resistant colonies into recipient rich media with kanamycin, diluted the culture back to a starting OD_{600} of 0.2 in 50-100 mL of recipient rich media with kanamycin, and grew the mutant library to a final OD_{600} of between 1.0 and 2.0. We added glycerol to a final volume of 10%, made multiple 1 or 2 mL -80ºC freezer stocks, and collected cell pellets to extract genomic DNA for TnSeq. For *Desulfovibrio vulgaris* Miyazaki F, we selected for G418-resistant transposon mutants in liquid media with no plating step (**Supplementary Table 20**).

Transposon insertion site sequencing (TnSeq). Given a pool of mutants, we performed TnSeq to amplify and sequence the transposon junction and to link the barcodes to a location in the genome [9]. In this study, we used the TnSeq data for two independent analyses. First, we used the mapped transposon insertions to identify genes that are likely to be essential for viability (or nearly so) in the conditions that we used to select the mutant library (see below). We performed this gene essentiality analysis independently of the DNA barcodes and we therefore considered transposon insertion locations that were not included in the mutant pool definition for BarSeq fitness assays. Second, we used the TnSeq data to define the mutant library for high-throughput mutant fitness assays. Here, we analyzed both the transposon insertion data and its associated random DNA barcode, as the RB-TnSeq approach requires an accurate association of the genomic insertion location of the transposon to a unique, random DNA barcode. We considered a barcode to be confidently mapped to a location if this mapping was supported by at least 10 reads and the barcode mapped primarily to one location [9]. (For *Shewanella* sp. ANA-3, the threshold was 8 reads.) The number of unique barcodes (strains) mapped in each mutant library is shown in **Supplementary Table 20**. Given this mapping, the abundance of the strains in each sample can be determined by a simpler and cheaper protocol: amplifying the barcodes with PCR followed by barcode sequencing [13].

Identifying essential or nearly essential genes. Genes that lack insertions or that have very low coverage in the start samples are likely to be essential or to be important for growth (nearly essential) in rich media, as except for *Synechococcus elongatus*, pools of mutants were produced and recovered in media that contained yeast extract. We used previously published heuristics [11] to distinguish likely-essential genes from genes that are too short or that are too repetitive to map insertions in. Briefly, for each protein-coding gene, we computed the total read density in TnSeq (reads / nucleotides across the entire gene) and the density of insertion sites within the central 10-90% of each gene (sites/nucleotides). We did not consider the DNA barcodes in this analysis of essential genes. Across the 32 bacteria, we mapped 5 – 66 insertions for the typical (median) protein-coding gene. We then excluded genes that might be difficult to map insertions within because they were very similar to other parts of the genome (BLAT score above 50) and also very-short genes of less than 100 nucleotides. Given the median insertion density and the median length of the remaining genes, we asked how short a gene could be and still be unlikely to have no insertions at all by chance ($P < 0.02$, Poisson distribution). Genes shorter than this threshold were excluded; the threshold varied from 100 nucleotides for *Phaeobacter inhibens, Caulobacter crescentus, and Synechococcus elongatus* to 675 nucleotides for *Shewanella loihica* PV-4. For the remaining genes, we normalized the read density by GC content by dividing by the running median of read density over a window of 201 genes (sorted by GC content). We normalized the insertion density so that the median gene's value was 1. Protein-coding genes were considered essential or important for growth (nearly essential) if we did not estimate fitness values for the gene and both the normalized insertion density and the normalized read density were under 0.2. A validation of this approach is described in **Supplementary Note 1**.

Mutant fitness assays. For each mutant library, we performed competitive mutant fitness assays under a large number of growth conditions that were chosen based on the results of high-throughput growth assays of wild-type bacteria (see above). For some bacteria, we also assayed growth at varying pH or temperature, motility on agar plates, or survival. The conditions that we profiled varied across the bacteria due to the differing growth capabilities of the bacteria as well as experimental limitations. For example, the heterotrophic bacteria we investigated showed a wide range in the number of compounds (either carbon or nitrogen) that they could utilize for growth in a defined media (**Supplementary Table 2, Supplementary Table 3**). For example, only 14 of the 32 bacteria were capable of utilizing D-xylose as the sole carbon source, and we successfully assayed mutant fitness in 12 of these strains. We did not perform carbon and nitrogen source experiments in *K. aquimarina* SW-154T and *P. actiniarum* as we could not culture these bacteria in a defined growth media. In addition, some stress experiments were not performed because of the native resistance of the bacteria to the compound (**Supplementary Table 4**).

In addition to biological reasons for the differences among exact set of conditions we profiled, for some bacteria we did not attempt certain assays. We did not systematically test *D. vulgaris* Miyazaki F for carbon source utilization and we did not perform nitrogen source experiments in two heterotrophic bacteria that grow in defined media: *D. suillum* PS and *S. loihica* PV-4. For 31 of the 32 bacteria, we attempted fitness assays for a core set of 32 stress compounds (see **Supplementary Table 4**), but for *D. suillum* PS, we attempted only 16 of them. We successfully studied motility in 12 of the 32 bacteria using a soft agar assay. For the other bacteria, we either

did not attempt a motility assay or the cells were motile but the mutant fitness data did not pass our thresholds for a successful experiment. Similarly, there were many growth-based fitness experiments (carbon and nitrogen source, and stressors) that we performed but did not pass our metrics for a successful experiment [9]. Although a few dozen of these samples had low read depth, which might indicate a problem during PCR of the barcodes or sequencing, we believe that most of these experiments failed due to biological factors that make them incompatible with a pooled fitness assay, such as intense positive selection or potentially stochastic exit from lag phase [9]. For example, across all 32 bacteria, there are 197 bacterium x stress combinations that lack fitness data because none of the experiments for that stress succeeded by our metrics. For 85 of these combinations, we attempted more than one experiment.

The full list of experiments performed for each mutant library along with detailed metadata is available in **Supplementary Table 5**, on figshare (https://doi.org/10.6084/m9.figshare.5134837) and at http://genomics.lbl.gov/supplemental/bigfit/. Our analysis includes 385 successful experiments from Wetmore *et al.* [9] and 36 successful experiments from Melnyk *et al.* [12]. The other 4,449 successful fitness assays are described here for the first time. In general, all growth assays with carbon sources, nitrogen sources, and inhibitors were done as previously described [9]. Briefly, an aliquot of the mutant library was thawed and inoculated into 25 mL of rich media with kanamycin or erythromycin and grown to mid-log phase in a flask. The only exception was *Klebsiella michiganensis*, which we recovered to stationary phase; experiments with the usual exponential phase start samples for this library had suspicious correlations of gene fitness with GC content and did not meet our quality metrics. Depending on the mutant library, this growth recovery took between 3 and 24 hours. After recovery, we collected pellets for genomic DNA extraction and barcode sequencing (BarSeq) of the start sample. We used the remaining cells to set up multiple mutant fitness assays with diverse carbon and nitrogen sources in defined media and diverse inhibitors in rich media, all at a starting OD_{600} of 0.02. In addition, for most bacteria, we profiled growth of the mutant library at different pH and at different temperatures. After the mutant library grew to saturation under the selective growth condition (typically 4 to 8 population doublings), we collected a cell pellet for genomic DNA extraction and BarSeq of the "end" sample. As described below, we calculate gene fitness from the barcode counts of the end sample relative to the start sample.

We used a number of different growth formats and media formations for mutant fitness assays across the 32 bacteria. A full list of compound components for each growth media are contained in **Supplementary Table 18**. Many fitness assays were done in 48-well microplates (Greiner) with 700 µL culture volume per well and grown in a Tecan Infinite F200 plate reader with OD_{600} measurements every 15 minutes. For these 48-well microplate assays, we combined the cultures from two replicate wells before genomic DNA extraction (total volume of experiment = 1.4 mL). For 24-well microplate experiments, we typically used deep-well plates with 1.5 mL (for inhibitors) or 2 mL (for carbon and nitrogen sources) total culture volume per well. For six bacteria (*Caulobacter crescentus*, *Dyella japonica* UNC79MFTsu3.2, *Echinicola vietnamensis*, *Klebsiella michiganensis*, *Pedobacter* sp. GW460-11-11-14-LB5, *Pontibacter actiniarum*), we performed all fitness assays in transparent 24-well microplates (Greiner) with 1.2 mL total volume per well. All 24-well microplate experiments were grown in a Multitron incubating shaker (Innova). For 24-well microplate experiments, we typically took the OD_{600} of each culture every 12 to 24 hours in a Tecan plate reader (after transferring the cells to a Greiner 96-well microplate for the 24-well deep well plate experiments). Because we used different methods for

measuring OD_{600} values (cuvettes with a spectrophotometer; 24, 48 and 96 well microplates in a plate reader), we used standard curves to interconvert microplate OD_{600} values to a common reference (cuvette with spectrophotometer). Over 1,000 experiments, primarily carbon source and temperature experiments, were done in glass test tubes with 5 mL culture volumes. For the test tube experiments, we monitored OD_{600} every 12 to 24 hours with a standard spectrophotometer and cuvettes.

For stress experiments, we aimed to use an inhibitory but sub-lethal concentration of the compound, typically at a concentration that resulted in a 50% reduction of the growth rate (IC_{50}). Both the growth and the inhibition are crucial for identifying mutants that have altered sensitivity to the compound. If the concentration of inhibitor is too high and there is no growth, then the abundance of the strains will not change and all of the fitness values will be near zero. If the concentration of inhibitor is too low and there is no growth inhibition, then the fitness pattern is likely to be as if the compound were not added at all. We identified these inhibitory concentrations using the growth curves with the wild-type bacteria described above. In addition to the calculated IC_{50} for a compound, we often used a few different concentrations above and below the IC_{50} to try and capture an inhibitory but sub-lethal concentration. We used multiple concentrations of a single compound because we found that the IC_{50} values determined with wild-type bacteria in 96-well microplates (see above) sometimes did not agree with the mutant fitness experiments, which were done with a complex transposon mutant library in 24 or 48-well microplates. For assays done in 48-well microplates and grown in a Tecan Infinite F200 plate reader, we could confirm that the culture was inhibited relative to a no stress control. For stress assays in 24-well microplates and grown in the Multitron shaker, we took OD readings approximately every 12 hours to estimate which cultures were inhibited. In practice, for a given mutant library, we often collected mutant fitness data with different concentrations of the same inhibitor. We also collected fitness data in plain rich media without an added inhibitory compound.

For some carbon source experiments in *Desulfovibrio vulgaris* Miyazaki F, which is strictly anaerobic, we grew the mutant pool in 18 x 150 mm hungate tubes with a butyl rubber stopper and an aluminum crimp seal (Chemglass Life Sciences, Vineland, NJ) with a culture volume of 10 mL and a headspace of about 15 mL. For the remainder of the *Desulfovibrio vulgaris* fitness experiments, we grew the mutant pool in 24-well microplates inside of the anaerobic chamber. We used OD_{600} measurements to determine which cultures were inhibited by varying concentrations of stress compounds. Similarly, for six of the other heterotrophs, we measured gene fitness during anaerobic growth. All anaerobic media was prepared within a Coy anaerobic chamber with an atmosphere of about 2% H_2, 5% CO_2, and 93% N_2.

For *Synechococcus elongatus* PCC 7942, which is strictly photosynthetic, we recovered the library from the freezer in BG-11 media at a light level of 7000 lux and we conducted fitness assays at 9250 lux. We used OD_{750} to measure the growth of *S. elongatus*. Most *S. elongatus* mutant fitness assays were done in the wells of a 12-well microplate (Falcon) with a 5 mL culture volume.

For motility assays, the mutant pool was inoculated into the center of a 0.3% agar rich media plate and “outer” samples with motile cells were removed with a razor after 24-48 hours. In many instances, we also removed an “inner” sample of cells from near the point of inoculation.

Not all bacteria we assayed were motile in this soft agar assay and others were motile but did not give mutant fitness results that passed our quality metrics.

In four bacteria, we also assayed survival. In these assays, a mutant pool was subjected to a stressful condition (either extended stationary phase, starvation, or a low temperature of 4°C) for a defined period; then, to determine which strains are still viable, they were recovered in rich media for a few generations. After recovery in rich media, the cells were harvested for genomic DNA extraction and BarSeq.

We performed replicate experiments (not necessary at the same concentration) for 25 of the 29 bacteria with carbon source data; the exceptions are *B. phytofirmans* PsJN, *P. fluorescens* FW300-N1B4, *P. fluorescens* FW300-N2E3, and *S. meliloti* 1021. Similarly, we performed replicate experiments (not necessarily at the same concentration) for 17 of the 28 bacteria with nitrogen source data. The bacteria with nitrogen source data but without biological replicates are: *Acidovorax* sp. GW101-3H11, *B. phytofirmans* PsJN, *C. basilensis* 4G11, *D. japonica* UNC79MFTsu3.2, *H. seropedicae* SmR1, *P. fluorescens* FW300-N1B4, *P. fluorescens* FW300-N2E3, *P. fluorescens* FW300-N2C3, *P. fluorescens* GW456-L13, *S. meliloti* 1021, and *P. simiae* WCS417. Lastly, we performed replicates (not necessarily at the same concentration) for the majority of stress experiments for 16 of the 32 bacteria: *C. crescentus* N1000, *E. vietnamensis*, *D. shibae* DFL-12, *K. aquimarina* SW-154T, *S. koreensis* DSMZ 15582, *K. michiganensis* M5a1, *M. adhaerens* HP15, *D. vulgaris* Miyazaki F, *S. oneidensis* MR-1, *P. inhibens* BS107, *P. actiniarum, D, suillum* PS, *P. stutzeri* RCH2, *Pedobacter* sp. GW460-11-11-14-LB5, *S. loihica* PV4, and *S. elongatus* PCC 7942.

Barcode sequencing (BarSeq). Genomic DNA extraction and barcode PCR were performed as described previously [9]. Most genomic DNA extractions were done in a 96-well format using a QIAcube HT liquid handling robot (QIAGEN). We used the 98°C BarSeq PCR protocol [9], which is less sensitive to high GC content. In general, we multiplexed 48 samples or 96 samples per lane of Illumina HiSeq. For *E. coli*, we sequenced 96 samples per lane. For sequencing runs on the Illumina HiSeq4000, we sequenced 96 samples per lane. For the HiSeq4000 runs, we used an equimolar mixture of four common P1 oligos for BarSeq, with variable lengths of random bases at the start of the sequencing reactions (2 – 5 nucleotides) (**Supplementary Table 15**). We did this to phase the amplicons for sequencing and to aid in cluster discrimination on the HiSeq4000.

Computation of fitness values. Fitness data was analyzed as previously described [9]. Briefly, the fitness value of each strain (an individual transposon mutant) is the normalized $\log_2$(strain barcode abundance at end of experiment/strain barcode abundance at start of experiment). The fitness value of each gene is the weighted average of the fitness of its strains. In this study, we restricted our analysis to the 123,255 different non-essential protein-coding genes for which we collected gene fitness data. Gene fitness values describe the relative abundance of the mutant strains in a condition, regardless of their fitness in rich media. Only strains containing insertions within the central 10-90% of the gene and with sufficient abundance, on average, in the start samples were included in these calculations; we used 3 to 26 mutant strains for the typical protein-coding gene in each bacterium (**Supplementary Table 20**). Specifically, in the average start samples, we require that each gene have at least 30 reads, and each individual strain for that gene to have at least 3 reads. Because we include the same set of strains in the analysis of

each experiment, there are no instances where a gene has a fitness value in one experiment but not in another. The gene fitness values were then normalized to remove the effects of variation in genes' copy number, for example due to variation in plasmid copy number relative to the chromosome or due to higher effective copy number near the origin of replication in actively dividing cells. The median for each scaffold is set to zero, and for large scaffolds (over 250 genes), the running median of the gene fitness values is subtracted. Also, for large scaffolds (over 250 genes), the peak of the distribution of gene fitness values is set to be at zero.

Fitness experiments were deemed successful using the quality metrics that we described previously [9]. These metrics ensure that the typical gene has sufficient coverage, that the fitness values of independent insertions in the same gene are consistent, and that there is no GC bias [9]. Experiments that did not meet these thresholds were excluded from our analyses. The remaining experiments show good agreement between exact replicates, with a median correlation of 0.87 for gene fitness values from defined media experiments. 95% of the replicate defined media experiments had $r >= 0.64$. Stress experiments sometimes have little biological signal, as they are usually done in rich media and mutants of genes that are important for growth in rich media may be absent from the pools. Nevertheless, the median correlation between replicate stress experiments was 0.74 and 95% replicate stress experiments had $r >= 0.35$.

To estimate the reliability of the fitness value for a gene in a specific experiment, we use a *t*-like test statistic which is the gene's fitness divided by the standard error [9]. The standard error is the maximum of two estimates. The first estimate is based on the consistency of the fitness for the strains in that gene. The second estimate is based on the number of reads for the gene.

Even mild phenotypes were quite consistent between replicate experiments if they were statistically significant. For example, if a gene had a mild but significant phenotype in one replicate ($0.5 < |\text{fitness}| < 2$ and $|t| > 4$), then the sign of the fitness value was the same in the other replicate 96.7% of the time. Because this comparison might be biased if the two replicates were compared to the same control sample, only replicates with independent controls were included.

Genes with statistically significant phenotypes. We averaged fitness values from exact replicate experiments. We combined *t* scores across replicate experiments with two different approaches. If the replicates did not share a start sample and were entirely independent, then we used $t_{comb} = \text{sum}(t) / \text{sqrt}(n)$, where n is the number of replicates. But if the replicates used the same start sample then this metric would be biased. To correct for this, we assumed that the start and end samples have similar amounts of noise. This is conservative because we usually sequenced the start samples with more than one PCR and with different multiplexing tags. Given this assumption and given that variance(A+B) = variance(A) + variance(B) if A and B are independent random variables, it is easy to show that the above estimate of t_{comb} needs to be divided by sqrt((n**2 + n)/(2*n)), where n is the number of replicates.

Our standard threshold for a statistically significant phenotype was |fitness| > 0.5 and |combined *t*| > 4, but this was increased for some bacteria to maintain a false discovery rate (FDR) of less than 5%. We use a minimum threshold on |fitness| as well as |*t*| to account for imperfect normalization or for other small biases in the fitness values. To estimate the number of false

positives, we used control experiments, that is, comparisons between different measurements of different aliquots of the same start sample. However we did not use some previously-published control comparisons (from [9]) that used the old PCR settings and had strong GC bias (to exclude these, we used the same thresholds that we used to remove biased experiments). The estimated number of false positive genes was then the number of control measurements that exceeded the thresholds, multiplied by the number of conditions and divided by the number of control experiments. As a second approach to estimate the number of false positives, we used the number of expected false positives if the t_{comb} values follow the standard normal distribution (2 * P(z > t) * #experiments * #genes). If either estimate of the false discovery rate was above 5%, we raised our thresholds for both |fitness| and $|t_{comb}|$ in steps of 0.1 and 0.5, respectively, until FDR < 5%. The highest thresholds used were |fitness| > 0.9 and $|t|$ > 6. Also, for *Pseudomonas fluorescens* FW300-N1B4, we identified six genes with large differences between control samples (|fitness| ≈ 2 and $|t|$ ≈ 6). These genes cluster on the chromosome in two groups and have high cofitness, and several of the genes are annotated as being involved in capsular polysaccharide synthesis. Because this bacterium is rather sticky, we suspect that mutants in these genes are less adherent and were enriched in some control samples due to insufficient vortexing, so these six genes were excluded when estimating the number of false positives.

Sequence analysis. To assign genes to Pfams [17] or TIGRFAMs [16], we used HMMer 3.1b1 [54] and the trusted score cutoff for each family. We used Pfam 28.0 and TIGRFAM 15.0. We used only the curated families in Pfam ("Pfam A").

To identify putative orthologs between pairs of genomes, we used bidirectional best protein BLAST hits with at least 80% alignment coverage both ways. We did not use any cutoff on similarity, as a similarity of phenotype can show that distant homologs have conserved functions. For the analysis of conserved specific phenotypes with cisplatin stress and xylose carbon catabolism, we used greedy clustering to go from pairs of orthologs to ortholog groups. In some instances, a single protein family was split into multiple ortholog groups and combined manually.

To estimate the evolutionary relationships of the bacteria that we studied (**Fig. 1b**), we used Amphora2 [55] to identify 31 highly-conserved proteins in each genome and to align them, we concatenated the 31 protein alignments, and we used FastTree 2.1.8 [56] to infer a tree.

Specific phenotypes. We defined a specific phenotype for a gene in an experiment as: |fitness| > 1 and $|t|$ > 5 in this experiment; |fitness| < 1 in at least 95% of experiments; and the fitness value in this experiment is noticeably more extreme than most of its other fitness values (|fitness| > 95th percentile(|fitness|) + 0.5). Our definition of specific phenotypes is sensitive to the conditions that we profiled. For example, an amino acid biosynthetic gene with an auxtrophic phenotype in defined media will likely not be identified with a specific phenotype in any of our experiments, as such a gene is expected to have a severe fitness defect in nearly all defined media fitness assays. To minimize this issue for the stress experiments, we chose stress compounds based on the variable genome-wide mutant fitness patterns they elicited from a prior study [3].

To demonstrate the reliability of specific phenotypes on a broader scale, we asked how often genes with specific phenotypes for the utilization of carbon sources could be assigned to their annotated SEED subsystems [35] (**Supplementary Table 21**). Across 32 carbon sources, we identified specific-important phenotypes (a specific phenotype with fitness < 0) for 643 genes that are linked to any SEED subsystem, and 388 of these (60%) were linked to the expected subsystem, which is far higher than the 1% that would be expected by chance ($P < 10^{-15}$, Fisher's exact test).

We considered a specific phenotype to be conserved if a potential ortholog had a specific phenotype with the same sign in a similar experiment with the same carbon source, nitrogen source, or stressful compound (but not necessarily using the same base media or the same concentration of the compound). For specific-important phenotypes (fitness < -1), we also considered a specific phenotype to be conserved if a potential ortholog had fitness < -1 and $t < -4$ in a similar experiment, regardless of how many experiments the ortholog had a phenotype in.

Across our entire dataset, many of the specific phenotypes were conserved, which supports their functional relevance. For specific-important phenotypes, if a putative ortholog had a fitness measurement in the corresponding condition, then 34% of the time, the ortholog had fitness < -1. Even for orthologs in the same genus, the phenotype was conserved just 46% of the time, so we believe that the lack of conservation is usually due to indirect phenotypes rather than due to the orthologs having different functions. For strong specific-important phenotypes (fitness < -2) in defined media conditions, the conservation rate rose to 54% (or 72% within a genus); it appears that strong and specific phenotypes in defined media are particularly likely to indicate a direct association. In contrast, the conservation rate of specific and detrimental phenotypes (fitness > 1) was just 6%, which suggests that most of these associations are indirect. So, we only consider a detrimental association to be conserved if an ortholog meets the criteria for a specific (and detrimental) phenotype in that condition.

Cofitness and conserved cofitness. We calculate cofitness (r) as the Pearson correlation of all of fitness values for a pair of genes within a single bacterium [3,4]. In contrast, conserved cofitness is identified by comparing cofitness scores for pairs of orthologous genes from two bacteria. Conserved cofitness is defined as minimum(cofitness in the source genome, max(cofitness of orthologs in other genomes)). In other words, a pair of genes is conserved cofit at 0.6 if cofitness > 0.6 in the source genome and any orthologous pair had cofitness > 0.6. Note that the calculations of cofitness and conserved cofitness do not take into account the experimental conditions, therefore it is common for cofit gene pairs to have different phenotypes in the two bacteria. We also note that the number of cofitness associations in our data is far higher than if the data were just noise. For example, we resampled the gene fitness values from the control comparisons (between "start" samples) to make a new dataset with the same number of experiments in each bacterium but with no biological signal. With these data, there were no cases of conserved cofitness above 0.6 and there were just 4 genes with cofitness in a bacterium above 0.8.

Our fitness calculations for stress experiments do not correct for the fitness values in the unstressed condition. This leads to the possibility of cofitness between genes that are important for fitness in every condition but might not have any functional relationship. However, even genes that are mildly important for fitness in most conditions often have different phenotypes in

a subset of conditions. For example, **Extended Data Figure 1b** compares the gene fitness values for *Shewanella loihica* PV-4 growing in LB or in LB with 0.8 mM zinc sulfate. There are 26 genes that have fitness < -2 in both conditions, and as you might expect, these genes are important for fitness in most of the LB stress experiments. So one might expect these genes to give spurious results for cofitness. However, of these 26 genes, 13 are significantly detrimental to fitness in some condition (fitness > 0). Even the remaining 13 genes seem to be important in most, but not all conditions. For example, Shew_1093 (peroxide stress resistance protein *yaaA*) is important for fitness in most conditions but has little or no defect with alpha-ketoglutarate as the carbon source (2 replicates, fitness = -0.3 and -0.4). Of the 13 genes that lack a detrimental phenotype, none have high-scoring cofitness (r > 0.8) with any gene in *S. loihica* PV-4. Four of these genes did meet our criteria for conserved cofitness, and three of the four cases are biologically plausible (*ruvC* with *recG*, which is also involved in recombinational DNA repair; *recG* with *ruvC*; and *gspL* with other *gsp* genes). More broadly, across 30 bacteria there are 1,797 protein-coding genes in our dataset that are important for fitness in our base media (average fitness < -1). (For the two other bacteria, we do not have fitness data from the base media.) Of those 1,797 genes, 758 have significant detrimental phenotypes in other conditions, so we do not expect spurious cofitness. Of the remaining 1,039 protein-coding genes, just 343 (33%) met our criteria for a functional association based on cofitness. As illustrated by our analysis of the 4 such cases in PV-4, we expect that many of these associations are genuine. Therefore, we believe that very few of the thousands of cofitness associations we identified are due to consistently deleterious and functionally unrelated gene pairs.

Comparison to other ways of computing cofitness. A related caveat in interpreting the cofitness values is that, if some genes are moderately important for fitness in every condition, they will have high cofitness due to the varying number of generations in the fitness assays, even if there is no functional relationship. This problem could be avoided if the fitness values were divided by the number of generations that the cells grew for, but we prefer not to rescale the fitness values in this way for several reasons. First, it is not clear what the rescaling should be for motility or survival experiments. Second, we expect that the rescaling would amplify the experimental noise in experiments with fewer generations. Third, any genes whose mutants have strong fitness defects should be absent from our dataset anyway, as these mutants should be at very low abundance after the construction of the mutant library and outgrowth from the freezer.

We tested the impact of rescaling the fitness values before computing cofitness for six bacteria (*C. crescentus* N1000, *E. vietnamensis*, *D. japonica* UNC79MFTsu3.2, *K. michiganensis* M5a1, *P. actiniarum,* and *Pedobacter* sp. GW460-11-11-14-LB5). For these bacteria, we have accurate measurements of the total number of generations of growth for every successful fitness assay. Across these six bacteria, if we consider the 2,905 query genes whose most-cofit gene had cofitness above 0.8 (without rescaling), then 95% of these pairs had cofitness above 0.8 after rescaling. Also, in all but 3 cases, the top cofitness hit was still in the top 5 hits after rescaling the cofitness values. The distributions of cofitness values for functionally-related gene pairs (same TIGR subrole) or unrelated pairs (different subrole) were also very similar with or without rescaling: for each organism and each set of gene pairs, the Kolmogorov-Smirnov distance (D) between the two distributions was between 0.01 and 0.02. (D is a non-parametric measure of the dissimilarity between two distributions and ranges from 0 for identical distributions to 1 for

distributions that do not overlap.) Overall, the cofitness values were very similar regardless of whether the fitness values were scaled by the number of generations or not (**Extended Data Fig. 10a**).

Also note that we compute cofitness across all experiments, without averaging the replicates. To quantify the impact of averaging the biological replicate fitness experiments on the cofitness values, we computed, for each of the 32 bacteria, the Kolmogorov-Smirnov distance (D) between the distribution of cofitness values for functionally unrelated genes (with different TIGR subroles) as computed with replicates averaged or not. For all 32 bacteria, $D < 0.08$. In other words, the distributions of cofitness are quite similar whether or not the biological replicates are averaged. We also verified that the vast majority of the top cofitness hits were still high-scoring hits if we averaged the replicates before computing cofitness. Of the top cofitness hits with $r >= 0.8$, 95% were in the top five hits and above 0.8 if we used replicate-averaged cofitness. Another 3% of these cases had replicate-averaged cofitness in the top 5 hits and between 0.7 and 0.8. Overall, the cofitness values were very similar regardless of whether replicate experiments were averaged or not (**Extended Data Fig. 10b**). In our dataset, 99% of bacterium x condition x concentration combinations have 1 to 5 replicates, which may explain why averaging the replicates did not have much effect on the cofitness values.

Functional associations and conserved functional associations. In a single bacterium, a gene has a functional association if it has a specific phenotype or high cofitness with another gene. A gene has a conserved functional association if it has a conserved specific phenotype or conserved cofitness.

"Predicting" TIGR subroles from cofitness or conserved cofitness. We only considered cofitness hits that were in the top 10 hits for a query gene, only hits with cofitness above 0.4 (or conserved cofitness above 0.4), and only hits that were at least 20 kilobases from the query gene. For this analysis, we did not consider nearby genes in the genome because (1) our data has subtle biases from chromosomal position [9], which can inflate cofitness [57]; (2) nearby genes may be in the same operon and may be spuriously cofit because of polar effects (i.e., when a transposon insertion in an upstream gene of an operon blocks transcription of the downstream genes); (3) functional relationships between nearby genes can be identified by comparative genomics [58] and may be less useful. We predict that the gene has the same subrole as the best-scoring hit. When testing cofitness, the hits with the highest cofitness were considered first. When testing conserved cofitness, the hits with the highest conserved cofitness (as defined above) were considered first.

Genome and gene annotations. For previously-published genomes, gene annotations were taken from MicrobesOnline, Integrated Microbial Genomes (IMG), or RefSeq. Newly-sequenced genomes were annotated with RAST [59], except that *S. koreensis* DSMZ 15582 was annotated by IMG and *P. fluorescens* FW300-N2C3 and *P. fluorescens* FW300-N2E3 were annotated by the NCBI pipeline. See **Supplementary Table 22** for a summary of genome annotations and their accession numbers.

Statistical analysis. For **Figure 2b**, we chose a sample size of 40 genes with non-conserved phenotypes to limit the manual effort for the analysis. We expected a large difference in the

proportions such as 80% versus 40%, and the power to detect such a difference was over 98% for alpha = 0.05.

Classification of how informative annotations are. To assess the existing computational annotations for the 32 genomes, we classified all of their proteins into one of four groups: detailed TIGR role, hypothetical, vague, or other detailed. (1) “Detailed TIGR role” includes proteins that belong to a TIGRFAM role other than "Unclassified", "Unknown function", or "Hypothetical proteins" and had a subrole other than "Unknown substrate", "Two-component systems", "Role category not yet assigned", "Other", "General", "Enzymes of unknown specificity", "Domain", or "Conserved". (2) “Hypothetical” includes proteins whose annotation contained the phrase "hypothetical protein", "unknown function", "uncharacterized", or if the entire description matched “TIGRnnnnn family protein” or “membrane protein”. (3) Proteins were considered to have “vague” annotations if the gene description contained "family", "domain protein", "related protein", "transporter related", or if the entire description matched common non-specific annotations ("abc transporter atp-binding protein", "abc transporter permease", "abc transporter substrate-binding protein", "abc transporter", "acetyltransferase", "alpha/beta hydrolase", "aminohydrolase", "aminotransferase", "atpase", "dehydrogenase", "dna-binding protein", "fad-dependent oxidoreductase", "gcn5-related n-acetyltransferase", "histidine kinase", "hydrolase" "lipoprotein", "membrane protein", "methyltransferase", "mfs transporter", "oxidoreductase", "permease", "porin", "predicted dna-binding transcriptional regulator", "predicted membrane protein", "probable transmembrane protein", "putative membrane protein", "response regulator receiver protein", "rnd transporter", "sam-dependent methyltransferase", "sensor histidine kinase", "serine/threonine protein kinase", "signal peptide protein", "signal transduction histidine kinase", "tonb-dependent receptor", "transcriptional regulator", "transcriptional regulators", or "transporter"). The remaining proteins were considered to have “other detailed” annotations.

To identify a subset of the proteins annotated as “hypothetical” or “vague” that do not belong to any characterized families, we relied on Pfam and TIGRFAM. A Pfam family was considered to be uncharacterized if its name began with either DUF or UPF (which is short for "uncharacterized protein family"). A TIGRFAM family was considered to be uncharacterized if it had no link to a role or if the top-level role was "Unknown function". To identify poorly-annotated proteins from diverse bacteria (for **Extended Data Fig. 9**), we used the rules for vague annotations only.

Protein re-annotation. We manually examined the genes with specific-important phenotypes (fitness < -1 and met criteria for a specific phenotype) on carbon sources in 25 bacteria (we excluded *C. crescentus* NA1000, *D. japonica* UNC79MFTsu3.2, *E. vietnamensis*, *H. seropedicae* SmR1, *K. michiganensis* M5a1, *Pedobacter* sp. GW460-11-11-14-LB5, and *P. actiniarum*). We focused on genes with strongly important phenotypes (fitness < -2) whose annotations (in SEED and/or KEGG) did not imply a role in utilizing that nutrient. In some cases, we also tried to identify the complete catabolic pathway and used cofitness to find additional proteins that were involved but did not meet the threshold for a specific phenotype. For putative transporters, we also considered strong specific-important phenotypes (fitness < -2) on nitrogen sources. The primary tools that we used were the Fitness Browser (which includes Pfam [17], TIGRFAM [16], SEED [35], and the public release of KEGG [34]), MetaCyc [60], PaperBLAST [61], and the Conserved Domains Database (CDD) [62]. (We only used SEED/RAST for the initial annotation of

a subset of the genomes in this study, but the Fitness Browser stores the SEED/RAST annotation of every protein sequence.) For proteins that appeared to lack correct annotations in the public release of KEGG, we checked the KEGG web site to see if the annotation had been updated.

Data and code availability. The Fitness Browser (http://fit.genomics.lbl.gov) provides access to information from the successful fitness experiments, including details of the experimental conditions, quality metrics for each experiment, per-strain fitness values, gene fitness scores, and *t* values. This Fitness Browser is available at http://fit.genomics.lbl.gov and is archived at https://doi.org/10.6084/m9.figshare.5134840. (This archive includes a SQLite relational database, a BLAST database of protein sequences, and tab-delimited files.) The Fitness Browser also contains the original gene annotations for each of the 32 bacteria we studied; the predicted protein sequences for the annotated genes; the results of various comparative genomics analyses (PFam, TIGRFam, SEED/RAST, and KEGG); and re-annotations based on the fitness data (**Supplementary Note 5** and **Supplementary Table 12**). The Fitness Browser also includes functional associations, but as more fitness experiments are conducted, the functional associations in the Fitness Browser may diverge from the analyses in this manuscript.

Most analyses in this manuscript are available from http://genomics.lbl.gov/supplemental/bigfit. (archived at https://doi.org/10.6084/m9.figshare.5134837). The R image contains all of the information in the Fitness Browser except for the per-strain fitness values. The R image also reports: how many insertions were found in each protein-coding gene; the likely-essential genes; metadata and quality metrics for the unsuccessful experiments; gene fitness values from control (start) experiments; which genes had statistically significant phenotypes; the conserved functional associations; the classification of how informative each protein's annotation is; the identifiers and BLAST scores for the proteins with hypothetical or vague annotations from diverse bacteria; and the cofitness of pairs of genes with the same TIGR subroles or different TIGR subroles. The web site and the archive also include: genome sequences; the mapping between the DNA barcode and the insertion location for each pool of transposon mutants; and which mutant strains were used for computing gene fitness. The analysis of cofitness values based on scaling by the number of generations (or not) is archived separately (https://doi.org/10.6084/m9.figshare.5146309.v1).

The BarSeq or TnSeq reads were analyzed with the RB-TnSeq scripts (https://bitbucket.org/berkeleylab/feba); we used statistics versions 1.0.3, 1.1.0, or 1.1.1 of the code; and version 32x1 of BLAT [63]. The R image was derived from these results using R code that is included in the archive.

All genomes sequenced for this study have been deposited in GenBank under the following accession numbers: *Acidovorax* sp. GW101-3H11 (LUKZ01000000), *Pseudomonas fluorescens* FW300-N1B4 (LUKJ01000000), *Pseudomonas fluorescens* FW300-N2C3 (CP012831.1), *Pseudomonas fluorescens* FW300-N2E2 (SAMN03294340), *Pseudomonas fluorescens* FW300-N2E3 (CP012830.1), *Pseudomonas fluorescens* GW456-L13 (LKBJ00000000.1), *Sphingomonas koreensis* DSMZ 15582 (PGEN01000001.1).

Strain availability. Most of the wild-type bacteria (25 of 32), including the 8 new isolates from the ORNL FRC, are available from public stock centers. The 8 new isolates are available from

DSMZ under the following accession numbers: *Acidovorax* sp. GW101-3H11 (DSM 106133), *Cupriavidus basilensis* 4G11 (DSM 106286), *Pedobacter* sp. GW460-11-11-14-LB5 (DSM 106514), *Pseudomonas fluorescens* FW300-N1B4 (DSM 106120), *Pseudomonas fluorescens* FW300-N2C3 (DSM 106121), *Pseudomonas fluorescens* FW300-N2E2 (DSM 106119), *Pseudomonas fluorescens* FW300-N2E3 (DSM 106124), *Pseudomonas fluorescens* GW456-L13 (DSM 106123). The other 7 bacteria were gifts from individuals (**Supplementary Table 14)**. Wild-type bacteria, mutant libraries, and the individual mutant strains that we generated for follow-up studies are available upon request.

Extended data figure legends

Extended Data Figure 1. Examples of nitrogen source and stress fitness experiments. (a) The utilization of D-alanine or cytosine by *Azospirillum brasilense* sp. 245. Each point shows the fitness of a gene in the two conditions. The data is the average of two biological replicates for each nitrogen source. Amino acid synthesis genes were identified using the top-level role in TIGRFAMs. The genes for D-alanine utilization were a D-amino acid dehydrogenase (AZOBR_RS08020), an ABC transporter operon (AZOBR_RS08235:RS08260), and a LysR family regulator (AZOBR_RS21915). The genes for cytosine utilization were cytosine deaminase (AZOBR_RS31895) and two ABC transporter operons (AZOBR_RS06950:RS06965 and AZOBR_RS31875:RS31885). (b) Zinc stress in *Shewanella loihica* PV-4. We compare fitness in rich media with added zinc (II) sulfate to fitness in plain rich media. The LB data is the average of two biological replicates. The highlighted genes include a putative heavy metal efflux pump (CzcCBA or Shew_3358:Shew_3356), a hypothetical protein at the beginning of the *czc* operon (CzcX), a zinc-responsive regulator (ZntR or Shew_3411), and another heavy metal efflux gene related to *arsP* or DUF318 (Shew_3410). *CzcX* lacks homology to any characterized protein, but homologs in other strains of *Shewanella* are also specifically important for resisting zinc stress. In both panels, the lines show x = 0, y = 0, and x = y.

Extended Data Figure 2. Phenotypes versus types of genes. We categorized proteins in our dataset by their type of annotation or by whether they have homologs in the same genome ("paralogs"). For each category, we show the fraction of genes that have statistically significant phenotypes, and more specifically the fractions that have strong phenotypes (|fitness| > 2 and statistically significant) or are significantly detrimental to fitness (fitness > 0). Genes with high or moderate similarity to another gene in the same genome (paralogs with alignment score above 30% of the self-alignment score) were less likely to have a phenotype (25% vs. 32%, $P < 10^{-15}$, Fisher's exact test), which likely reflects genetic redundancy.

Extended Data Figure 3. Known DNA repair genes are important for cisplatin resistance. We compared the growth of a gene deletion strain and the wild-type bacterium under varying cisplatin concentrations. We show all replicate growth curves for each genotype. We believe the higher overall growth for some of the wild-type experiments (for example, in top middle panel) is random. We observe this phenomenon consistently for some bacteria and we speculate that this is due to varying oxygen content across the microplate. (a) *E. coli radD* (n = 6 independent experiments per strain). (b) *Dinoroseobacter shibae* Dshi_2244 (n = 3 independent experiments for wild-type and n = 6 independent experiments for the mutant). (c) *Phaeobacter inhibens* PGA1_c08960 (n = 4 independent experiments for wild-type and n = 6 independent experiments

for the mutant). Dshi_2244 and PGA1_c08960 are orthologs of MmcB (DUF1052) from *Caulobacter crescentus* [24].

Extended Data Figure 4. EndA, DUF3584, and a FAN1-like VRR-NUC domain protein are important for cisplatin resistance. Same as Extended Data Figure 3, comparing cisplatin sensitivity of a gene deletion mutant to wild-type. (a) *E. coli endA* knockout. *cycA* encodes an amino acid transporter and is not expected to have a phenotype on cisplatin and is used as a control. Each growth curve is the average of 12 replicate wells and the dashed lines show 95% confidence intervals from the *t* test. (b) A deletion of *Shewanella oneidensis* MR-1 SO4008, a member of the DUF3584 protein family (n = 6 independent experiments per strain). (c) A deletion of *Pseudomonas stutzeri* RCH2 Psest_2235 (n = 4 independent experiments per strain). Psest_1636 is not expected to be involved in DNA repair and is used here as a control. Psest_2235 is a FAN1-like VRR-NUC domain protein [25].

Extended Data Figure 5. The nuclease domain of EndA is important for cisplatin resistance. We assayed the growth of an *E. coli endA*- Keio collection deletion mutant carrying one of three different vectors: an empty vector with no insert (endA- + empty), a complementation vector carrying a wild-type copy of *endA* (endA- + endA), and a complementation vector with a mutant version of *endA* with an alanine at position 84 instead of histidine (endA- + mutant endA). A mutation of this conserved histidine residue in a close homolog from *Vibrio vulnificus* is reported to eliminate nearly all nuclease catalytic activity [64]. As a control, we assayed the wild-type, parental *E. coli* strain carrying a vector with no insert (wt + empty). We performed these growth assays on three separate microplates (labeled as Plate #1, #2, #3). n = 3 independent experiments per strain in Plate #1; n = 4 independent experiments per strain in Plates #2 and #3. We added 20 μg/mL of gentamicin to each assay to maintain selection for the plasmids (pBBR1-MCS5 and derivatives). Although the catalytic activity of EndA (endonuclease I) appears to be important for resisting cisplatin, it is not clear how EndA would be involved in DNA repair if it is located in the periplasm, as previously believed [65-67]. We speculate that EndA relocates to the cytoplasm upon DNA damage or that EndA degrades broken DNA that enters the periplasm and would otherwise damage the membrane.

Extended Data Figure 6. Members of protein family UPF0126 are important for growth on glycine. Growth comparison of gene deletion mutants in UPF0126 versus wild-type bacteria in minimal defined media. (a) SO1319 from *Shewanella oneidensis* MR-1, with either ammonium chloride (n = 6 independent experiments per strain) or glycine as the sole source of nitrogen (n = 12 independent experiments per strain). (b) PGA1_c00920 from *Phaeobacter inhibens*, with glycine as the sole source of carbon (n = 8 independent experiments for wild-type and n = 16 independent experiments for the mutant). (c) Psest_1636 from *Pseudomonas stutzeri* RCH2, with either ammonium chloride (n = 4 independent experiments per strain) or glycine (n = 8 independent experiments per strain) as the sole source of nitrogen. The Psest_2235 deletion strain is used as a control and is not expected to have a phenotype in these conditions.

Extended Data Figure 7. PGA1_c00920 partially rescues the glycine growth defect of an *E. coli cycA* mutant. CycA is a glycine transporter from *E. coli* and a mutant in this gene has reduced uptake of glycine [37]. We investigated whether a member of the UPF0126 protein family could rescue the glycine growth defect of an *E. coli cycA* deletion strain. We introduced different plasmids into the *E. coli cycA* Keio collection deletion background: an empty plasmid with no

insert (cycA- + empty), a plasmid with a wild-type allele of the *E. coli cycA* gene (cycA- + cycA), and a plasmid with PGA1_c00920 from *Phaeobacter inhibens* (cycA- + PGA1_c00920). We compared the growth of these strains and a wild-type *E. coli* control (wt + empty) in defined media with either ammonium chloride (n = 2 independent experiments per strain) or glycine as the sole source of nitrogen (n = 4 independent experiments per strain). PGA1_c00920 partially rescues the glycine-specific growth defect of the *cycA*- deletion strain.

Extended Data Figure 8. Overexpression of members of protein family UPF0060 confer resistance to thallium. We introduced three plasmids into wild-type *E. coli*: a plasmid control with no insert (Empty vector), a plasmid carrying RR42_RS34240 from *Cupriavidus basilensis* 4G11, and a plasmid carrying Pf6N2E2_2547 from *Pseudomonas fluorescens* FW300-N2E2. We assayed the growth of these strains in LB at 30°C with varying concentrations of thallium(I) acetate (n = 6 independent experiments per strain). We added 50 µg/mL of kanamycin to each assay to maintain selection for the plasmids (pFAB2286 and derivatives). RR42_RS34240 and Pf6N2E2_2547 are members of the UPF0060 protein family.

Extended Data Figure 9. Relevance to all bacteria. We selected 2,593 hypothetical or vaguely-annotated proteins from diverse bacterial species, we compared them to the protein-coding genes that we have fitness data for (using protein BLAST), and we identified potential orthologs as best hits that were homologous over at least 75% of each protein's length. We show the fraction of these proteins that have a potential ortholog with each type of phenotype and that is above a given level of amino acid sequence similarity. Similarity was defined as the ratio of the alignment's bit score to the score from aligning the query to itself.

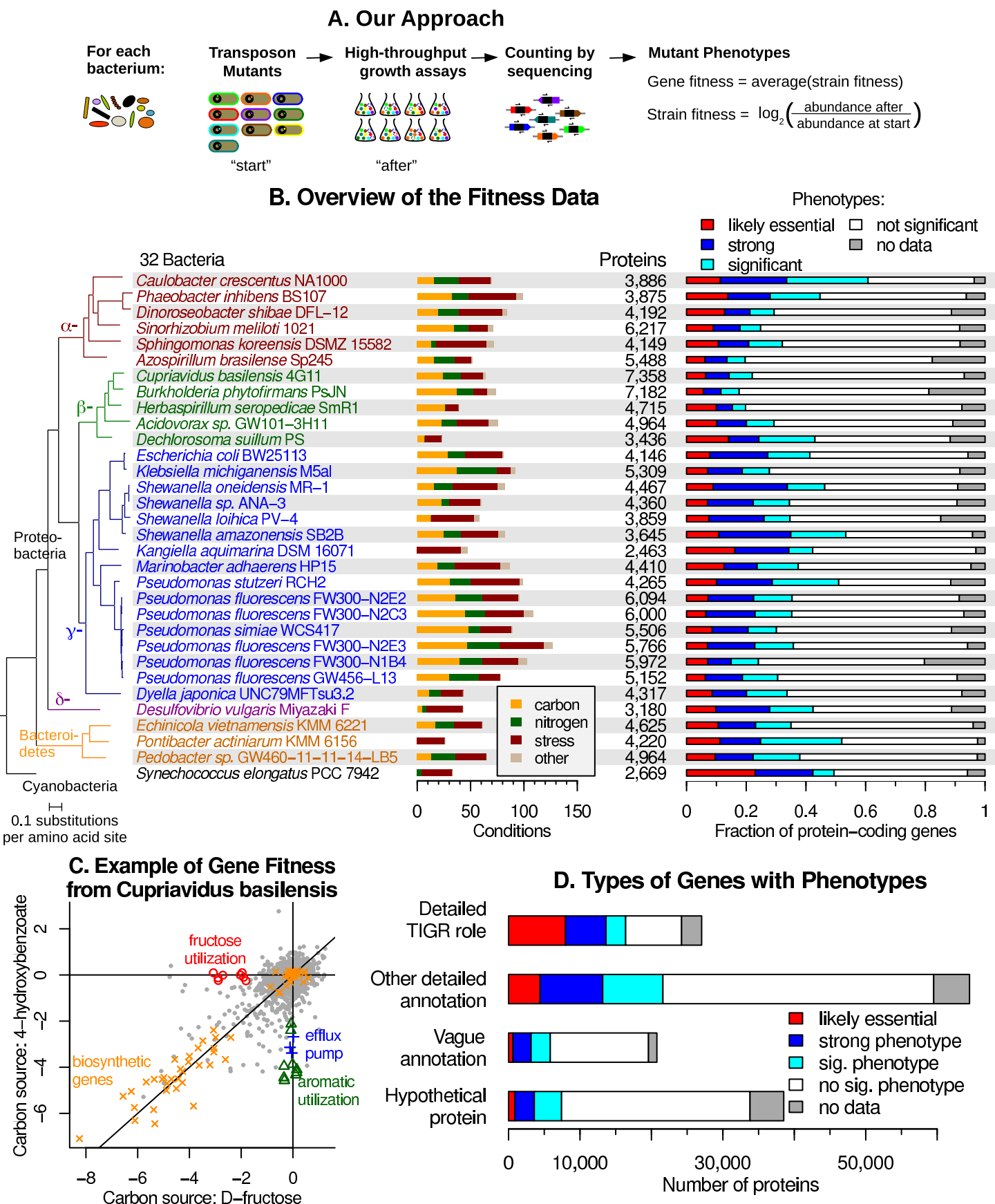
Extended Data Figure 10. Alternative ways of computing cofitness. (a) The effect of rescaling the cofitness values by the number of generations in six bacteria. For each of the six bacteria, we identified all pairs of protein-coding genes that were assigned to the same TIGR subrole, were more than 20 kB apart, and had fitness data. This gave 1,711 to 9,406 pairs per bacterium. We also selected a random subset of pairs that were assigned to different TIGR subroles, were more than 20 kB apart, and had fitness data (1,559 to 8,881 pairs per bacterium). For each pair, we compared the original cofitness values to the rescaled cofitness (computed from fitness values that were divided by the number of generations). (b) The effect of averaging fitness scores from replicate experiments on the cofitness values.

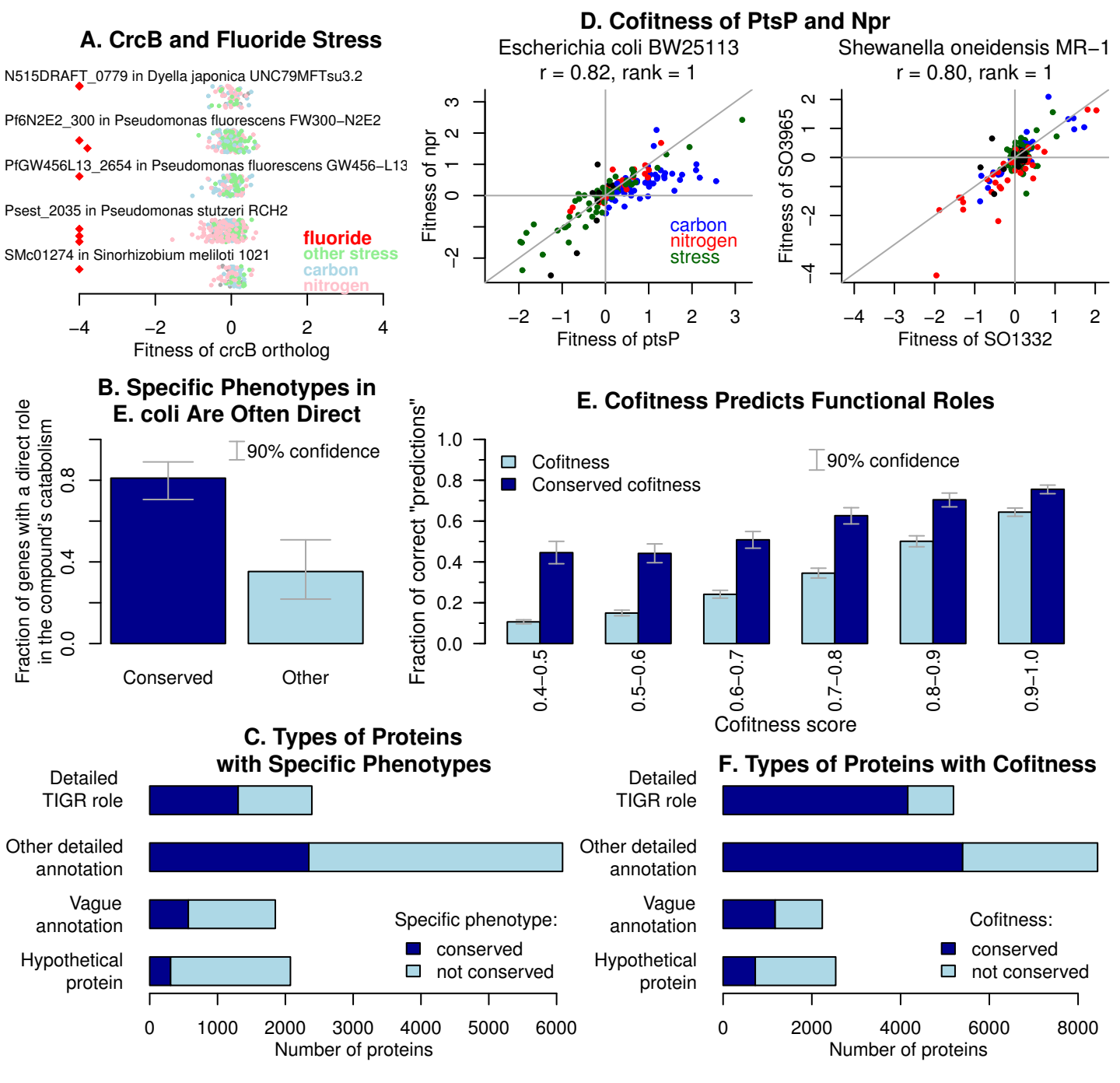
References Methods and Extended Data Figures

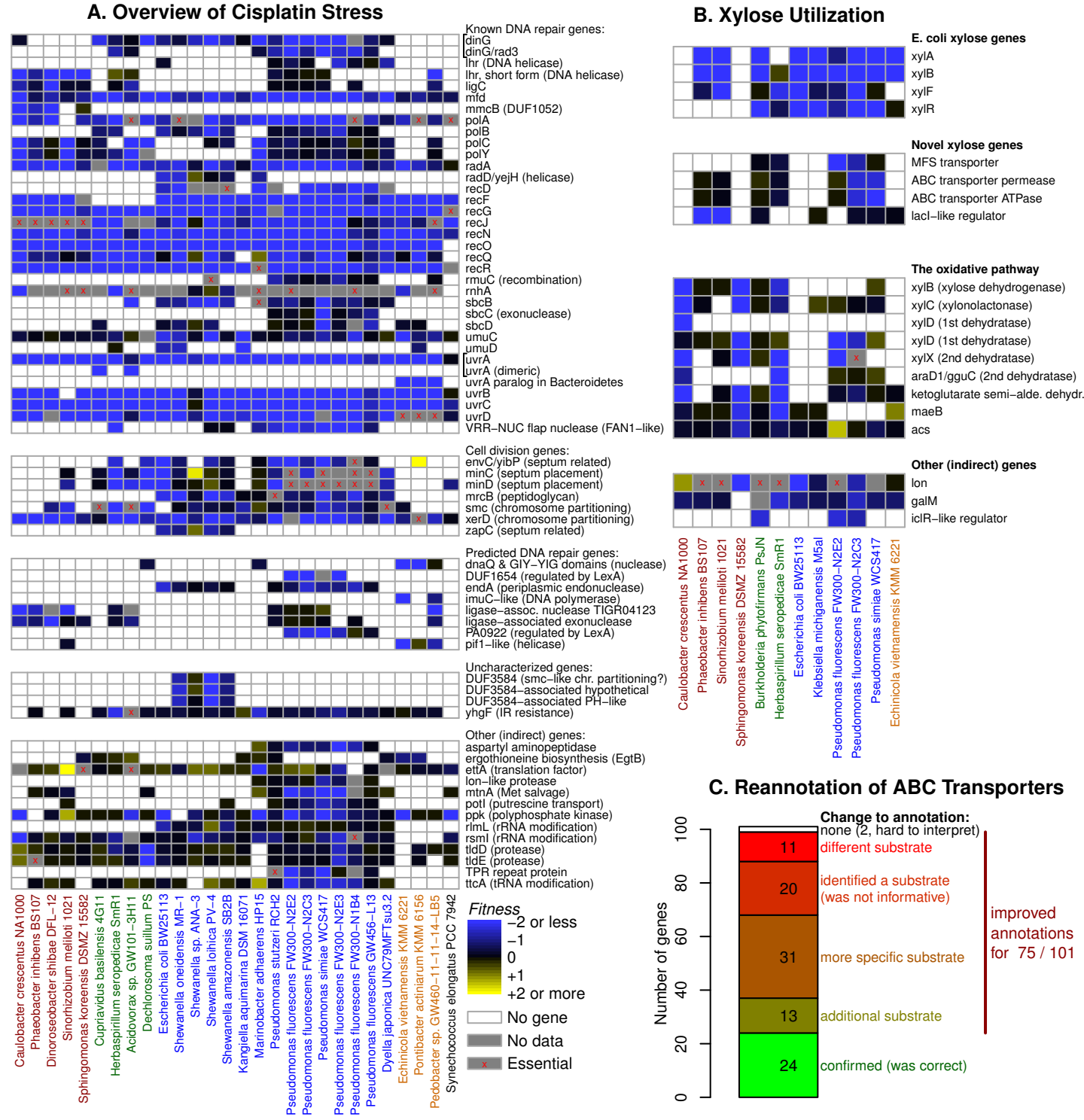
40. Thorgersen, M. P. *et al.* Molybdenum Availability is Key to Nitrate Removal in Contaminated Groundwater Environments. *Appl Environ Microbiol* AEM.00917–15 (2015). doi:10.1128/AEM.00917-15
41. Ray, J. *et al.* Complete Genome Sequence of Cupriavidus basilensis 4G11, Isolated from the Oak Ridge Field Research Center Site. *Genome Announc* **3,** e00322–15 (2015).
42. Vaccaro, B. J. *et al.* Novel Metal Cation Resistance Systems from Mutant Fitness Analysis of Denitrifying Pseudomonas stutzeri. *Appl Environ Microbiol* **82,** 6046–6056 (2016).
43. Kovach, M. E. *et al.* Four new derivatives of the broad-host-range cloning vector

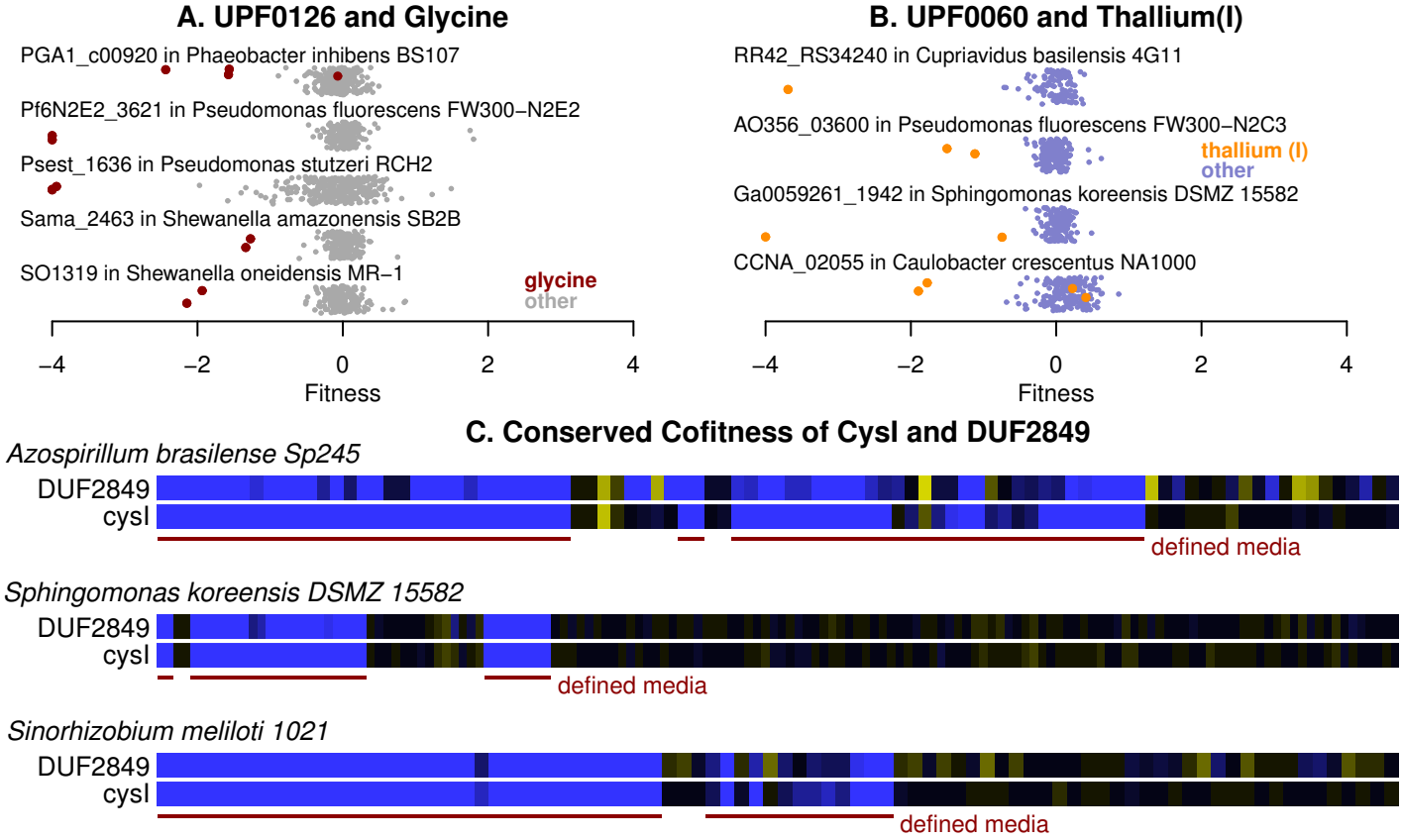
pBBR1MCS, carrying different antibiotic-resistance cassettes. *Gene* **166,** 175–176 (1995).
44. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 2006 0008 (2006).
45. Kuehl, J. V. *et al.* Functional genomics with a comprehensive library of transposon mutants for the sulfate-reducing bacterium Desulfovibrio alaskensis G20. *MBio* **5,** e01041–14 (2014).
46. Zane, G. M., Yen, H. C. & Wall, J. D. Effect of the deletion of qmoABC and the promoter-distal gene encoding a hypothetical protein on sulfate reduction in Desulfovibrio vulgaris Hildenborough. *Appl Environ Microbiol* **76,** 5500–5509 (2010).
47. Kahm, M., Hasenbrink, G., Lichtenberg-Frate, H., Ludwig, J. & Kschischo, M. grofit: Fitting Biological Growth Curves with R. *Journal of Statistical Software* **33,** (2010).
48. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
49. Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.* **30,** 701–707 (2012).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9,** 357–359 (2012).
51. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9,** e112963 (2014).
52. Tritt, A., Eisen, J. A., Facciotti, M. T. & Darling, A. E. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One* **7,** e42304 (2012).
53. Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16,** 294 (2015).
54. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7,** e1002195 (2011).
55. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28,** 1033–1034 (2012).
56. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).
57. Sagawa, S., Price, M. N., Deutschbauer, A. M. & Arkin, A. P. Validating regulatory predictions from diverse bacteria with mutant fitness data. *PLoS One* **12,** e0178258 (2017).
58. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 2896–2901 (1999).
59. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75 (2008).
60. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44,** D471–80 (2016).
61. Price, M. N. & Arkin, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2,** e00039–17 (2017).
62. Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43,** D222–6 (2015).
63. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12,** 656–664 (2002).
64. Li, C.-L. *et al.* DNA binding and cleavage by the periplasmic nuclease Vvn: a novel structure with a known active site. *Embo J* **22,** 4014–4025 (2003).
65. Ananthaswamy, H. N. The release of endonuclease I from Escherichia coli by a new cold

shock procedure. *Biochem. Biophys. Res. Commun.* **76,** 289–298 (1976).
66. Lopes, J., Gottfried, S. & Rothfield, L. Leakage of periplasmic enzymes by mutants of Escherichia coli and Salmonella typhimurium: isolation of "periplasmic leaky" mutants. *J. Bacteriol.* **109,** 520–525 (1972).
67. Nossal, N. G. & Heppel, L. A. The release of enzymes by osmotic shock from Escherichia coli in exponential phase. *J Biol Chem* **241,** 3055–3062 (1966).


A. Our Approach
For each bacterium:
Transposon Mutants
High-throughput growth assays
Counting by sequencing
Mutant Phenotypes
Gene fitness = average(strain fitness)
Strain fitness = log2(abundance after / abundance at start)
"start"
"after"
B. Overview of the Fitness Data
Phenotypes:
likely essential
strong
significant
not significant
no data
32 Bacteria
Proteins
Caulobacter crescentus NA1000 3,886
Phaeobacter inhibens BS107 3,875
Dinoroseobacter shibae DFL–12 4,192
Sinorhizobium meliloti 1021 6,217
Sphingomonas koreensis DSMZ 15582 4,149
Azospirillum brasilense Sp245 5,488
Cupriavidus basilensis 4G11 7,358
Burkholderia phytofirmans PsJN 7,182
Herbaspirillum seropedicae SmR1 4,715
Acidovorax sp. GW101–3H11 4,964
Dechlorosoma suillum PS 3,436
Escherichia coli BW25113 4,146
Klebsiella michiganensis M5al 5,309
Shewanella oneidensis MR–1 4,467
Shewanella sp. ANA–3 4,360
Shewanella loihica PV–4 3,859
Shewanella amazonensis SB2B 3,645
Kangiella aquimarina DSM 16071 2,463
Marinobacter adhaerens HP15 4,410
Pseudomonas stutzeri RCH2 4,265
Pseudomonas fluorescens FW300–N2E2 6,094
Pseudomonas fluorescens FW300–N2C3 6,000
Pseudomonas simiae WCS417 5,506
Pseudomonas fluorescens FW300–N2E3 5,766
Pseudomonas fluorescens FW300–N1B4 5,972
Pseudomonas fluorescens GW456–L13 5,152
Dyella japonica UNC79MFTsu3.2 4,317
Desulfovibrio vulgaris Miyazaki F 3,180
Echinicola vietnamensis KMM 6221 4,625
Pontibacter actiniarum KMM 6156 4,220
Pedobacter sp. GW460–11–11–14–LB5 4,964
Synechococcus elongatus PCC 7942 2,669
α-
β-
γ-
δ-
Proteo-bacteria
Bacteroi-detes
Cyanobacteria
0.1 substitutions per amino acid site
carbon
nitrogen
stress
other
0 50 100 150
Conditions
0 0.2 0.4 0.6 0.8 1
Fraction of protein–coding genes
C. Example of Gene Fitness from Cupriavidus basilensis
fructose utilization
efflux pump
aromatic utilization
biosynthetic genes
Carbon source: 4–hydroxybenzoate
Carbon source: D–fructose
2 0 –2 –4 –6
–8 –6 –4 –2 0
D. Types of Genes with Phenotypes
Detailed TIGR role
Other detailed annotation
Vague annotation
Hypothetical protein
likely essential
strong phenotype
sig. phenotype
no sig. phenotype
no data
0 10,000 30,000 50,000
Number of proteins


A. CrcB and Fluoride Stress
N515DRAFT_0779 in Dyella japonica UNC79MFTsu3.2
Pf6N2E2_300 in Pseudomonas fluorescens FW300−N2E2
PfGW456L13_2654 in Pseudomonas fluorescens GW456−L13
Psest_2035 in Pseudomonas stutzeri RCH2
SMc01274 in Sinorhizobium meliloti 1021
fluoride
other stress
carbon
nitrogen
−4
−2
0
2
4
Fitness of crcB ortholog
D. Cofitness of PtsP and Npr
Escherichia coli BW25113
r = 0.82, rank = 1
Fitness of npr
Fitness of ptsP
carbon
nitrogen
stress
Shewanella oneidensis MR−1
r = 0.80, rank = 1
Fitness of SO3965
Fitness of SO1332
B. Specific Phenotypes in E. coli Are Often Direct
Fraction of genes with a direct role in the compound's catabolism
90% confidence
0.0
0.4
0.8
Conserved
Other
E. Cofitness Predicts Functional Roles
Fraction of correct "predictions"
Cofitness
Conserved cofitness
90% confidence
0.0
0.2
0.4
0.6
0.8
1.0
0.4–0.5
0.5–0.6
0.6–0.7
0.7–0.8
0.8–0.9
0.9–1.0
Cofitness score
C. Types of Proteins with Specific Phenotypes
Detailed TIGR role
Other detailed annotation
Vague annotation
Hypothetical protein
Specific phenotype:
conserved
not conserved
0
1000
2000
3000
4000
5000
6000
Number of proteins
F. Types of Proteins with Cofitness
Detailed TIGR role
Other detailed annotation
Vague annotation
Hypothetical protein
Cofitness:
conserved
not conserved
0
2000
4000
6000
8000
Number of proteins


A. Overview of Cisplatin Stress
Known DNA repair genes:
dinG
dinG/rad3
Ihr (DNA helicase)
Ihr, short form (DNA helicase)
ligC
mfd
mmcB (DUF1052)
polA
polB
polC
polY
radA
radD/yejH (helicase)
recD
recF
recG
recJ
recN
recO
recQ
recR
rmuC (recombination)
rnhA
sbcB
sbcC (exonuclease)
sbcD
umuC
umuD
uvrA
uvrA (dimeric)
uvrA paralog in Bacteroidetes
uvrB
uvrC
uvrD
VRR−NUC flap nuclease (FAN1−like)
Cell division genes:
envC/yibP (septum related)
minC (septum placement)
minD (septum placement)
mrcB (peptidoglycan)
smc (chromosome partitioning)
xerD (chromosome partitioning)
zapC (septum related)
Predicted DNA repair genes:
dnaQ & GIY−YIG domains (nuclease)
DUF1654 (regulated by LexA)
endA (periplasmic endonuclease)
imuC−like (DNA polymerase)
ligase−assoc. nuclease TIGR04123
ligase−associated exonuclease
PA0922 (regulated by LexA)
pif1−like (helicase)
Uncharacterized genes:
DUF3584 (smc−like chr. partitioning?)
DUF3584−associated hypothetical
DUF3584−associated PH−like
yhgF (IR resistance)
Other (indirect) genes:
aspartyl aminopeptidase
ergothioneine biosynthesis (EgtB)
ettA (translation factor)
lon−like protease
mtnA (Met salvage)
potI (putrescine transport)
ppk (polyphosphate kinase)
rlmL (rRNA modification)
rsmI (rRNA modification)
tldD (protease)
tldE (protease)
TPR repeat protein
ttcA (tRNA modification)
Caulobacter crescentus NA1000
Phaeobacter inhibens BS107
Dinoroseobacter shibae DFL−12
Sinorhizobium meliloti 1021
Sphingomonas koreensis DSMZ 15582
Cupriavidus basilensis 4G11
Herbaspirillum seropedicae SmR1
Acidovorax sp. GW101−3H11
Dechlorosoma suillum PS
Escherichia coli BW25113
Shewanella oneidensis MR−1
Shewanella sp. ANA−3
Shewanella loihica PV−4
Shewanella amazonensis SB2B
Kangiella aquimarina DSM 16071
Marinobacter adhaerens HP15
Pseudomonas stutzeri RCH2
Pseudomonas fluorescens FW300−N2E2
Pseudomonas fluorescens FW300−N2C3
Pseudomonas simiae WCS417
Pseudomonas fluorescens FW300−N2E3
Pseudomonas fluorescens FW300−N1B4
Pseudomonas fluorescens GW456−L13
Dyella japonica UNC79MFTsu3.2
Echinicola vietnamensis KMM 6221
Pontibacter actiniarum KMM 6156
Pedobacter sp. GW460−11−11−14−LB5
Synechococcus elongatus PCC 7942
Fitness
−2 or less
−1
0
+1
+2 or more
No gene
No data
Essential
B. Xylose Utilization
E. coli xylose genes
xylA
xylB
xylF
xylR
Novel xylose genes
MFS transporter
ABC transporter permease
ABC transporter ATPase
lacI−like regulator
The oxidative pathway
xylB (xylose dehydrogenase)
xylC (xylonolactonase)
xylD (1st dehydratase)
xylD (1st dehydratase)
xylX (2nd dehydratase)
araD1/gguC (2nd dehydratase)
ketoglutarate semi−alde. dehydr.
maeB
acs
Other (indirect) genes
lon
galM
iclR−like regulator
Caulobacter crescentus NA1000
Phaeobacter inhibens BS107
Sinorhizobium meliloti 1021
Sphingomonas koreensis DSMZ 15582
Burkholderia phytofirmans PsJN
Herbaspirillum seropedicae SmR1
Escherichia coli BW25113
Klebsiella michiganensis M5al
Pseudomonas fluorescens FW300−N2E2
Pseudomonas fluorescens FW300−N2C3
Pseudomonas simiae WCS417
Echinicola vietnamensis KMM 6221
C. Reannotation of ABC Transporters
Change to annotation:
none (2, hard to interpret)
different substrate
identified a substrate (was not informative)
more specific substrate
additional substrate
confirmed (was correct)
11
20
31
13
24
improved annotations for 75 / 101
Number of genes
0
20
40
60
80
100


A. UPF0126 and Glycine
PGA1_c00920 in Phaeobacter inhibens BS107
Pf6N2E2_3621 in Pseudomonas fluorescens FW300−N2E2
Psest_1636 in Pseudomonas stutzeri RCH2
Sama_2463 in Shewanella amazonensis SB2B
SO1319 in Shewanella oneidensis MR−1
glycine
other
−4
−2
0
2
4
Fitness
B. UPF0060 and Thallium(I)
RR42_RS34240 in Cupriavidus basilensis 4G11
AO356_03600 in Pseudomonas fluorescens FW300−N2C3
thallium (I)
other
Ga0059261_1942 in Sphingomonas koreensis DSMZ 15582
CCNA_02055 in Caulobacter crescentus NA1000
−4
−2
0
2
4
Fitness
C. Conserved Cofitness of CysI and DUF2849
Azospirillum brasilense Sp245
DUF2849
cysI
defined media
Sphingomonas koreensis DSMZ 15582
DUF2849
cysI
defined media
Sinorhizobium meliloti 1021
DUF2849
cysI
defined media

Supplementary Information for "Mutant Phenotypes for Thousands of Bacterial Genes of Unknown Function"

Morgan N. Price[1], Kelly M. Wetmore[1], R. Jordan Waters[2], Mark Callaghan[1], Jayashree Ray[1], Hualan Liu[1], Jennifer V. Kuehl[1], Ryan A. Melnyk[1], Jacob S. Lamson[1], Yumi Suh[1], Hans K. Carlson[1], Zuelma Esquivel[1], Harini Sadeeshkumar[1], Romy Chakraborty[3], Grant M. Zane[4], Benjamin E. Rubin[5], Judy D. Wall[4], Axel Visel[2,6], James Bristow[2], Matthew J. Blow[2,*], Adam P. Arkin[1,7,*], Adam M. Deutschbauer[1,8,*]

[1]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory
[2]Joint Genome Institute, Lawrence Berkeley National Laboratory
[3]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory
[4]Department of Biochemistry, University of Missouri
[5]Division of Biological Sciences, University of California, San Diego
[6]School of Natural Sciences, University of California, Merced
[7]Department of Bioengineering, University of California, Berkeley
[8]Department of Plant and Microbial Biology, University of California, Berkeley
[*]To whom correspondence should be addressed:
MJB (MJBlow@lbl.gov)
APA (APArkin@lbl.gov)
AMD (AMDeutschbauer@lbl.gov)

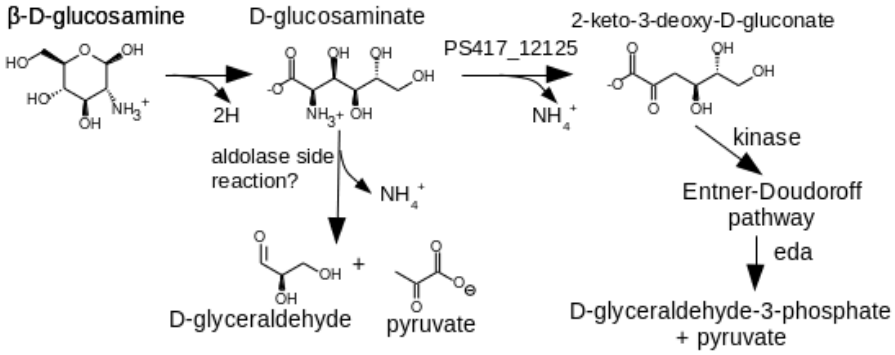
Website for interactive analysis of mutant fitness data:
http://fit.genomics.lbl.gov/

Website with supplementary information and bulk data downloads:
http://genomics.lbl.gov/supplemental/bigfit/

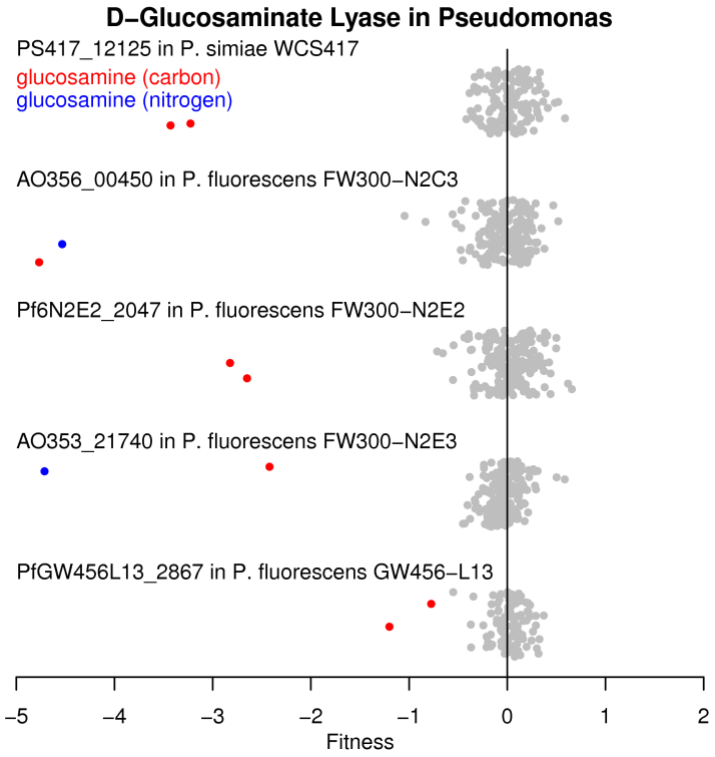
Contents

In separate excel file:

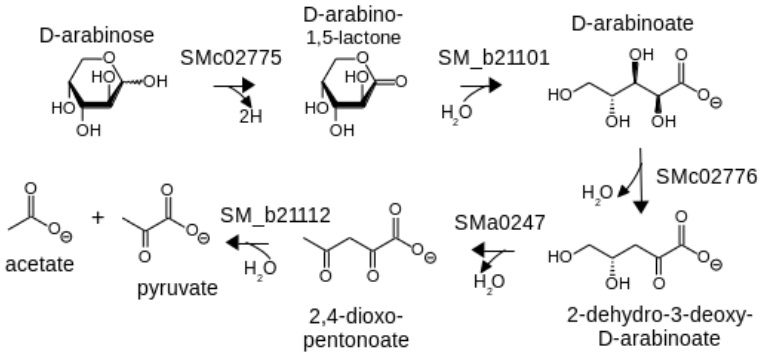
a



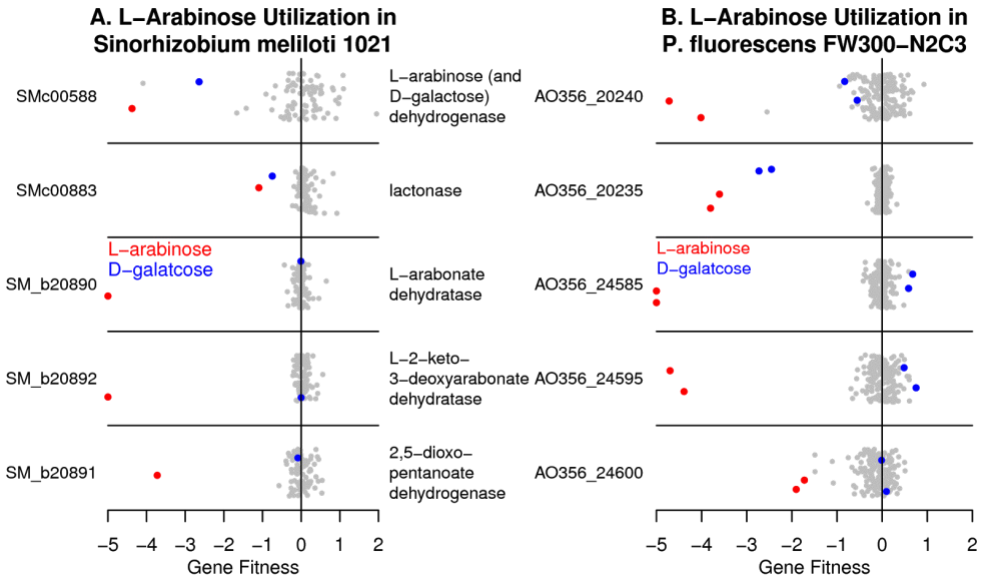
b



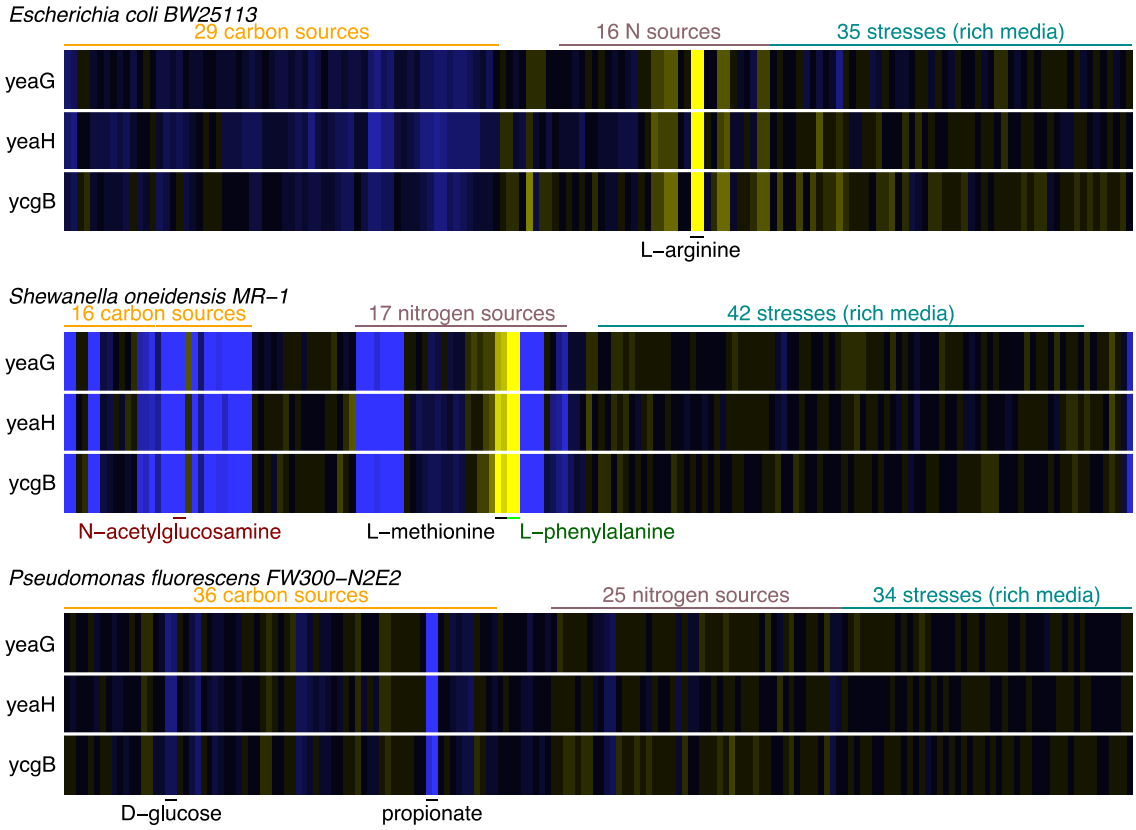
Supplementary Figure 1. Identification of the putative gene encoding D-glucosaminate ammonia-lyase. (a) The pathway for D-glucosamine catabolism. The proposed D-glucosaminate ammonia-lyase from *P. simiae* WCS417 (PS417_12125) is highlighted. (b) We show the fitness data for 5 orthologous genes from different Pseudomonads. Each point represents the fitness value for the gene in a given condition, with experiments using D-glucosamine as the sole source of carbon or nitrogen highlighted. The y-axis is arbitrary for each gene.



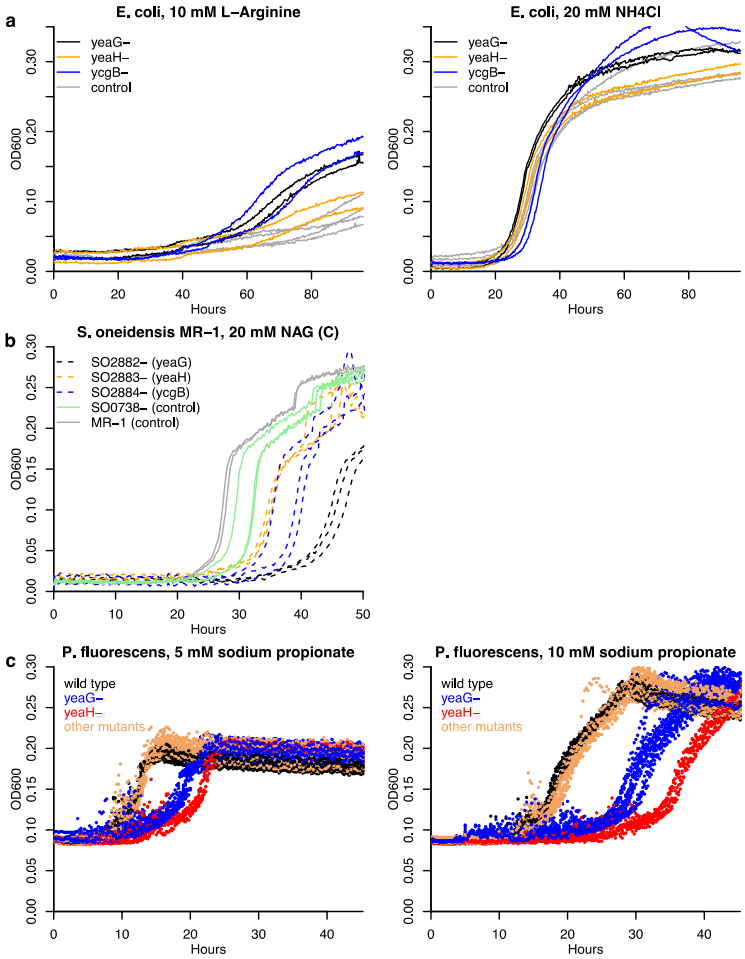
Supplementary Figure 2. The proposed oxidative pathway for D-arabinose catabolism in *S. meliloti*.



Supplementary Figure 3. L-arabinose catabolic genes. Mutant fitness for genes involved in L-arabinose utilization in (a) *S. meliloti* and (b) *P. fluorescens* FW300-N2C3. Each point represents the fitness of a gene under an experimental condition that we profiled for that bacterium, with L-arabinose and D-galactose carbon source utilization experiments highlighted. The y-axis is arbitrary for each gene.



Supplementary Figure 4. Conserved cofitness of YeaG, YeaH, and YcgB. Heatmap of fitness data for *yeaG*, *yeaH*, and *ycgB* from three bacteria. Certain experimental conditions with statistically significant phenotypes are highlighted. The color scale is the same as in Fig. 3a.



Supplementary Figure 5. Growth of signaling mutants. We grew mutants of *yeaG*, *yeaH*, and *ycgB* from different bacteria under conditions identified by the genome-wide mutant fitness assays. (a) We grew deletion strains from the Keio collection [1] in M9 glucose media with varying nitrogen sources. Control mutants are deletions of two pseudogenes, *agaA* or *ygaY*. The signaling mutants had a strong growth advantage on L-arginine, as expected from fitness assays with 10 mM L-arginine as the nitrogen source. n = 2 independent experiments per strain. (b) *S. oneidensis* MR-1 mutants grown in a defined medium (ShewMM_noCarbon) with 20 mM N-acetylglucosamine (NAG) as the carbon source. These mutants have transposons inserted within the signaling pathway (SO2882:SO2884) or in a pseudogene (SO0738, identified as such by Romine and colleagues [2]). Insertions in the signaling pathway had an increased lag and a slower growth rate, as expected from fitness assays in this condition. The mutants are from a previously described collection [3] and were generated independently from the mutants that were used to generate the fitness data. n = 2 or 3 independent experiments per strain. (c) We grew gene deletion mutants of Pf6N2E2_4800 (*yeaH*) and Pf6N2E2_4801 (*yeaG*) from *Pseudomonas fluorescens* FW300-N2E2 in defined media with varying concentrations of sodium propionate as the sole source of carbon. Compared to wild-type *P. fluorescens* FW300-N2E2, the mutants have a growth defect on sodium propionate. The “other mutants” are gene deletions strains of Pf6N2E2_1270 and Pf6N2E2_2534, neither of which is expected to have a phenotype in this condition. n = 4 or more independent experiments per strain.

Supplementary Note 1. Validation of likely-essential genes

Our approach to identify essential or nearly-essential genes was initially validated for *Synechococcus elongatus* [4]. To verify that this approach gave reasonable results for other bacteria, we compared our list of putatively essential protein-coding genes in *E. coli* K-12 to genes that are essential [5] or important for growth in LB [1,6]. We also compared the list of essential genes in *Shewanella oneidensis* MR-1 to a previous analysis based on a different set of transposon mutants [3]. Finally, for all of the genomes, we examined the functional categories according to TIGRFAMs [7] and whether essentiality was conserved.

By combining the profiling of *E. coli* chromosome database (PEC) and the results of systematically attempting to delete every gene in *E. coli* [1,6], we obtained a list of 330 *E. coli* genes that were previously reported to be essential. 257 of these 330 genes (78%) were in our list of likely-essential genes from TnSeq analysis. Another 13 of the known-essential genes are under 250 nucleotides and were too short to be predicted to be essential by the criteria of our TnSeq analysis. Our list of essential genes also includes 67 non-essential genes, which corresponds to a false discovery rate (FDR) of 21%. However, some of these "false positives" are likely to be essential, or nearly so, in the condition that we used to isolate our mutant library, namely growth on LB plates. For example, our data suggested that the lysyl-tRNA synthetase encoded by *lysS* is essential; a mutant in this gene grows slowly in LB [8]. As another example, our data suggested that *cydABCD*, which are required for the activity of cytochrome bd oxidase, are all essential. The Keio gene deletion project [1,6] was able to delete *cydB* and *cydD* but was unable to delete *cydA* or *cydC* (despite eight attempts for each). None of these genes is truly essential, but the growth of mutants in any of these genes is inhibited by the presence of other strains of *E. coli* [9]. CydABCD is probably nearly-essential in the conditions we used to generate the mutant library. For another 12 of the false positives, deletion strains have been previously reported to have a reduced growth yield (at most 2/3rds of a typical strain's yield) in liquid LB medium at 37˚C [1]. These genes are also likely to be nearly-essential under our growth conditions. If we consider these 15 cases as true positives then the FDR drops to 16%, but there may be other genes in our list that are nearly-essential for growth on LB plates at 37˚C. Additional discrepancies between the genes that lack insertions in our TnSeq data and the genes whose activity is required for growth could reflect the inhibition of Tn5 transposase activity in some genomic regions, such as by nucleoid proteins [10]. To estimate how many false positives we should expect in our list of likely-essential genes given that we mapped transposon insertions at 114,948 locations, we simulated data with no essential genes. Specifically, we randomized the location of each insertion event; to account for the varying coverage across the genome, we split the genome into 100 bins of about 46 kB each, and we kept each insertion within the same bin. We repeated this procedure 10 times and observed 20 essential genes on average in the absence of any biological information. (This may be a slight overestimate because our analysis of likely-essential genes also considers

whether the genes' mutants have significant abundance in the BarSeq start samples but we did not consider that for our shuffled analysis.) This analysis implies a false discovery rate of about 20/324 = 6%. So, we expect that the true rate of false positives in our list of *E. coli* proteins that are essential, or nearly so, for growth in rich media is somewhere between 6% and 16%.

In *S. oneidensis* MR-1, our list contained 397 protein-coding genes. We compared this to a list of likely-essential genes that we had previously generated using a different transposon and Sanger sequencing [3]. The previous analysis was conservative as genes that were not expected to be essential (based on orthology to *E. coli* or *Acinetobacter*, which are also γ-Proteobacteria) were required to be adjacent to another essential gene or to be conserved across most other *Shewanella* genomes. Of the 397 genes in our new list, 298 were previously classified as essential and 83 were classified as unknown, just 16 were previously identified as dispensable, and 2 of these are genes that are essential in *E. coli*. This implies a false discovery rate of around 16/(397-83) = 5%, with the caveat that insertions in a gene might be selected against even though the gene is not itself important (i.e., polar effects). We also examined the 15 genes that are putatively essential in *S. oneidensis* MR-1 but lacked clear orthologs in the closely related strain *S. sp.* ANA-3, as these might be more likely to be false positives. Of these 15, at least seven are plausibly essential, including three prophage repressors that are probably required to prevent prophage excision; a gene adjacent to one of these repressors; the RepA protein that is probably required to maintain the megaplasmid, a tRNA synthetase (SO3128.2) with a putative internal stop codon that nevertheless forms full-length protein [11], and ribosomal protein L25 (whose ortholog was missed because it seems to be annotated with the wrong start codon).

We then considered TIGR roles that are likely to be essential. The roles we chose were: DNA metabolism; transcription; protein synthesis; protein fate; energy metabolism; cell envelope; fatty acid and phospholipid metabolism; purines, pyrimidines, nucleosides, and nucleotides; amino acid biosynthesis; and biosynthesis of cofactors, prosthetic groups, and carriers. We included biosynthetic and energetic genes as likely-essential roles because many bacteria cannot take up as wide a range of nutrients as *E. coli* can or have more limited ways of creating energy. In *E. coli* and *S. oneidensis*, these categories account for 62-67% of putatively essential genes but just 12-18% of other genes. In the 25 bacteria, these categories accounted for 46-69% of putatively essential genes but just 6-17% of other genes. This confirms that in each bacterium, most of these genes are essential.

Finally, we asked whether the putatively essential genes had orthologs in other bacteria that were essential. Overall, 86% of the likely-essential genes were confirmed by conservation (they had an ortholog that was also essential). The organisms with the lowest proportions of conserved essentials were *S. elongatus* and *D. vulgaris* (60% and 68%, respectively). This might reflect their unique sources of energy (photosynthesis

and dissimilatory sulfate reduction, respectively) or their evolutionary distance from the other bacteria that we studied.

Supplementary Note 2. Rationales for annotating DNA repair proteins

This supplementary note describes our rationale for annotating genes with conserved and specific phenotypes on the DNA-damaging agent cisplatin, as shown in **Figure 3a**. As mentioned in the main text, of the 67 protein families that were specifically important for resisting cisplatin in more than one bacterium, 33 are known to be involved in DNA repair. Seven others are potentially involved in chromosome segregation or cell division and the cisplatin phenotype of these genes may be due to an indirect effect of DNA damage inhibiting cellular division [12].

Among the remaining 27 families, we considered 12 as putatively new DNA repair families because they contain DNA-related domains or because similar proteins are regulated by the DNA damage response regulator LexA in some bacteria [13-15]. 6 of these 12 protein families contain domains that suggest an involvement in DNA repair. However, among these is DUF3584, which is distantly related to the chromosome segregation protein Smc as well as to DNA repair proteins (RecF and RecN). We confirmed the cisplatin sensitivity of a DUF3584 mutant (SO4008 from *Shewanella oneidensis*; **Extended Data Fig. 4**), but this protein (and the two proteins with conserved proximity) are also required for motility, which requires separated cells. So, DUF3584 (and the two associated proteins) may be involved in DNA segregation instead of DNA repair. We predict that the other five families are involved in DNA repair because of their domain content or because of conserved proximity to DNA repair proteins (see **Supplementary Table 9** for a rationale for each family).

Among the four remaining families, two are uncharacterized (DUF1654 and PA0922-like) but because they are regulated by LexA, we predict that they are also involved in DNA repair. While regulation by LexA alone is not sufficient to confidently predict a role in DNA repair, we argue that LexA regulation together with a specific-important phenotype on cisplatin is very suggestive of a role in DNA repair. In support of this view, of the 40 *E. coli* genes regulated by LexA that are not involved in DNA repair [16], none have a specific-important phenotype on cisplatin. We also predict that the periplasmic nuclease EndA, which is present in *E. coli*, is involved in DNA repair. In *E. coli*, EndA is also important for resistance to ionizing radiation [17], which also damages DNA. We confirmed that a mutant of *E. coli endA* is sensitive to cisplatin (**Extended Data Fig. 4**). Furthermore, through genetic complementation assays, we showed that the catalytic residue of EndA's nuclease domain is important for cisplatin resistance (**Extended Data Fig. 5**). Finally, in Pseudomonads, *endA* is regulated by LexA [14]. Nevertheless, EndA's precise role remains unclear. The remaining potential DNA repair family is YhgF, which contains two putative DNA-binding domains (that are not necessarily linked to DNA repair). Although YhgF is required for ionizing radiation resistance in *E. coli* [17], a mutant

in this gene has increased stop codon read-through and has genetic interactions with translation-related proteins [18], so YhgF may not be involved in DNA repair.

The remaining 15 families of proteins that have a conserved specific-important phenotype in cisplatin stress are probably not involved in DNA repair (**Supplementary Table 9**). Three are involved in tRNA or rRNA metabolism, which could be related to RNA damage by cisplatin [19]. Overall, our cisplatin experiments provide an overview of the proteins involved in DNA repair across diverse bacteria, including 33 protein families with a known role, 8 protein families that we predict are involved in DNA repair, and 4 ambiguous cases.

Supplementary Note 3. Improving annotations of ABC transporter proteins

We systematically examined 101 ABC transporter proteins with strong and specific-important phenotypes on different carbon and nitrogen sources. 20 of the 101 proteins only have vague annotations, and for these we made novel substrate predictions based on the specific phenotype data. For example, Dshi_0548 and Dshi_0549 from *Dinoroseobacter shibae* are annotated with no substrate specificity yet are important for utilizing xylitol. For another 31 proteins with moderately-specific substrate annotations, such as "amino acids", we predicted a specific substrate within that group of compounds. For example, Ac3H11_2942 and Ac3H11_2943 from *Acidovorax* sp. 3H11 are annotated as transporting "various polyols", whereas our data shows they are important for utilizing the polyol D-sorbitol but not the polyol D-mannitol. Another 11 proteins had incorrect annotations: for example, in three Pseudomonads, the D-glucosamine transporter is annotated as an amino acid transporter (i.e., AO353_21715:AO353_21720), but D-glucosamine is not an amino acid. In 13 cases, the specific phenotypes indicate that the protein transports a substrate that was not included in the annotation along with a substrate that was expected. 24 proteins had correct annotations that were confirmed by our data; these included 22 cases in which the specific phenotype(s) were expected given the annotation, and 2 cases in which the gene has the expected mutant phenotype(s) but the association to a specific condition was misleading. Overall, we improved the annotations for 75 of the 101 examined proteins.

Supplementary Note 4. Novelties in carbon catabolic pathways

Most of our re-annotations of catabolic enzymes are for proteins whose homologs are known to act on different substrates (see **Supplementary Table 12**), and the genetic information helped us identify a substrate that was physiologically relevant and was not consistent with the original annotation. Here, we describe some more challenging cases where we identified novel enzymatic reactions or linked poorly-characterized enzymes to genes. We also describe an example of an enzyme that was known, biochemically, to act on multiple substrates, and our genetic data found that both activities are relevant *in vivo*.

The gene for glucosaminate ammonia-lyase in Pseudomonads. Some bacteria catabolize D-glucosamine by oxidation to D-glucosaminate followed by the action an enzyme known as D-glucosaminate ammonia-lyase or D-glucosaminate dehydratase, which forms 2-keto-3-deoxy-D-gluconate and ammonia (**Supplementary Fig. 1a**) [20]. The 2-keto-3-deoxy-D-gluconate is then phosphorylated and enters the Entner-Doudoroff pathway [20]. Alternatively, there is a report that the enzyme also has an aldolase activity to form glyceraldehyde, pyruvate, and ammonia [21].

Although purified D-glucosaminate ammonia-lyase from *Pseudomonas* has been studied biochemically, it does not appear in any of the sequence databases. As far as we can tell, this enzyme had never been linked to a gene. Iwamoto et al. did identify the sequence of a putative α_2 enzyme but this turned out to be thioredoxin reductase [22], which is essential for viability based on our transposon data, and the reported K_m of 12 mM seems too high for the physiological enzyme. They also mention purifying a 90 kD β_2 enzyme [22].

We identified putative D-amino acid deaminases that were specifically important for utilizing D-glucosamine in several Pseudomonas (AO356_00450, Pf6N2E2_2047, AO353_21740, PfGW456L13_2867, and PS417_12125; **Supplementary Fig. 1b**). Also, D-glucosaminate ammonia-lyase enzyme is reported to require a 5'-pyridoxal-phosphate cofactor, as expected for this gene family. Finally, the expected molecular weight of this protein is around 43 kD, which is consistent with a 90 kD β_2 dimer. Together, this strongly suggests that these genes encode D-glucosaminate ammonia-lyase.

Catabolism of D-arabinose via 2,4-dioxopentanoate (2 novel reactions). In *Sinorhizobium meliloti* 1021, D-arabinose is catabolized by an oxidative pathway that has some overlap with L-fucose oxidation. This proposed pathway has some similarity to the oxidative pathway for D-arabinose consumption in *Sulfolobus solfataricus*, which also proceeds via D-arabinoate and 2-dehydro-3-deoxy-D-arabinoate [23]. But in *S. solfataricus*, a different dehydratase enzyme forms 2,5-dioxopentanoate, which is oxidized to α-ketoglutarate (an intermediate in the citric acid cycle). Our proposal includes two putative novel reactions in *S. meliloti*: the dehydration of 2-dehydro-3-deoxy-D-arabinoate to 2,4-dioxopentanoate (instead of 2,5-dioxopentanoate) and the hydrolysis of 2,4-dioxopentanoate (**Supplementary Fig. 2**).

All of the genes in this pathway have specific phenotypes, either for D-arabinose utilization alone or for L-fucose utilization as well. The catabolism of L-fucose by *S. meliloti* also has some novel aspects and is discussed later.

SMc02775: D-arabinose → D-arabinolactone + 2H. This protein is annotated as L-fucose dehydrogenase, but it has little phenotype on L-fucose. The product is probably the 1,5 lactone rather than the 1,4 lactone. A complication is the 1,5-lactone might spontaneously rearrange to the 1,4-lactone at significant rates, as reported for L-fucono-

1,5-lactone [24]. Fungal enzymes are reported to produce the 1,4-lactone but are distantly related.

SM_b21101: D-arabinolactone + H_2O → D-arabinoate. A related protein (29% identical) was shown by Hobbs et al. [24] to be a L-fucono-1,5-lactonase. SM_b21101 is mildly important for fitness on L-fucose as well, so it apparently acts on both D-arabinolactone and L-fuconolactone.

SMc02776: D-arabinoate → 2-dehydro-3-deoxy-D-arabinoate + H_2O. It is not certain which dehydratase performs which step, but SMc02776 is annotated as altronate dehydratase, which is a similar substrate. SMc02776 is also important for fitness on L-fucose. And it is predicted to be in an operon with the D-arabinose dehydrogenase SMc02775. SMc02776 is 36% identical to D-galactarate dehydratase (GarD) from *E. coli*.

SMa0247: 2-dehydro-3-deoxy-D-arabinoate → 2,4-dioxo-pentanoate + H_2O. SMa0247 is related to 2-keto-3-deoxyxylonate dehydratase (XylX), which acts on the same substrate. Also, a close homolog (HSERO_RS19360, 59% identical) is important for D-xylose utilization. However, the product of XylX would be 2,5-dioxo-pentanoate, which does not make sense for the next step.

SM_b21112: 2,4-oxo-pentanoate + H_2O → pyruvate + acetate. A similar protein (58%) hydrolyses L-2,4-diketo-3-deoxyrhamnonate to pyruvate and L-lactate (LRA6) [25]. This substrate is similar to 2,4-dioxo-pentanoate but with an additional -CHOH- group (hence lactate instead of acetate as the product). This 6-carbon hydrolase reaction is also proposed to be involved in L-fuconate degradation [26]. Yew and colleagues refer to it as 2,4-diketo-3-deoxy-L-fuconate hydrolase, but it is the same reaction as in rhamnose catabolism. Indeed, SM_b21112 is also important for L-fucose utilization. This may be the first direct evidence for the involvement of this hydrolase in L-fucose catabolism. Note that SM_b21112 belongs to the fumarylacetoacetate hydrolase family, which hydrolyses substrates of the form R-CO-CH2-CO-R' via an alkoxy intermediate (which involves one CO group) and a carbanion leaving group in the enolate form (which involves the other CO group [27]). This is consistent with 2,4-dioxo-pentanoate as the substrate, but not 2,5-dioxo-pentanoate.

Additional dehydratase steps in the oxidation of L-fucose. *Sinorhizobium meliloti* 1021 appears to use an oxidative pathway for L-fucose catabolism that was identified in *Xanthomonas campestris* [26]. This pathway should require one dehydratase, to convert L-fuconate to 2-keto-3-deoxy-L-fuconate. But we identified three putative dehydratase genes as specifically important for fitness during growth on L-fucose in *S. meliloti.*

In the characterized version of this pathway [26], L-fucose is converted to the α-pyranose form by a mutarotase, oxidized to L-fucono-lactone, hydrolyzed to L-fuconate, dehydrated to 2-keto-3-deoxy-fuconate, oxidized to 2,4-diketo-3-deoxy-fuconate, and

hydrolyzed to pyruvate and L-lactate. Fitness data identified most of the genes from this pathway as important for utilizing L-fucose in *S. meliloti*. SM_b21108 is the mutarotase; SM_b21109 is the fucose dehydrogenase; SM_b21101 is the lactonase; SM_b21111 is the ketodeoxyfuconate dehydrogenase (59% identical to XCC4067); and SM_b21112 is the hydrolase. Finally, a L-lactate dehydrogenase (SM_b20850) is important for growth on L-fucose. SM_b20850 is also specifically important for growth on L-lactate, which confirms that it is annotated correctly and that L-lactate is an intermediate, as in *X. campestris*.

The fitness data identified three dehydratases as being important for L-fucose catabolism in *S. meliloti* (SMc02776, SM_b21107, and SM_b21110). Of these, SM_b21107 was studied biochemically and was reported to have some L-fuconate dehydratase activity, but "not at rates that would be physiologically relevant" (http://hdl.handle.net/2142/14751). This suggests that SM_b21107, which was expected to be the L-fuconate dehydratase, plays a slightly different role. SMc02776 is related to D-galactarate dehydratase and is also important for D-arabinose utilization as discussed above. SM_b21110 is similar to the (R)-enoyl-CoA hydratase portion of *E. coli* MaoC (36% identical). We propose that two of the dehydratases work together to switch the stereochemistry of L-fuconate (i.e., dehydration followed by rehydration to a stereoisomer of L-fuconate), and the third dehydratase forms 2-keto-3-deoxy-L-fuconate (or an isomer thereof). Alternatively, the additional dehydratases could be involved in the formation of another molecule that induces expression of the pathway, instead of being part of the main branch of catabolism. Also the phenotype of SM_b21110 could possibly be due to a polar effect (but then the requirement for a second dehydratase would be difficult to explain).

Divergent Arginine Deiminases. We identified several proteins from the amidinotransferase family (PF02274) as specifically important for growth on L-citrulline (PGA1_c16380 from *Phaeobacter inhibens*, AO353_25635 from *Pseudomonas fluorescens* FW300-N2E3, and PS417_17580 from *Pseudomonas simiae* WCS417). These proteins are similar to each other (over 40% identity) and are distantly related to characterized arginine deiminases, which are also classified in PF02274 but are not found as homologs by BLAST. We propose that that these three proteins allow growth on L-citrulline by catalyzing the L-arginine deiminase reaction in reverse (L-citrulline + NH_3 → L-arginine + H_2O). For example, in *P. inhibens*, the resulting arginine appears to be cleaved by an arginase (PGA1_c16370) to ornithine, ornithine cyclodeaminase (PGA1_c16390) converts it to L-proline, and the L-proline is catabolized via PutA (PGA1_c11750). All of these genes are required for L-citrulline utilization in *P. inhibens.*

Role of GguC-like proteins in L-arabinose catabolism. Some bacteria and archaea catabolize L-arabinose by an oxidative pathway (L-arabinose degradation III in MetaCyc) that includes oxidation to a lactone by a dehydrogenase, conversion to L-arabinonate by a lactonase, and two dehydratase reactions to 2,5-dioxo-pentanoate (also known as α-ketoglutarate semialdehyde), which can be oxidized to α-ketoglutarate and enter the tricarboxylic acid cycle. The second of these dehydratase reactions is the conversion of L-2-keto-3-deoxyarabonate to 2,5-dioxo-pentanoate (EC 4.2.1.43). As far as we know, the only protein that is biochemically confirmed to have this activity is AraD (Q1JUQ0) from *Azospirillum brasilense*, which belongs to the dihydrodipicolinate synthase (DapA) family [25]. However, GguC (Atu2345) from *Agrobacterium tumefaciens* is also proposed to have this activity [28]. GguC belongs to the fumarylacetoacetate hydrolase superfamily, which includes dehydratases that act on a stereoisomer of L-2-keto-3-deoxyarabonate (D-2-keto-3-deoxyarabonate), and a *gguC* mutant has reduced growth on L-arabinose [28]. GguC might not be entirely required for the utilization of L-arabinose because *A. tumefaciens* also contains a homolog of AraD [28].

We found that GguC-like proteins are required for L-arabinose utilization in *Sinorhizobium meliloti* and in five strains of *Pseudomonas.* The genes are SM_b20892, AO356_24595, Pf6N2E2_611, PfGW456L13_3317, Pf1N1B4_4623, and PS417_11010. These proteins are 53-76% identical to GguC. These bacteria do not contain AraD-like proteins, which may explain why, unlike in *A. tumefaciens*, mutants in these genes have strong phenotypes. Furthermore, genes for all of the other biochemical steps in the L-arabinose oxidation pathway were identified as important for its utilization (**Supplementary Fig. 3)**. Thus, our data confirms that GguC-like proteins from diverse bacteria are L-2-keto-deoxyarabonate dehydratases.

Bifunctional L-arabinose/D-galactose 1-dehydrogenases. Biochemical studies have reported bifunctional L-arabinose/D-galactose 1-dehydrogenases in *Azospirillum brasilense* [25] or *Rhizobium leguminosarum* [29]. We found that a related protein, SMc00588 (80% identical to the enzyme from *R. leguminosarum*, which is UniProt B5ZWY9) was important for the utilization of both L-arabinose and D-galactose (**Supplementary Fig. 3**). This confirms that the bifunctionality is relevant *in vivo*. These enzymes belong to the MviM family (COG0673; PF01408).

We also identified putative bifunctional L-arabinose/D-galactose 1-dehydrogenases in the short chain dehydrogenase family (PF00106). These include the similar proteins Ac3H11_614 from *Acidovorax* sp. GW101-3H11 and BPHYT_RS16920 from *Burkholderia phytofirmans* PsJN and a distantly-related group of proteins in *Pseudomonas* (PfGW456L13_2119, Pf1N1B4_412, Pf6N2E2_5967). All of these genes are important for utilization of L-arabinose and D-galactose and not in most other conditions. Furthermore, there is no other dehydrogenase step that would explain the requirement for these enzymes, except for α-ketoglutarate semialdehyde dehydrogenase, which was identified as another protein. Also, Ac3H11_614 and BPHYT_RS16920 are similar to the xylose 1-dehydrogenase (XylB or CC0821) from *C.*

crescentus [30] (BPHYT_RS16920 is 42% identical). One unexplained aspect of the pathway in *B. phytofirmans* is that an adjacent short-chain dehydrogenase, BPHYT_RS16940, is also important for utilizing both substrates.

Supplementary Note 5. Rationales for protein function predictions for domains of unknown function

Specific Predictions:

DUF485 (PF04341): component of actP-like carboxylate transporters. Several representatives of DUF485 (which is known as *yjcH* in *E. coli*) are cofit with an adjacent ActP-like permease. Examination of the per-strain data did not show any evidence of polar effects (when a transposon insertion in an upstream gene in an operon blocks transcription of the downstream genes), so we suggest that DUF485 is required for the activity of the permease. The paper that characterized ActP in *E. coli* did not rule out the requirement of another gene [31]. DUF485 seems to be a membrane protein (i.e., AZOBR_RS02935 has two transmembrane helices and a possible signal sequence). Together, this suggests that DUF485 is a component of the transporter. The clearest phenotypes are for pyruvate utilization (i.e. AZOBR_RS02935).

DUF1513 (PF07433): outer membrane component of ferrous iron uptake. DUF1513 is in a conserved operon and is cofit with the other genes in that operon in multiple bacteria. The operon contains EfeO-like, DUF1111, a second EfeO-like, and DUF1513. EfeO has an unknown role in iron uptake by *efeUOB* and is a putative periplasmic lipoprotein. DUF1111 is proposed to be a homolog of EfeB, a di-heme peroxidase involved in ferrous iron uptake [32]. Thus, the operon seems to be involved in ferrous iron uptake. Related operons in α-Proteobacteria often contain bacterioferritin, which is consistent with that role.

DUF1513 has a putative signal peptide (PSORTb) and has similarity to beta propellers with 6-8 repeats (Pfam clans). This suggests that it is the outer membrane component of this system.

Although the fitness data shows that DUF1513 is involved in this process, it is not certain that ferrous iron is the substrate. Several representatives of DUF1513 are pleiotropic, but AO356_18450 is specifically important for chlorite resistance and Psest_1156 is sensitive to various metals. Inhibiting iron uptake would plausibly create these phenotypes.

DUF1656 (PF07869): component of efflux pump with MFP and FUSC. Several representatives of DUF1656 are in an operon with and cofit with a RND efflux pump and a fusaric acid resistance-like protein. DUF1656 is related to *E. coli* YdhI and AaeX/YhcR. Homologs of these proteins are sometimes annotated as inner membrane efflux pump

components or as Na+-dependent SNF-like transporters but we could not find the rationale for these annotations. (As of March 2016, EcoCyc reports that the functions of *aaeX* and *ydhI* are not known.)

This family has a variety of stress sensitivity phenotypes. In *Cupriavidus basilensis*, this efflux pump is specifically important for the utilization of 4-hydroxybenzoate. It could be involved in uptake, or 20 mM of 4-hydroxybenzoate may have been inhibitory and it is involved in efflux. Similarly, a number of representatives are important for the utilization of octanoate and it is not clear if this reflects uptake or efflux. In several strains of *Pseudomonas fluorescens*, the efflux pump is either important for or detrimental during acetate utilization. Finally, in *Zymomonas mobilis*, we found that a similar system (ZMO1432-ZMO1431-ZMO1430) is involved in resisting hydrolysate [33].

DUF1854 (PF08909): subunit of transporter for efflux of an amino acid polymer. In two β-Proteobacteria, *Acidovorax* sp. 3H11 and *Cupriavidus basilensis* 4G11, representatives of DUF1854 (i.e., RR42_RS04420) are cofit with two nearby genes that are annotated as cyanophycin synthetases as well as an ABC-like transporter. Cyanophycin is a copolymer of aspartate and arginine that many cyanobacteria use to store nitrogen; it is also formed by some heterotrophic bacteria [34]. The putative cyanophycin synthetases (ChpA and ChpA') have been studied in another strain of *Cupriavidus necator* and do not form cyanophycin but may produce a different light-scattering polymer [35]. Also, the genomes of many β-Proteobacteria contain *chpA* and *chpA'* but do not appear to encode the cyanophycinase (*chpB*) to break down the polymer, which hints at a different role for these genes [34]. As *chpA* and *chpA'* were cofit with an ABC transporter in both organisms, we propose that the polymer is being exported to form part of the cell wall rather than serving as a storage compound. This can also explain why the genes are important for motility (in *Acidovorax*) or have pleotropic stress phenotypes (in both organisms). (We did not succeed in measuring fitness during motility in *C. basilensis*.) The ABC transporter contains an ABC transporter transmembrane domain and an ATPase domain. DUF1854 is usually found in an apparent operon with the ABC transporter, but in a few genomes the two proteins are fused together, as pointed out by the Pfam curators. Together with the cofitness, this suggests that DUF1854 is involved in the transport and forms a complex with the ABC transporter.

DUF2849 (PF11011): electron source for sulfite reductase. This family is usually upstream of *cysI*, the beta subunit of sulfite reductase. These sulfite reductases are important for fitness in our defined media, which confirms that they are indeed sulfite reductase and not nitrite reductase, as sulfate is the sulfur source in these media and sulfate must be reduced to sulfite and then to sulfide before it is assimilated. (In *S. meliloti*, *cysI* or SMc02124 was misannotated as "nitrite reductase.") However, these genomes do not seem to contain the alpha flavoprotein subunit (*cysJ*). Instead, DUF2849 is found upstream. Several representatives of DUF2849 have similar fitness patterns as the downstream *cysI* (SMc01054, AZOBR_RS10130, Ga0059261_1497).

The phenotypes of DUF2849 do not seem to be due to polar effects, as strains with insertions in either orientation have low fitness in defined media. Also, DUF2849 is found fused to *cysI* in *Pseudomonas putida* GB-1. Altogether this suggests that DUF2849 is required for the activity of sulfite reductase. Usually *cysJ* is the electron source for *cysI* so we propose that in its absence, DUF2849 fulfills this role. (The relationship between the representatives of DUF2849 with similar cofitness was missed by the automated analysis of orthologs, as these alignments just missed the cutoffs of coverage > 80% or E < 10^{-5}. So this case is absent from **Supplementary Table 13**.)

DUF4212 (PF13937, TIGR03647): small subunit of transporter for D-alanine, lactate. DUF4212 was predicted by the TIGRFAM curators to be the small subunit of a solute:sodium symporter because of its hydrophobicity and conserved gene context. Multiple members of this family (e.g., Sama_1522 or Psest_0346) are cofit with a putative large subunit of a symporter that is downstream (e.g., Sama_1523 or Psest_0347). Sama_1522 from *Shewanella amazonensis* SB2B is important for fitness when D,L-lactate is the carbon source; similarly, a homolog in *S. oneidensis* MR-1 (SO2858) is more mildly important in some D,L-lactate conditions. Psest_0346 in *Pseudomonas stutzeri* RCH2 (51% identical to Sama_1523) is very important for D-alanine utilization and strongly detrimental during L-alanine utilization. Note that alanine and lactate are chemically analogous three-carbon organic acids, with alanine having an amino group where lactate has a hydroxyl group. Overall, our data confirms that that DUF4212 is required for the transporter's activity, so it is probably an additional subunit. Some members of DUF4212 are cofit with a nearby member of COG2905 (RNAse T domain protein) (e.g., Sama_1525 or Psest_0349); we speculate that COG2905 is required for expression of the symporter.

UPF0060 (PF02694, COG1742): efflux pump for thallium (I) ions. UPF0060 is specifically important for resisting thallium (I) stress in *Cupriavidus basilensis* (RR42_RS34240), in *Sphingomonas koreensis* (Ga0059261_1942), and in two strains of *Pseudomonas fluorescens*. This family includes E. coli YnfA, which is an integral membrane protein. Nir Hus (PhD dissertation, 2005) proposed that YnfA belongs to the SMR (small multi-drug-resistance) family and SCOOP [36] (a family-family relationship finder) shows similarity to transporters and efflux pumps. UPF0060 is sometimes adjacent to cation efflux genes related to *czcD* or *zntA*. So, we propose that these members of UPF0060 are efflux pumps for thallium (I).

UPF0126 (PF03458, COG2860): glycine transporter. A number of representatives of UPF0126 are specifically important for utilizing glycine (PGA1_c00920, SO1319, Sama_2463, Psest_1636, AO353_13110, RR42_RS33360, CA265_RS15435, Pf6N2E2_3621, PfGW456L13_161, PS417_02455, Shewana3_1204). SO1319 was not identified as an ortholog of Sama_2463 or the genes from Pseudomonas by our automated approach, but SO1319 is homologous and has a specific phenotype in glycine-containing media. These genes contain a pair of UPF0126 domains and are predicted to be membrane proteins. SCOOP identified similarity between UPF0126 and

TRIC channels, so we propose that this is a family of glycine transporters. This family includes *yicG* from *Escherichia coli*, which is not characterized (and which we did not identify any significant phenotypes for).

Pathway-level Predictions:

DUF444 (yeaH, PF04285) and SpoVR (ycgB, PF04293): signaling with serine kinase yeaG. As discussed in the main text, we identified strong and conserved cofitness between *yeaG*, *yeaH*, and *ycgB* in multiple bacteria, including *Escherichia coli*. These proteins are predicted to be functionally related because of conserved proximity, co-occurrence, and coexpression [37]. These genes form a single operon in most bacteria, but in *E. coli* and *Klebsiella michiganensis* M5a1 they are broken up into two operons (*yeaGH* and *ycgB*). While these three genes are more cofit with each other than with any other gene in each of the seven bacteria, the phenotypes are not conserved across species (**Supplementary Fig. 4**). We validated some of these phenotypes in growth assays with individual mutants. These experiments confirmed that in *E. coli*, mutants in all three genes have a growth advantage when L-arginine is the sole nitrogen source; in *S. oneidensis*, mutants in all three genes grow slowly when N-acetylglucosamine is the carbon source; and in *P. fluorescens* FW300-N2E2, mutants in *yeaG* and *yeaH* grow slowly on propionate as a carbon source (**Supplementary Fig. 5**). The physiological target of YeaG is not known but it is a PrkA-like protein kinase and can phosphorylate itself or casein *in vitro* [38]. In *E. coli*, YeaG is reported to be important for survival after nitrogen starvation via its effect on the expression of the toxins YafQ and MqsA [39]. Also, YeaH appears to be involved in this process, as a *yeaH* deletion strain had a similar phenotype as a *yeaG* deletion strain or a *yeaGH* deletion strain [39]. However, little is known about the molecular mechanisms of these proteins, except that the catalytic activities of both the kinase domain of YeaG and the AAA+ domain of YeaG are required [39]. SpoVR is so named because a member of this family is involved in sporulation in *Bacillus subtilis,* but nothing is known about its biochemical function and the organisms that we studied do not sporulate. Since YeaG appears to be a signaling protein, we infer that YeaH and YcgB are involved in this signaling pathway as well. Most of the bacteria in which we identified cofitness of *yeaG*/*yeaH*/*ycgB* do not contain orthologs of *yafQ* or *mqsA*, so the targets in other bacteria are probably different. The mutant phenotypes of these genes are not conserved in different bacteria, but there is a potential commonality relating to the utilization of amino acids: these genes are detrimental for the utilization of L-arginine as the nitrogen source by *E. coli*, detrimental for the utilization of L-tryptophan or L-lysine as the nitrogen source in *K. michiganensis* M5a1, detrimental for the utilization of L-methionine or L-phenylalanine as the nitrogen source by *S. oneidensis* MR-1, and detrimental for the utilization of L-serine as the carbon source by *P. fluorescens* FW300-N2E3. This signaling pathway seems to be related to nitrogen metabolism in all of these bacteria, as observed in *E. coli* by Figueira and colleagues [39]. On the other hand, these genes are important for propionate utilization in *P. fluorescens* FW300-N2E2, which has no obvious relationship to amino acid metabolism or nitrogen starvation.

DUF466 (PF04328, COG2879): accessory to pyruvate transport by CstA-like. Three representatives of DUF466 were cofit with a nearby CstA-like protein. In *E. coli*, CstA is reported to be a peptide transporter. In *Desulfovibrio alaskensis* G20, a CstA-like protein (Dde_2007) is specifically important for fitness when pyruvate is the carbon source (data from [40,41]). The only strong phenotype for the DUF466 and *cstA* in our data were in *C. basilensis*, where they were specifically important for pyruvate utilization. So, we propose that DUF466 is required for pyruvate transport by a CstA-like protein.

DUF692 (PF05114), DUF2063 (PF09836): chlorite stress signaling proteins. DUF692 and DUF2063 form a conserved operon that is important for chlorite resistance in *Pseudomonas stutzeri* RCH2 (Psest_0116:Psest_0117), *Kangiella aquimarina* (B158DRAFT_1333:B158DRAFT_1334), and *Shewanella amazonensis* SB2B (Sama_1305 = DUF692; DUF2063 seems to have been replaced by a hypothetical protein, Sama_1304, which also has this phenotype). In *Shewanella amazonensis*, DUF692 is cofit with the hypochlorite scavenging system *yedYZ* (Sama_1893:Sama_1892, see [42,43]), so DUF692 probably has another role rather than detoxifying hypochlorite directly. DUF692 is a putative xylose isomerase or epimerase-like and DUF2063 is a putative DNA binding protein. We propose that these genes are involved in sensing an aspect of chlorite stress or metal stress. Some homologs are in an operon with an extracellular type sigma factor that responds to heavy metal stress (e.g. NGO1944 from *Neisseria gonorrhoeae*, which regulates methionine sulfoxide reductase). This suggests that DUF2063 might be an anti-anti-sigma factor rather than a DNA-binding protein. Consistent with this, in *Neisseria*, there is a putative anti-sigma factor downstream of the sigma factor (e.g., NMB2145), which is not similar to DUF2692 or DUF2063.

DUF934 (PF06073): accessory protein for sulfite reduction. This uncharacterized protein family is conserved downstream of *cysI*, the β subunit of sulfite reductase. In SEED, members of this family are annotated as CysX or "Oxidoreductase probably involved in sulfite reduction," but we were not able to find any published experimental evidence. They were probably annotated based on conserved gene proximity. (DUF934 does not seem to be homologous to the small ferredoxin-like CysX of *Corynebacterium glutamicum* described by Ruckert et al. [44]; furthermore, it lacks the CxxC motifs that are conserved in CysX.) We found that DUF934 was strongly cofit with various genes in the sulfate assimilation pathway in various *Pseudomonas* (i.e., Psest_2088) as well as in *Marinobacter adhaerens* HP15 and *C. basilensis* 4G11. In *M. adhaerens* and in *C. basilensis*, there are other sulfate assimilation genes downstream, but not in the *Pseudomonas* species, so this is not a polar effect.

DUF971 (PF06155): FeS cluster maintenance protein. In several *Shewanella* or *Pseudomonas* species, a representative of DUF971 is cofit with Mrp, BolA, and/or YggX. All of these proteins are related to FeS cluster maintenance: Mrp is an FeS loading protein; BolA is related to the Fra2 protein of *Saccharomyces cerevisiae* which is an FeS cluster protein and regulates iron levels; and YggX plays a

role in oxidation resistance of FeS clusters. DUF971 is pleiotropic but a number of representatives are important for resisting cobalt (II) or paraquat; either of these might disrupt FeS clusters. Nothing is known about the biochemical function of DUF971, but in eukaryotes, DUF971 is found as a domain within gamma-butyrobetaine hydroxylase and trimethyllysine dioxygenase proteins, and it is also found within the chloroplast 4Fe4S cluster scaffold protein HCF101 (fused with an Mrp-like domain). These occurrences are consistent with DUF971 having a role in FeS cluster maintenance.

DUF1302 (PF06980), DUF1329 (PF07044): export of cell wall component. We identified cofitness for a gene cluster comprising DUF1302 (e.g., Sama_1588 in *Shewanella amazonensis* SB2B), DUF1329 (Sama_1589), a BNR repeat protein (COG4447; Sama_1590), and a putative RND export protein (COG1033; Sama_1591). The BNR repeat protein is related to a photosystem II stability/assembly factor (ycf48 or hcf136) and has a beta-propeller fold (PDB 2xbg). These genes were conserved near each other and are cofit in *Kangiella aquimarina* DSM 16071 as well. In *Pseudomonas stutzeri* RCH2, the cluster is broken up into two pieces, Psest_1122:_1123 (DUF1302 and DUF1329) and Psest_1923:_1924 (BNR repeat protein and RND export protein). Again these proteins are cofit, although in one condition (tyrosine as a carbon source), the two groups have strong and opposing phenotypes; this could imply some separation of function, or it could relate to the existence of paralogs for DUF1329 in this organism. These genes were cofit in some strains of *Pseudomonas fluorescens* as well. In both *S. amazonensis* and *P. stutzeri*, their phenotypes are pleotropic.

Several lines of evidence suggest that these proteins are in the outer membrane and affect the cell wall. DUF1329 has some similarity to the outer membrane protein LolB (e.g., the C-terminal part of Sama_1589 is similar to CATH superfamily 2.50.20.10). In *Delftia sp.* Cs1-4, both DUF1302 and DUF1329 are reported to be components of extended outer membrane vesicles or nanopods [45]. BNR repeat proteins are often found in the outer membrane. And a mutant of a member of DUF1329 in *P. fluorescens* F113 is reported to have increased swimming motility [46]. We propose that these four proteins work together to export a component of the cell wall.

Incidentally, a number of other studies have discussed the DUF1329 family but without identifying a biochemical function or a mutant phenotype. In *Delftia sp.* Cs1-4, a DUF1329 family member (*phnK*) is in a genomic island for phenanthrene catabolism, where it is also near a RND efflux protein [47]. In *Thauera aromatica* T1, expression of a member of this family (*pipB*) is induced by p-cresol (M. Chatterjee, PhD Thesis 2012). (A DUF1302 member, *pipA*, was also induced.) In *Comamonas testosteroni*, ORF61 (DUF1329) is in a gene cluster for steroid degradation but is not required for it [48]. The presence of DUF1329 in these gene clusters seems to suggest a specific role in the maturation of an outer membrane transporter, or directly in the transport of aromatic compounds, but this is not what we observed in *S. amazonensis* or *P. stutzeri*.

DUF1654 (PF07867): DNA repair protein. In several strains of *Pseudomonas fluorescens*, members of DUF1654 (PF07867) are specifically important for resisting the DNA-damaging agent cisplatin (e.g., AO356_00980). Furthermore, these genes are upstream of an endonuclease precursor, and close homologs of these proteins are predicted to be regulated by LexA, which controls the DNA damage response (e.g., PFL_2098 from *P. fluorescens* Pf-5). These imply a role for DUF1654 in DNA repair, but we have little idea of its molecular function. Polar effects seem unlikely because DUF1654 has a strong phenotype relative to the downstream endonuclease, but we cannot rule them out entirely. DUF1654 has some similarity (SCOOP) to an anti-parallel beta barrel family (calycin_like, PF13944).

DUF2946 (PF11162): copper homeostasis protein. In two strains of *Pseudomonas fluorescens*, DUF2946 is in an operon with and cofit with a TonB-dependent copper receptor (TIGR01778) and a Cox17-like copper chaperone (i.e., AO356_08325 with _08320 and _08330; the copper chaperone is also referred to as CopZ). Some members of DUF2946 are labeled as being homologous to COG2132, which includes *E. coli* CueO and SufI, but those are much longer proteins (around 500 amino acids, while DUF2946 is around 100 amino acids). We propose that DUF2946 is also involved in copper uptake or homeostasis.

DUF3584 (PF12128): Smc-like chromosome partitioning protein. In three species of *Shewanella*, DUF3584 was specifically important for resisting cisplatin stress (i.e., Sama_0592) and also for motility. According to HHSearch, DUF3584 is related to the N-terminal domain of Smc, or RecF or RecN, which are all involved in DNA repair or recombination. Together, this strongly suggests that DUF3584 is involved in chromosome segregation or DNA repair, but we do not have a specific proposal for its biochemical role. DUF3584 is conserved cofit with two hypothetical proteins (e.g., Sama_0594 and Sama_0593) that do not belong to any annotated families, and the three genes probably act together.

UPF0227: hydrolase affecting the cell envelope. UPF0227 includes the *E. coli* protein YqiA. *In vitro*, YqiA has esterase activity on palmitoyl-CoA and p-nitrophenylbutyrate [49], but little is known about its biological role. Mutants in this family have diverse phenotypes, including sensitivity to bacitracin, which inhibits synthesis of a peptidoglycan precursor (*E. coli* yqiA, Sama_3044, and Psest_0482), and sensitivity to carbenicillin, which a β-lactam antibiotic and inhibits peptidoglycan synthesis (SO3900, Psest_0482). In some *Shewanella* species (but not in *S. oneidensis* MR-1), this family is important for fitness in many defined media conditions. These phenotypes are consistent with an effect on cell wall or membrane composition. Given its esterase activity on palmitoyl-CoA, an effect on lipid composition seems most likely. A caveat is that these genes are upstream of the essential gene *parE*, and it difficult to rule out a phenotype from reducing *parE* expression (a polar effect). Although we cannot rule out this explanation, we did find that insertions within *E. coli*'s *yqiA* or within Sama_3044 have strong phenotypes in either orientation, which makes it less likely.

Supplementary Note 6. Relevance to all bacteria

Beyond the 32 bacteria experimentally studied in the present study, combining fitness data with comparative genomics can provide phenotype-derived insights into protein function across all sequenced bacterial genomes. Specifically, we asked if our dataset could help illuminate the potential functions of hypothetical or vaguely-annotated proteins from phylogenetically diverse bacterial genomes. To measure the relevance of our fitness data to diverse bacteria, we started with 1,752 bacterial genomes from MicrobesOnline [50]. We grouped together closely related genomes if they were separated from their common ancestor by less than 0.01 substitutions per site in highly conserved proteins (using the MicrobesOnline species tree). These groups correspond roughly to species. We selected one representative of each group at random, which gave us 1,236 divergent bacterial genomes. We selected 5 proteins at random, regardless of their annotation, from each of these genomes to form our sample of bacterial protein-coding genes. 2,593 of these proteins had hypothetical or vague annotations and were included in the analysis of **Extended Data Figure 9**. With such a large sample, the standard error in a proportion of 10% would be just 0.6%. The best-scoring BLAST hit to one of the 32 bacteria that was studied was considered as a potential ortholog if the alignment coverage was 75% or more; we used a range of threshold for similarity but our recommended cutoff is that the ratio of the BLAST alignment score to the self-score be above 0.3 (as in [51]).

For poorly-annotated proteins from diverse bacteria, 16% have a potential ortholog (with at least 75% coverage and 30% sequence similarity) with a statistically significant phenotype in our dataset (**Extended Data Fig. 9**). Furthermore, for 12% of the poorly-annotated proteins, we can associate a potential ortholog to a specific condition or to another protein with cofitness (**Extended Data Fig. 9**). Given that function is often conserved at 30% similarity [52], our data should provide functional insight for many poorly characterized bacterial proteins. Not surprisingly, the probability of finding a potential ortholog with a functional association in our dataset is much higher for poorly-annotated proteins from bacterial divisions that we studied multiple representatives of (α,β,γ-Proteobacteria and Bacteroidetes) than for such proteins from other bacteria (21% vs. 7%). For bacteria from well-covered divisions, we provide potential orthologs with functional associations for about 320 poorly-annotated proteins per average genome (3,876 proteins per genome * 39% poorly-annotated * 21%). To enable rapid access to the phenotypes and functional associations of the homologs of a protein of interest, we provide a "Fitness BLAST" web service (http://fit.genomics.lbl.gov/images/fitblast_example.html). The results from Fitness BLAST are available within the protein pages at IMG/M [53] and MicrobesOnline [50]. We also provide a web page for comparing all of the proteins in a bacterium to the fitness data. In summary, our data and comparative analysis tools are valuable resources to aid in the annotation of protein function across the bacterial tree of life.

References

1. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 2006 0008 (2006).
2. Romine, M. F., Carlson, T. S., Norbeck, A. D., McCue, L. A. & Lipton, M. S. Identification of mobile elements and pseudogenes in the Shewanella oneidensis MR-1 genome. *Appl Environ Microbiol* **74,** 3257–3265 (2008).
3. Deutschbauer, A. *et al.* Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. *PLoS Genet.* **7,** e1002385 (2011).
4. Rubin, B. E. *et al.* The essential gene set of a photosynthetic organism. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E6634–43 (2015).
5. Kato, J.-I. & Hashimoto, M. Construction of consecutive deletions of the Escherichia coli chromosome. *Mol. Syst. Biol.* **3,** 132 (2007).
6. Yamamoto, N. *et al.* Update on the Keio collection of Escherichia coli single-gene deletion mutants. *Mol. Syst. Biol.* **5,** 335 (2009).
7. Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* **41,** D387–95 (2013).
8. Lévêque, F., Gazeau, M., Fromant, M., Blanquet, S. & Plateau, P. Control of Escherichia coli lysyl-tRNA synthetase expression by anaerobiosis. *J. Bacteriol.* **173,** 7903–7910 (1991).
9. Cook, G. M. *et al.* A factor produced by Escherichia coli K-12 inhibits the growth of E. coli mutants defective in the cytochrome bd quinol oxidase complex: enterochelin rediscovered. *Microbiology (Reading, Engl.)* **144 (Pt 12),** 3297–3308 (1998).
10. Kimura, S., Hubbard, T. P., Davis, B. M. & Waldor, M. K. The Nucleoid Binding Protein H-NS Biases Genome-Wide Transposon Insertion Landscapes. *MBio* **7,** e01351–16 (2016).
11. Romine, M. F. *et al.* Validation of Shewanella oneidensis MR-1 small proteins by AMT tag-based proteome analysis. *OMICS* **8,** 239–254 (2004).
12. Justice, S. S., Hunstad, D. A., Cegelski, L. & Hultgren, S. J. Morphological plasticity as a bacterial survival strategy. *Nature Publishing Group* **6,** 162–168 (2008).
13. da Rocha, R. P., Paquola, A. C. de M., Marques, M. D. V., Menck, C. F. M. & Galhardo, R. S. Characterization of the SOS regulon of Caulobacter crescentus. *J. Bacteriol.* **190,** 1209–1218 (2008).
14. Abella, M., Campoy, S., Erill, I., Rojo, F. & Barbé, J. Cohabitation of two different lexA regulons in Pseudomonas putida. *J. Bacteriol.* **189,** 8855–8862 (2007).
15. Cirz, R. T., O apos Neill, B. M., Hammond, J. A., Head, S. R. & Romesberg, F. E. Defining the Pseudomonas aeruginosa SOS response and its role in the global response to the antibiotic ciprofloxacin. *J. Bacteriol.* **188,** 7101–7110 (2006).
16. Keseler, I. M. *et al.* EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **41,** D605–12 (2013).
17. Byrne, R. T., Chen, S. H., Wood, E. A., Cabot, E. L. & Cox, M. M. Escherichia coli

genes and pathways involved in surviving extreme exposure to ionizing radiation. *J. Bacteriol.* **196,** 3534–3545 (2014).
18. Gagarinova, A. *et al.* Systematic Genetic Screens Reveal the Dynamic Global Functional Organization of the Bacterial Translation Machinery. *Cell Rep* **17,** 904–916 (2016).
19. Hostetter, A. A., Osborn, M. F. & DeRose, V. J. RNA-Pt adducts following cisplatin treatment of Saccharomyces cerevisiae. *ACS Chem. Biol.* **7,** 218–225 (2012).
20. Iwamoto, R. & Imanaga, Y. Direct evidence of the Entner-Doudoroff pathway operating in the metabolism of D-glucosamine in bacteria. *J. Biochem.* **109,** 66–69 (1991).
21. Iwamoto, R., Taniki, H., Koishi, J. & Nakura, S. D-glucosaminate aldolase activity of D-glucosaminate dehydratase from Pseudomonas fluorescens and its requirement for Mn2+ ion. *Biosci Biotechnol Biochem* **59,** 408–411 (1995).
22. Iwamoto, R., Amano, C., Ikehara, K. & Ushida, N. The D-glucosaminate dehydratase alpha-subunit from Pseudomonas fluorescens exhibits thioredoxin reductase activity. *Biochim Biophys Acta* **1647,** 310–314 (2003).
23. Brouns, S. J. J. *et al.* Identification of the missing links in prokaryotic pentose oxidation pathways: evidence for enzyme recruitment. *J Biol Chem* **281,** 27378–27388 (2006).
24. Hobbs, M. E. *et al.* Discovery of an L-fucono-1,5-lactonase from cog3618 of the amidohydrolase superfamily. *Biochemistry* **52,** 239–253 (2013).
25. Watanabe, S., Shimada, N., Tajima, K., Kodaki, T. & Makino, K. Identification and characterization of L-arabonate dehydratase, L-2-keto-3-deoxyarabonate dehydratase, and L-arabinolactonase involved in an alternative pathway of L-arabinose metabolism. Novel evolutionary insight into sugar metabolism. *J Biol Chem* **281,** 33521–33536 (2006).
26. Yew, W. S. *et al.* Evolution of enzymatic activities in the enolase superfamily: L-fuconate dehydratase from Xanthomonas campestris. *Biochemistry* **45,** 14582–14597 (2006).
27. Bateman, R. L. *et al.* Mechanistic inferences from the crystal structure of fumarylacetoacetate hydrolase with a bound phosphorus-based inhibitor. *J Biol Chem* **276,** 15284–15291 (2001).
28. Zhao, J. & Binns, A. N. Characterization of the mmsAB-araD1 (gguABC) genes of Agrobacterium tumefaciens. *J. Bacteriol.* **193,** 6586–6596 (2011).
29. Aro-Kärkkäinen, N. *et al.* L-arabinose/D-galactose 1-dehydrogenase of Rhizobium leguminosarum bv. trifolii characterised and applied for bioconversion of L-arabinose to L-arabonate with Saccharomyces cerevisiae. *Appl. Microbiol. Biotechnol.* **98,** 9653–9665 (2014).
30. Stephens, C. *et al.* Genetic analysis of a novel pathway for D-xylose metabolism in Caulobacter crescentus. *J. Bacteriol.* **189,** 2181–2185 (2007).
31. Gimenez, R., Nuñez, M. F., Badia, J., Aguilar, J. & Baldoma, L. The gene yjcG, cotranscribed with the gene acs, encodes an acetate permease in Escherichia coli. *J. Bacteriol.* **185,** 6448–6455 (2003).
32. Goonesekere, N. C. W., Shipely, K. & O apos Connor, K. The challenge of

annotating protein sequences: The tale of eight domains of unknown function in Pfam. *Comput Biol Chem* **34,** 210–214 (2010).
33. Skerker, J. M. *et al.* Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates. *Mol. Syst. Biol.* **9,** 674–674 (2013).
34. Krehenbrink, M., Oppermann-Sanio, F.-B. & Steinbüchel, A. Evaluation of non-cyanobacterial genome sequences for occurrence of genes encoding proteins homologous to cyanophycin synthetase and cloning of an active cyanophycin synthetase from Acinetobacter sp. strain DSM 587. *Arch Microbiol* **177,** 371–380 (2002).
35. Adames, K., Euting, K., Bröker, A. & Steinbüchel, A. Investigations on three genes in Ralstonia eutropha H16 encoding putative cyanophycin metabolizing enzymes. *Appl. Microbiol. Biotechnol.* **97,** 3579–3591 (2013).
36. Bateman, A. & Finn, R. D. SCOOP: a simple method for identification of novel protein superfamily relationships. *Bioinformatics* **23,** 809–814 (2007).
37. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
38. Tagourti, J., Landoulsi, A. & Richarme, G. Cloning, expression, purification and characterization of the stress kinase YeaG from Escherichia coli. *Protein Expr. Purif.* **59,** 79–85 (2008).
39. Figueira, R. *et al.* Adaptation to sustained nitrogen starvation by Escherichia coli requires the eukaryote-like serine/threonine kinase YeaG. *Sci Rep* **5,** 17524 (2015).
40. Meyer, B. *et al.* The energy-conserving electron transfer system used by Desulfovibrio alaskensis strain G20 during pyruvate fermentation involves reduction of endogenously formed fumarate and cytoplasmic and membrane-bound complexes, Hdr-Flox and Rnf. *Environ. Microbiol.* n–a–n–a (2014). doi:10.1111/1462-2920.12405
41. Price, M. N. *et al.* The genetic basis of energy conservation in the sulfate-reducing bacterium Desulfovibrio alaskensis G20. *Front Microbiol* **5,** 577 (2014).
42. Gennaris, A. *et al.* Repairing oxidized proteins in the bacterial envelope using respiratory chain electrons. *Nature* **528,** 409–412 (2015).
43. Melnyk, R. A. *et al.* Novel mechanism for scavenging of hypochlorite involving a periplasmic methionine-rich Peptide and methionine sulfoxide reductase. *MBio* **6,** e00233–15 (2015).
44. Rückert, C. *et al.* Functional genomics and expression analysis of the Corynebacterium glutamicum fpr2-cysIXHDNYZ gene cluster involved in assimilatory sulphate reduction. *BMC Genomics* **6,** 121 (2005).
45. Shetty, A., Chen, S., Tocheva, E. I., Jensen, G. J. & Hickey, W. J. Nanopods: a new bacterial structure and mechanism for deployment of outer membrane vesicles. *PLoS One* **6,** e20725 (2011).
46. Navazo, A. *et al.* Three independent signalling pathways repress motility in Pseudomonas fluorescens F113. *Microb Biotechnol* **2,** 489–498 (2009).
47. Hickey, W. J., Chen, S. & Zhao, J. The phn Island: A New Genomic Island

Encoding Catabolism of Polynuclear Aromatic Hydrocarbons. *Front Microbiol* **3,** 125 (2012).
48. Horinouchi, M., Kurita, T., Hayashi, T. & Kudo, T. Steroid degradation genes in Comamonas testosteroni TA441: Isolation of genes encoding a Δ4(5)-isomerase and 3α- and 3β-dehydrogenases and evidence for a 100 kb steroid degradation gene hot spot. *J. Steroid Biochem. Mol. Biol.* **122,** 253–263 (2010).
49. Kuznetsova, E. *et al.* Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS Microbiol. Rev.* **29,** 263–279 (2005).
50. Dehal, P. S. *et al.* MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.* **38,** D396–400 (2010).
51. Lerat, E., Daubin, V., Ochman, H. & Moran, N. A. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* **3,** e130 (2005).
52. Clark, W. T. & Radivojac, P. Analysis of protein function and its prediction from amino acid sequence. *Proteins* **79,** 2086–2096 (2011).
53. Markowitz, V. M. *et al.* IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* **42,** D560–7 (2014).