JB
Journals.ASM.org

Towards an Informative Mutant Phenotype for Every Bacterial Gene

Adam Deutschbauer,[a] Morgan N. Price,[a] Kelly M. Wetmore,[a] Daniel R. Tarjan,[b,c] Zhuchen Xu,[d] Wenjun Shao,[a] Dacia Leon,[b,c] Adam P. Arkin,[a,c,d] Jeffrey M. Skerker[a,c,d]

Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA[a]; Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA[b]; Energy Biosciences Institute, University of California, Berkeley, California, USA[c]; Department of Bioengineering, University of California, Berkeley, California, USA[d]

Mutant phenotypes provide strong clues to the functions of the underlying genes and could allow annotation of the millions of sequenced yet uncharacterized bacterial genes. However, it is not known how many genes have a phenotype under laboratory conditions, how many phenotypes are biologically interpretable for predicting gene function, and what experimental conditions are optimal to maximize the number of genes with a phenotype. To address these issues, we measured the mutant fitness of 1,586 genes of the ethanol-producing bacterium *Zymomonas mobilis* ZM4 across 492 diverse experiments and found statistically significant phenotypes for 89% of all assayed genes. Thus, in *Z. mobilis*, most genes have a functional consequence under laboratory conditions. We demonstrate that 41% of *Z. mobilis* genes have both a strong phenotype and a similar fitness pattern (cofitness) to another gene, and are therefore good candidates for functional annotation using mutant fitness. Among 502 poorly characterized *Z. mobilis* genes, we identified a significant cofitness relationship for 174. For 57 of these genes without a specific functional annotation, we found additional evidence to support the biological significance of these gene-gene associations, and in 33 instances, we were able to predict specific physiological or biochemical roles for the poorly characterized genes. Last, we identified a set of 79 diverse mutant fitness experiments in *Z. mobilis* that are nearly as biologically informative as the entire set of 492 experiments. Therefore, our work provides a blueprint for the functional annotation of diverse bacteria using mutant fitness.

Assigning function to the millions of hypothetical and uncharacterized genes identified by genome sequencing projects is a substantial challenge in the postgenome era (1, 2). This problem is compounded in bacteria due to the ease of genome sequencing and the vast reservoir of genetic diversity contained in prokaryotes. Therefore, high-throughput experimental approaches are necessary to bridge the gap between genome sequencing and genome characterization (3). One promising strategy is the use of high-throughput mutagenesis to predict gene function based on the observation that genes with similar functions tend to have similar growth phenotypes (4–9). In single-cell organisms, the most commonly used approach for large-scale mutagenesis and phenotyping involves pooling thousands of individual mutant strains and parallel analysis of their abundance using either DNA microarrays (10, 11) or sequencing (12–14). These approaches are advantageous because they produce quantitative measures of fitness for all nonessential genes in a single-pot assay. In the single-cell eukaryote *Saccharomyces cerevisiae*, assaying mutant fitness for all nonessential genes under hundreds of laboratory conditions (primarily growth in the presence of drugs and other small-molecule inhibitors) identified a significant phenotype for nearly every protein-coding gene (7). However, for bacteria, it remains unclear what fraction of the genome has a phenotype under laboratory conditions, with estimates ranging from 50% in *Escherichia coli* (8) to 70% in *Shewanella oneidensis* MR-1 (4). In addition, the number of genes with a pattern of mutant fitness that is biologically interpretable for predicting gene function and the optimal set of experimental conditions for maximizing new gene annotations have yet to be established.

To address these issues, we performed 492 genome-wide mutant fitness assays in *Zymomonas mobilis* ZM4, a fermentative, ethanol-producing bacterium (15). We find that 89% of all assayed *Z. mobilis* genes, including many genes without a specific annotation, have a statistically significant phenotype when disrupted in the laboratory. However, many genes have subtle phenotypes under just a few conditions, and it is not obvious how these phenotypes relate to each gene's function. To determine whether our findings in *Z. mobilis* are generalizable to other bacteria, we calculated the fraction of *Shewanella oneidensis* MR-1 genes with a detectable phenotype by using the same experimental strategy. *S. oneidensis* has diverse respiratory abilities, including metal reduction, whereas *Z. mobilis* obtains energy only by fermentation, and the genome of *S. oneidensis* is substantially larger than the genome of *Z. mobilis* (4,467 versus 1,892 protein-coding genes). *S. oneidensis* also has a wider range of metabolic abilities: we have confirmed the growth of *S. oneidensis* on 25 carbon sources, compared to just 3 for *Z. mobilis* (16). By analyzing 296 *S. oneidensis* fitness experiments, we found that 75% of assayed genes exhibited a significant phenotype.

Previous studies have noted that genes with related functions often have similar fitness patterns (4, 8), and we observed this in our data as well. Thus, to estimate the fraction of genes that have a biologically informative phenotype, we identified genes that have a strong mutant phenotype under at least one condition and also

Received 8 May 2014 Accepted 1 August 2014
Published ahead of print 11 August 2014
Address correspondence to Adam P. Arkin, APArkin@lbl.gov, or Jeffrey M. Skerker, skerker@berkeley.edu.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JB.01836-14.

doi:10.1128/JB.01836-14
The authors have paid a fee to allow immediate free access to this article.

have correlated fitness patterns (cofitness) with another gene. With these criteria, we found that 41% of *Z. mobilis* genes are candidates for functional annotation using mutant fitness, including 174 genes currently without a specific annotation. For 57 of these poorly annotated genes, we used a combination of comparative genomics and homology evidence to demonstrate that the cofitness-derived gene-gene associations are biologically meaningful. Last, many of the 492 experiments were conducted under similar conditions and thus gave similar results. After removing redundant experiments, only 79 diverse mutant fitness experiments remained, and these sufficed to identify the majority of strong phenotypes and biological associations. Similarly, we found that 296 fitness experiments for *S. oneidensis* could be reduced to 61 experiments. In sum, this work provides a blueprint for mutant fitness-based gene annotation in a wide range of bacteria.

MATERIALS AND METHODS

Strains and media. *Zymomonas mobilis* ZM4 (ATCC 31821) and *Shewanella oneidensis* MR-1 (ATCC 700550) were purchased from ATCC. For typical culturing, we grew *Z. mobilis* in rich ZRMG medium (25 g/liter glucose, 10 g/liter yeast extract, and 2 g/liter KH_2PO_4) and *S. oneidensis* in Luria-Bertani broth (LB). When necessary, we added kanamycin to a final concentration of 100 μg/ml for *Z. mobilis* and 50 μg/ml for *S. oneidensis*. Unless indicated otherwise, we grew both bacteria aerobically at 30°C. We used previously described transposon mutant collections for *Z. mobilis* (17) and *S. oneidensis* (4). These transposon mutants contain DNA bar codes (tags) that enable the pooling and parallel analysis of mutant fitness (11).

Pooled mutant fitness assays. Pooled mutant fitness assays were performed as previously described for both *Z. mobilis* (17) and *S. oneidensis* (4). Briefly, for each bacterium, we assayed two pools of transposon mutants per experimental condition, before and after growth (usually for six to eight population doublings). For all experiments, DNA bar code abundance was monitored with an Affymetrix microarray (GenFlex_16K_v2) containing the tag sequence complements (18). In this study, the majority of mutant fitness assays were performed in rich media with an inhibitory but sublethal concentration of a single chemical. For each inhibitor, we identified the appropriate concentration(s) for the pooled fitness assay by measuring the growth of the wild-type bacterium in a 96-well microplate. These prescreen assays were cultured in a microplate reader (either Tecan Sunrise or Infinite F200) with absorbance (optical density at 600 nm [OD_{600}]) readings every 15 min. We typically aimed for a concentration of inhibitor that resulted in a 50% reduction of wild-type growth rate. In practice, we often profiled the fitness of the mutants at multiple concentrations of the same compound. Data sets S1 and S2 in the supplemental material contain detailed information on all *Z. mobilis* and *S. oneidensis* fitness experiments. All data are available at http://genomics.lbl.gov/supplemental/phenotypes2013/.

Mutant fitness data analysis. Raw data processing, the calculation of strain fitness, the calculation of gene fitness, and data normalization were performed as previously described for *Z. mobilis* and *S. oneidensis* (4, 17). Briefly, for each experiment, we would like to estimate the functional consequence of disrupting each gene, given the data for various strains with insertions in that gene. As we will show, independent insertions in the same gene tend to give similar results, so we believe that most of the effects that we observe reflect the consequence of disrupting the gene. A related concern is that insertions in a gene could affect the expression of downstream genes via polar effects, but we will show that polar effects do not predominate. Thus, to quantify the functional consequence of disrupting each gene, we calculate the "gene fitness," which is the average of each mutant strain for the gene. The fitness of a strain is its change in $\log_2$ abundance during the course of an experiment (typically 6 to 8 generations) and is analogous to a $\log_2$ ratio in expression experiments (4).

These fitness values are normalized so that a typical gene or strain has a fitness of zero. We normalized the fitness values to control for effects from chromosomal position, artifacts from mutant pool construction, and scaffold effects (plasmid versus chromosome). Additionally, for the main chromosome of both bacteria, we set the mode of the strain fitness distribution to zero. All fitness data for both *Z. mobilis* and *S. oneidensis* are publically available in MicrobesOnline (19) and are available as Data sets S3 and S4 in the supplemental material.

Because the strain fitness values for independent transposon insertions in the same gene are highly correlated ($R = 0.86$ for strains with central insertions in the 5 to 80% of coding region, average of 30 unamended rich medium fitness experiments for *Z. mobilis*), our "gene fitness" values reflect the mutant fitness of the individual strains and the impact of knocking out the gene, which is the focus of this study. As another test of the agreement of independent insertions in the same gene, we measured the correlation of fitness for pairs of strains with insertions in the same gene (strain cofitness) across all 492 experiments in *Z. mobilis*. We found a strong correlation (median $R = 0.61$) for insertions in *Z. mobilis* genes with reduced fitness phenotypes (see Fig. S1 in the supplemental material).

After the fitness values were normalized, we calculated a test statistic for each gene in each experiment that takes into account the consistency of measurements within that experiment, as previously described (4). The test statistic (t) was calculated as follows:

$$t = \mu / \sqrt{V/n}$$

$$V = \{\Psi^2 + \Sigma(x - \mu)^2\}/n$$

where x is the measurement(s) for the gene, μ is their average, n is the number of measurements, V is the variance of strain fitness, and Ψ is median[STD(x)], the median across all genes with more than one measurement of the standard deviation (STD) of that gene's measurements (4). This test statistic was transformed to P values using 17 independent start experiments as a comparison; this was done separately for *Z. mobilis* and *S. oneidensis*. The test statistic was transformed independently for genes with $n = 1, 2, 3$, or 4 or more. These P values represent the significance of the gene's fitness within a single experiment.

To increase our sensitivity for detecting more mild phenotypes, we grouped the fitness experiments (separately for each bacterium) by overall similarity. Specifically, fitness experiments were grouped using hierarchical agglomerative clustering with complete linkage (hclust in R) and with "1 − correlation" as the distance metric. The clustering was cut at a depth of 0.25 (cutree in R), which corresponds to requiring that each pair of experiments in a group have a correlation of 0.75 or greater.

After the test statistic was transformed, the P value for a gene in a given experiment ranges from 0 to 1, with values close to 0 or 1 indicating confidence that the gene's fitness was negative or positive, respectively. To increase the sensitivity for detecting phenotypes, the significance values for each gene within a group were combined using Fisher's combined probability test to give a combined P value (P_{comb}). For each gene and each group of experiments, we used a two-sided test and corrected for the number of experiment groups. Specifically, a gene's phenotype was considered significant if $P_{comb} < 0.05/(2 \times \text{number of groups})$ or $P_{comb} > 1 - [0.05/(2 \times \text{number of groups})]$. Because of uncertainty in the normalization of the fitness data, which implies that the typical strain that has no phenotype might be assigned a fitness slightly below or above zero, we also required that the average fitness of the gene within a group be below −0.2 or above 0.2.

As a second method to assess the significance of phenotypes in *Z. mobilis*, we used a two-tailed t test that does not depend on the test statistic or its transformation. For this analysis, we used the 54 *Z. mobilis* experimental groups with three or more experiments (see Data set S1 in the supplemental material). The t test was used to generate a P value for the hypothesis that the average fitness value for a gene within that group is equal to zero. The source code used for all statistical analyses is available at http://genomics.lbl.gov/supplemental/phenotypes2013/.

***Z. mobilis* gene expression.** We measured gene expression using high-density tiling Nimblegen microarrays for *Z. mobilis* grown in rich ZRMG and defined ZMMG media (17). We harvested total RNA during exponential growth using the RNeasy kit (Qiagen). For the tiling arrays, enrichment for mRNA, cDNA synthesis, labeling, and hybridization were performed as previously described (20). The tiling array data were normalized so that the median probe has a log level of zero (20). As most of the genome is expressed on one strand or the other, zero will correspond to the high end of background expression.

Comparative genomics. Orthologs between *Z. mobilis* ZM4 and *Caulobacter crescentus* NA1000 were determined using MicrobesOnline tree orthologs (19). Likewise, the analyses of conserved synteny and of InterPro hits, including hits to Pfam domains that are annotated as domains of unknown function (http://pfam.janelia.org/), are taken from MicrobesOnline.

Microarray data accession number. The *Z. mobilis* tiling microarray data are publically available (GEO accession no. GSE51870).

RESULTS

Eighty-nine percent of assayed *Zymomonas mobilis* genes have a phenotype. To determine the fraction of bacterial genes with an identifiable phenotype, we used the alphaproteobacterium *Zymomonas mobilis* ZM4, which has the advantages of a small genome size (1,892 protein-coding genes) (21) and the availability of a DNA bar-coded transposon mutant collection for the quantitative and parallel analysis of mutant fitness (17). Using two previously described mutant pools of *Z. mobilis* covering 1,586 genes (83% of protein-coding genes) and genome-wide fitness data in 202 growth experiments as the starting point (17, 22), we performed an additional 290 pooled fitness assays, including growth during inhibition with various antibiotics, metals, and salts, and growth with alternative carbon and nitrogen sources, anaerobic growth, and survival after UV irradiation (for a full list and annotation of the 492 *Z. mobilis* experiments, see Data set S1 in the supplemental material). The 290 additional *Z. mobilis* fitness experiments were chosen to be diverse, including stresses with different modes of action, to maximize the likelihood of identifying phenotypes for all genes. The entire *Z. mobilis* fitness data set is clustered and summarized as a heat map in Fig. 1A. Gene fitness is defined as the $\log_2$ change in the abundance of strains with insertions in the gene: negative values indicate that the gene is beneficial for fitness and that strains with the mutated gene have reduced fitness, while positive fitness values indicate that mutating the gene leads to improved fitness relative to the typical strain in the pools and that the gene's activity is detrimental to fitness.

As a representative illustration of the *Z. mobilis* mutant fitness data, a genome-wide comparison of "gene fitness" for two conditions, rich medium supplemented with the DNA-damaging agent cisplatin and rich medium with no supplements, is highlighted in Fig. 1B. The nucleotide excision repair complex genes *uvrABCD* (23) and the RecA-mediated double-strand break repair genes *recFGORX* (24) are beneficial for optimal fitness in the presence of the inhibitor but not in rich medium without cisplatin. In *E. coli*, strains with mutations in genes in both the double-strand break recombination and nucleotide excision repair pathways are also hypersensitive to cisplatin (25). *Z. mobilis recA* mutants have reduced fitness in both the presence and absence of cisplatin (Fig. 1B), which likely reflects the multiple biochemical roles of RecA protein in recombination, DNA repair, and regulation (26).

An examination of the Fig. 1A heat map reveals a large block of 481 genes with strongly reduced fitness in nearly all experiments (blue at the top of the heat map in Fig. 1A). The median gene in this block has a strongly reduced-fitness phenotype (fitness less than −1) in 350 experiments (out of a possible 492). This block includes many ribosomal proteins and other genes that are expected to be essential. *Z. mobilis* appears to be polyploid, and insertions are as likely to occur in essential genes as in other genes (17). The mutants with insertions in essential genes do not have segmental duplications; rather, they are unstable heterozygotes (17), which explains why these strains drop in abundance during the fitness experiment. The high rate of insertions in essential genes also implies that the 306 *Z. mobilis* genes without data are not significantly enriched for essential genes but rather reflect a largely random group of genes that by chance we did not map transposon mutants in (17). Many of these predicted essential genes are beneficial for fitness in nearly all experiments: 196 of the 481 frequently beneficial genes in the blue block at the top of Fig. 1A are predicted essential genes (based on orthology to essential *Caulobacter* genes [27]; for a full list of the genes in this cluster and whether they are predicted to be essential, see Data set S5 in the supplemental material). More broadly, genes with strong reduced fitness phenotypes in many conditions are clearly important for organismal fitness, and detecting phenotypes for these genes is straightforward. However, previous work has established that most bacterial genes do not have such obvious phenotypes (4, 8). Rather, we expect many phenotypes to be subtle and manifested in only a subset of our 492 experiments.

To increase sensitivity for detecting mild phenotypes, we clustered 95% (465 of 492) of the fitness experiments into 79 groups (Fig. 1A). These groups represent experiments with highly correlated genome-wide fitness (pairwise correlations greater than 0.75 for all members of the group) and are listed in Data set S1 in the supplemental material. The two largest experimental groups, with 40 and 29 experiments respectively, are rich medium with no stress and rich medium with little stress (i.e., low concentration of added inhibitor). The next biggest group (23 experiments) includes a variety of alcohols and aldehydes and growth at 40°C. Overall, the groups frequently contain structurally related compounds, compounds with similar modes of action, or the same compound at different concentrations. For example, group 31 includes two doxycycline and three minocycline experiments (doxycycline and minocycline are structurally similar tetracycline antibiotics), group 25 includes six aminoglycoside antibiotic experiments (tobramycin, sisomicin, or gentamicin), and group 73 contains two bacitracin experiments at different concentrations. Our finding that compounds with similar structures or modes of actions have correlated genome-wide fitness patterns is consistent with previous findings in both bacteria (8, 17) and yeast (28).

To systematically determine the fraction of the *Z. mobilis* genome with a statistically significant phenotype, we used a test statistic for each gene under each condition that takes into account the consistency of measurements for that gene as well as for other genes in that experiment (4). We converted this test statistic to *P* values by using control experiments, we combined these *P* values across similar experiments, and we corrected for multiple testing across 79 groups (see Materials and Methods for details). At a cutoff of $P < 0.05$, we found that 1,090 (69%) genes are beneficial to fitness and 855 (54%) are detrimental to fitness in at least one of the 79 groups of experiments. Overall, 1,409 genes, or 89% of the genes we have data for, have either a significant reduced fitness or enhanced-fitness phenotype based on this analysis. The false dis-

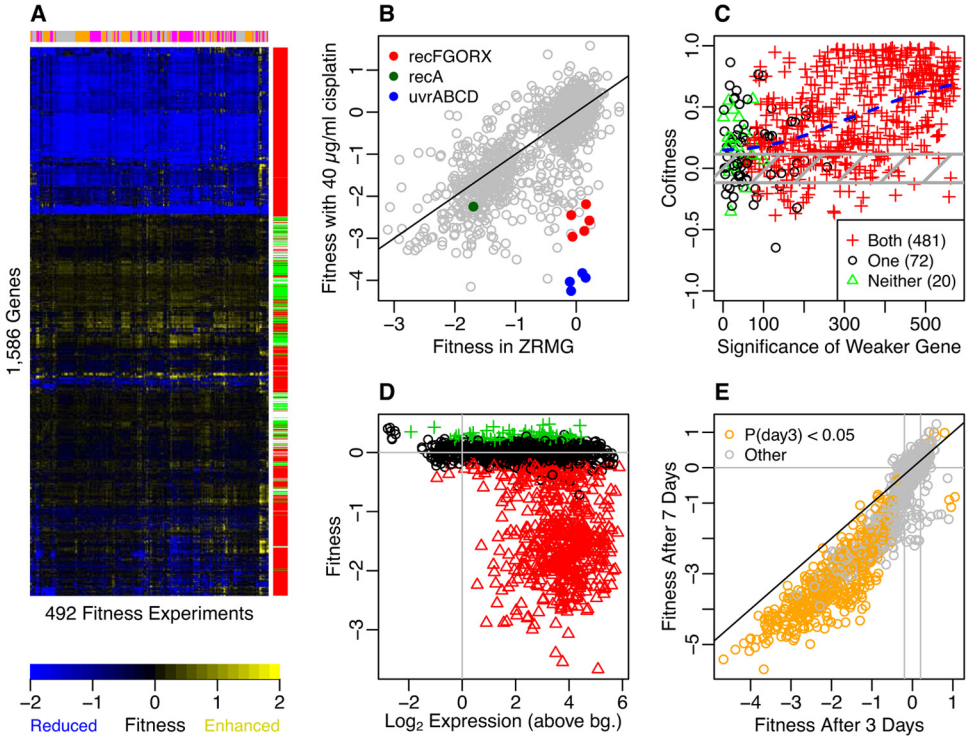

FIG 1 Identifying a phenotype for most *Z. mobilis* genes. (A) Heat map of clustered mutant fitness data for 1,586 genes (*y* axis) across 492 experiments (*x* axis). Reduced fitness values are shown in blue, and enhanced fitness values are shown in yellow (see color key). The experiments are binned into 79 groups (alternating colors on the *x* axis) to increase statistical power for detecting subtle phenotypes (see Materials and Methods). Genes are color-coded to the right of the heat map according to whether they are beneficial for fitness in any group of experiments (red), detrimental to fitness in a group of experiments and never beneficial (green), or have no statistically significant phenotype in any group of experiments (no color). (B) Scatterplot of gene fitness values in rich medium (ZRMG medium; *x* axis) versus rich medium supplemented with an inhibitory concentration of cisplatin (*y* axis). Negative values are indicative of reduced fitness relative to the typical strain in the mutant pools. Genes encoding members of the UvrABCD nucleotide excision repair system, RecA, and RecFGORX are highlighted. The solid black line shows $x = y$. (C) Correlation of fitness (cofitness on the *y* axis) for 573 pairs of adjacent genes that are predicted to be cotranscribed in an operon. The pairs are ranked by the most significant phenotype of the weaker gene in any of the 79 groups of experiments (from weakest to strongest phenotype; *x* axis). Cofitness values are colored according to whether both genes in the pair have a significant phenotype (red), only one gene in the pair has a significant phenotype (black), or neither gene has a significant phenotype (green). The gray hatched region covers 99% of the cofitness distribution from shuffled data (−0.117 to 0.115). The dashed blue line represents the best-fit smooth line through the data (local regression from loess). (D) Comparison of gene fitness in rich medium (ZRMG medium; *y* axis) and expression level in the same condition (*x* axis). Expression was determined using a high-resolution tiling microarray and is plotted as the $\log_2$ level relative to background (bg.) (see Materials and Methods). Genes with significantly reduced (red) or enhanced (green) phenotypes after 1 day (~6 population doublings) of growth in ZRMG medium ($P < 0.001$ by Fisher test with 30 replicates) are indicated. (E) Comparison of gene fitness for 1,586 genes after 3 days (~18 population doublings) (*x* axis) or 7 days (~42 population doublings) (*y* axis) of batch transfer growth in rich medium (cells were diluted back in fresh medium each day). The solid black line shows $x = y$. The vertical gray lines represent fitness of −0.2 and 0.2. Genes with a significant phenotype after 3 days of growth in rich medium ($P < 0.05$, based on the transformed test statistic for this single experiment) are shown in orange.

covery rate for *Z. mobilis* genes with phenotypes using this analysis is 5.6% or less than 80 genes.

To illustrate how the grouping of experiments in the above analysis provides increased statistical power for detecting phenotypes, we highlight a specific example. Mutations in the dehydrogenase *ZMO0226* have reduced fitness in three experiments with different concentrations of the uncoupling agent carbonyl cyanide-*p*-trifluoromethoxyphenylhydrazone (FCCP), with fitness values of −1.4 to −2.1 (average measurement for five different strains with transposon insertions in *ZMO0226*). The lowest *P* value from any single experiment was 0.0071, but after correcting for multiple testing across 492 experiments, this is not meaningful (corrected $P > 1$). The combined *P* value from the three FCCP experiments (which clustered into a single group) was 0.00022, or 0.018 after correcting for multiple testing across 79 groups.

To control for potential bias in the above analysis, we performed a two-tailed *t* test on the normalized fitness values that is independent of the test statistic and *P* value transformation described above (see Materials and Methods). With a correction for multiple testing and a *P* value cutoff of less than 0.05, 75% of the genes have a reduced-fitness phenotype and 43% have an enhanced fitness phenotype. Based on this *t* test analysis, 1,492 genes, or 94% of *Z. mobilis* genes we have data for, have a significant phenotype (reduced or enhanced) with a false discovery rate of 5.3%. Therefore, regardless of the statistical test, the vast majority of *Z. mobilis* genes assayed (89% or 94%) have a detectable phenotype in our large mutant fitness compendium.

Multiple lines of evidence suggest that the statistically significant but subtle phenotypes identified by our analyses are bona fide phenotypes and not artifacts of our experimental strategy or analysis. First, genes with significant phenotypes tend to have higher correlations in fitness (cofitness) with genes in the same operon across all experiments, as expected given that genes in the same operon often have related functions (Fig. 1C). Even genes with the

weakest significant phenotypes tend to have higher cofitness within operons than expected compared to shuffled data (Fig. 1C). Because operon gene pairs can also have unrelated functions (29–31), there are multiple instances where one or both adjacent genes have significant phenotypes but have near zero or negative cofitness (Fig. 1C). Operon cofitness could also be due to polarity effects in our data set. Polarity occurs when a transposon insertion in an upstream gene of the operon leads to transcriptional termination and reduced expression of a downstream gene(s). Polarity can be detected genome-wide by a significantly higher fraction of instances where only an upstream gene has a phenotype relative to instances where only the downstream gene has a phenotype (4). Applying this test to our fitness data set, we find a moderate increase of upstream-only reduced fitness relative to downstream-only reduced fitness (5,709 versus 4,345 in individual fitness experiments; $P < 10^{-15}$ by a binomial test). Therefore, while polarity influences our data set, it is not an overwhelming effect and does not substantially change our estimate of the number of genes in *Z. mobilis* with a phenotype.

A second line of evidence in support of our estimate of genes with phenotypes is that, for a single condition, even those genes with subtle phenotypes tend to be well expressed (Fig. 1D). The fact that many genes with mild phenotypes are well expressed supports the results of our global analysis, as expression should be a prerequisite for a gene to exert a phenotypic effect.

Last, we validated the genes with weak but significant phenotypes by performing a long-term growth experiment in rich medium with batch transfer of the mutant libraries once per day. If weak phenotypes are real, then the fitness defects of strains with mutations in these genes should become more pronounced at later transfers. For weakly beneficial genes, there is a clear bias for reduced fitness phenotypes to become more severe after 7 days of growth relative to 3 days of growth (Fig. 1E). Taken together, the results of our analysis and experimental validation strongly support our estimate that 89% of the *Z. mobilis* genome has a detectable phenotype under laboratory conditions.

To determine whether a similar fraction of genes have a detectable phenotype in a second bacterium, we supplemented our previously reported 219 genome-wide *Shewanella oneidensis* fitness experiments (4, 22, 32) by performing an additional 77 experiments. Combined, the data from the 296 *S. oneidensis* fitness experiments represent a diverse range of metabolic and stress conditions that are comparable in size and diversity to the *Z. mobilis* data set. A subset of the fitness experiments is very similar (primarily the same stresses in rich medium) in both bacteria, while most experiments are unique to either *Z. mobilis* or *S. oneidensis*. Using the same test statistic, combined *P* value analysis, and thresholds for significance as applied to the *Z. mobilis* data set, we grouped 243 of the *S. oneidensis* fitness experiments into 61 groups by hierarchical clustering and identified 1,805 beneficial genes (out of 3,355 total genes with data, or 54%) with reduced fitness phenotypes and 1,895 (56%) detrimental genes with enhanced fitness phenotypes in at least one group of experiments. For the complete list of *S. oneidensis* mutant fitness experiments, medium compositions, and groups of experiments, see Data set S2 in the supplemental material. In total, 2,507 or 75% of the *S. oneidensis* genes that we have data for have a significant phenotype. The fraction of genes with a phenotype in *S. oneidensis* is only moderately less than that of *Z. mobilis* and may be explained by the larger size of the *Z. mobilis* fitness data set or the larger size of the *S. oneidensis* genome. Thus, our finding that the vast majority of genes in bacterial genomes have a detectable phenotype under laboratory conditions is likely generalizable.

Characteristics of *Z. mobilis* phenotypes. To uncover broader trends in the identified *Z. mobilis* phenotypes, we characterized the phenotypes based on their strength, directionality, occurrence in multiple conditions (pleiotropy), and the functional category of the genes. Overall, reduced-fitness phenotypes are much stronger than enhanced-fitness phenotypes, which fits the expectation that most mutations are detrimental to fitness (Fig. 2A). For example, 880 genes have significantly reduced fitness under −1 (in one or more of the 79 groups of experiments), but just 345 genes have significantly enhanced fitness above +1. Indeed, many of the significant enhanced-fitness phenotypes are weak: 201 of the 855 detrimental genes have a maximum fitness across groups of experiments less than 0.4. While not as pronounced, some of the reduced-fitness phenotypes are also weak; 77 of the 1,090 beneficial genes have a minimal fitness greater than −0.4. Assuming a large effective population size, a significant number of bacterial genes with weak reduced-fitness phenotypes should be expected (33). In this view, selection will maintain genes with very small beneficial effects that are difficult to measure in the laboratory. Alternatively, genes with no or only subtle phenotypes in our laboratory-based fitness compendium may play a crucial role under natural conditions, such as mediating interactions with other microorganisms. In this view, performing fitness assays under more-ecological conditions would uncover strong phenotypes for those genes with no or weak phenotypes in our laboratory data set.

Surprisingly, we found that 54% (855 of 1,586) of the *Z. mobilis* genes had a significant enhanced-fitness phenotype, and for 319 of these 855 genes, we identified an increase in fitness only for insertions in these genes. For a list of these 319 genes with only enhanced-fitness phenotypes, see Data set S6 in the supplemental material. Although these findings are consistent with recent reports that selection for increased laboratory fitness can drive gene loss in bacteria (34, 35), the extent and scale to which loss-of-function mutations lead to increased fitness in bacteria are only starting to be appreciated at the genome-wide level (36). In a study of *E. coli* mutant fitness data, Hottes and colleagues found that beneficial mutations were identified in nearly all conditions and that these mutations were enriched in genes encoding enzymes and regulatory proteins, suggesting that metabolic and regulatory rewiring via loss of function is a prevalent mechanism for fitness increases in the absence of new genes (36). However, in contrast to the recent *E. coli* results, *Z. mobilis* regulators, which are defined as transcription factors in the DNA-binding domain (DBD) database (37), are not significantly enriched among the detrimental gene set (odds ratio, 1.17; $P > 0.5$ by Fisher exact test). Furthermore, we find that *Z. mobilis* enzymes (defined as genes with an EC [Enzyme Commission] number assigned) are significantly less likely to be detrimental to fitness in the laboratory (odds ratio, 0.69; $P = 0.0005$). Rather, genes associated with amino acid transport and metabolism (COG [clusters of orthologous groups of proteins] function code E; false discovery rate = 0.04, after correcting for testing 20 functional categories) are significantly more likely to be detrimental to fitness in *Z. mobilis*. In *S. oneidensis*, genes with enhanced fitness phenotypes are significantly enriched for motility genes (COG function code N [cell motility]; false discovery rate of 3×10^{-5}) but not for regulators or enzymes ($P > 0.05$). Overall, among *E. coli*, *S. oneidensis*, and *Z. mobilis*, there

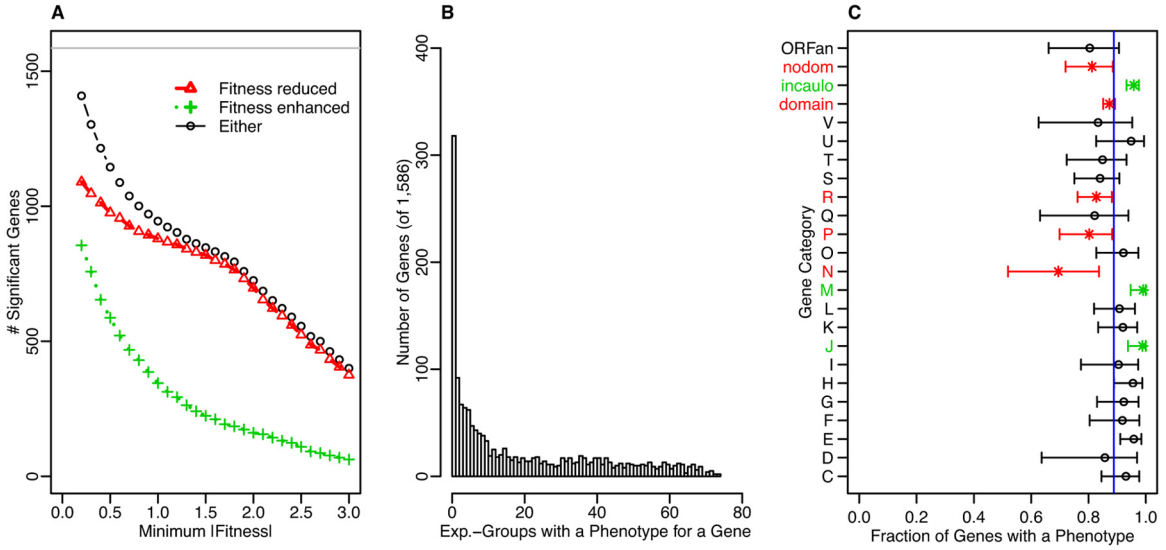

FIG 2 Characteristics of *Z. mobilis* phenotypes. (A) Comparison of the number of genes with a significant phenotype at different absolute fitness thresholds for genes with reduced fitness phenotypes (red), enhanced fitness phenotypes (green), or any phenotype (either; black). For example, at a fitness threshold of less than −1.0 in any of the 79 experimental groups, there are 880 beneficial genes (reduced fitness). Similarly, at a fitness threshold of greater than 1.0, there are 345 detrimental genes (enhanced fitness). The gray horizontal line marks 1,586, the total number of *Z. mobilis* genes we have data for. (B) Histogram of the number of genes (*y* axis) and their frequency of significant phenotypes among the 79 groups of experiments (*x* axis). (C) The fraction of *Z. mobilis* genes (*x* axis) with a significant phenotype among different categories (*y* axis). Genes are categorized as follows: "ORFan," no close homologs in any other bacterial genome; "nodom," no significant InterPro domain; "incaulo," presence of an ortholog in *Caulobacter crescentus*; "domain," other genes that contain an InterPro domain. The single letters indicate the COG (clusters of orthologous groups of proteins) categories: C (energy production and conversion), D (cell cycle control, cell division, and chromosome partitioning), E (amino acid transport and metabolism), F (nucleotide transport and metabolism), G (carbohydrate transport and metabolism), H (coenzyme transport and metabolism), I (lipid transport and metabolism), J (translation, ribosomal structure, and biogenesis), K (transcription), L (replication, recombination, and repair), M (cell wall/membrane/envelope biogenesis), N (cell motility), O (posttranslational modification, protein turnover, chaperones), P (inorganic ion transport and metabolism), Q (secondary metabolite biosynthesis, transport, and catabolism), R (general function prediction only), S (function unknown), T (signal transduction mechanisms), U (intracellular trafficking, secretion, and vesicular transport), and V (defense mechanisms). The vertical blue line represents the fraction of all *Z. mobilis* genes with a phenotype (0.89). The error bars show the 95% confidence intervals. Categories marked in green are significantly enriched for phenotypes (Fisher exact test, false discovery rate of <0.05), while those in red are significantly less likely to have phenotypes relative to the entire genome.

was no clear consistency as to which gene classes are more likely to be detrimental to laboratory fitness. However, the observation that over half of the genes we assayed in *Z. mobilis*, representing all functional categories, were detrimental in some condition strongly suggests that many mechanisms can lead to an enhanced-fitness phenotype in the laboratory (see Data set S6). For example, we found that a number of flagellar genes were detrimental to fitness only in our laboratory conditions. Increased fitness of motility mutants has been previously observed in other bacteria (12, 34), and may reflect an energetic advantage of being nonmotile in well-shaken laboratory experiments. As a second example, the putative metal ion transporter *ZMO0230* only has significant enhanced-fitness phenotypes, including in the presence of cobalt stress.

Given the diversity of experimental conditions we assayed, it is expected that some genes will exhibit pleiotropy in our large-scale fitness data set, as previously observed in yeast (5). Because many of our experiments are biological replicates or otherwise similar (structurally similar compounds or compounds with similar modes of actions), we investigated pleiotropy in *Z. mobilis* using the 79 groups of clustered experiments described above. We find that the mean *Z. mobilis* gene has a significant phenotype in 20 of the 79 experimental groups (Fig. 2B), with significant reduced- and enhanced-fitness phenotypes in 17 and 3 groups of conditions, respectively (Fig. 2B). Only 141 genes have a significant phenotype in just one experiment group, demonstrating that pleiotropy in bacteria under laboratory conditions is common if enough experimental conditions are assayed and that many bacterial genes have multiple functions or have a single functional role of key importance to multiple processes. One caveat of this analysis is that groups with more experiments will contain more genes with significant phenotypes (due to increased statistical power). To illustrate this point, we find that a relatively large percentage of all assayed *Z. mobilis* genes (43% [676 of 1,586 genes with data]) have a significant phenotype across 30 experiments in rich medium with no supplements ($P < 0.01$ by the combined P value test).

To identify classes of genes that are more or less likely to exhibit phenotypes, we calculated the fraction of genes with a significant phenotype among different categories (Fig. 2C). We find that nearly all of the *Z. mobilis* genes that have an ortholog in *Caulobacter crescentus*, also an alphaproteobacterium, have a phenotype in our compendium (95% [630 of 664]). Conversely, ORFans (genes without identifiable homologs in other bacteria) and hypothetical genes (with or without InterPro domains) are less likely to have significant phenotypes (Fig. 2C). However, even though ORFans and other hypothetical genes are less likely to have phenotypes relative to evolutionarily conserved genes, we still identified significant phenotypes for 80% of ORFans and 81% of domain-free proteins (Fig. 2C). Among functional categories, genes

associated with amino acid metabolism, translation, and the cell wall are significantly more likely to have a phenotype. Conversely, genes associated with inorganic ion transport/metabolism and motility, and genes with only a general function prediction are somewhat less likely to have a phenotype (Fig. 2C). The relative lack of phenotypes for motility-related genes might be attributable to the loss (or reduction) of this activity in the parental *Z. mobilis* ZM4 strain used in this study. Despite a myriad of motility experiments with the *Z. mobilis* ZM4 mutant pools, we identified clear phenotypes for only a fraction of the expected motility genes, and these phenotypes were typically less severe than those identified for motility genes in *S. oneidensis* MR-1 using similar assays (4). In addition to reduced motility of the parental strain [potentially due to an unknown mutation(s)], it is also possible that we did not identify the ideal conditions to induce motility in *Z. mobilis*.

To more systematically explore the properties of genes without a phenotype, we manually examined the 157 chromosomal, protein-coding *Z. mobilis* genes with no significant phenotypes (for a list of these genes, see Data set S7 in the supplemental material). Forty-seven of these genes are adjacent to another gene without a phenotype, a fraction significantly greater than expected by chance ($P = 1.7 \times 10^{-6}$ by Fisher exact test). Of these 47 genes, 19 are involved in either secretion (*ZMO0799* to *ZMO0801*, *ZMO1482*, and *ZMO1483*), antibiotic synthesis (*ZMO1779* and *ZMO1780*), or phage defense (*ZMO0680*, *ZMO0681*, and *ZMO0683* to *ZMO0685*) or encode components of prophage (*ZMO0387* to *ZMO0390* and *ZMO0397* to *ZMO0399*). Given that we did not challenge the *Z. mobilis* mutant libraries with viral infection or microbial competitors, it is not surprising that we did not identify significant phenotypes for these genes.

We considered several other reasons why these 157 genes may lack a phenotype in our data set. One possibility is that we do not have insertions in the central part of each gene and hence are not inactivating its function, or we have only a single transposon mutant and hence have insufficient data. Only 13 of the genes with no phenotype lack insertions in the central part of the gene, and another 13 have a single transposon mutant in our pools, so these explanations cannot explain the majority of the cases. Of these 157 genes with no phenotype, only 9 are ORFans, so incorrect gene calls contribute minimally. A related possibility is that the genes without phenotypes are recent pseudogenes and not functional. To address this issue, we analyzed tiling microarray gene expression data for *Z. mobilis* and compared the expression of genes with and without phenotypes under the expectation that pseudogenes are less likely to be expressed on the correct strand (defined as twofold-greater expression on the sense strand relative to the antisense strand). We find that chromosomal genes without phenotypes are only slightly less likely to be significantly expressed on the correct strand in either rich or minimal medium (86% [135 of 157]) than protein-coding genes on the chromosome with a phenotype (95% [1,292 of 1,362]). An additional explanation is genetic or functional redundancy at the gene or pathway level, whereby a single gene mutation would be expected to have no fitness consequence. One outcome of this hypothesis is that the percentage of duplicated genes (paralogs) among these 157 genes without phenotypes should be larger than for the genes with significant phenotypes. While the number of paralogs (21 of 157) is enriched among the no-phenotype class ($P = 0.0053$ by Fisher exact test), it accounts for few of the genes without phenotypes. Therefore, it is unlikely that the absence of phenotypes for these 157 protein-coding genes is solely due to insufficient data, pseudogenes, lack of expression, or functional redundancy. Alternatively, we did not profile conditions that would lead to a detectable phenotype for these genes. Last, the phenotypes for these genes may not be detectable by our competitive growth assay, which were typically run for 6 generations. For example, a gene fitness value of −0.1 (which would not match our significance criteria), corresponds to a selection coefficient (s) of $0.01 \cong 0.1 \times \ln(2)/6$. Given the effective population size of bacteria, an s of 0.01 may correspond to very strong selection in the wild.

Forty-one percent of *Z. mobilis* genes have biologically informative patterns of fitness. Previous work has established the utility of genome-wide mutant fitness data to annotate the functions of poorly characterized genes in bacteria and yeast using gene-gene associations (4–6, 8). For example, using a large *S. oneidensis* fitness data set, we previously proposed specific functional annotations for 40 genes or operons with poor or incomplete annotations (4). Here, we estimate the fraction of all bacterial genes that are amenable to informative gene-gene associations using high-throughput genetics. To determine the number of genes with biologically meaningful gene-gene associations using mutant fitness, it is important to differentiate between whether a gene has a significant phenotype at all (as discussed above) and whether a gene's pattern of phenotypes is sufficiently strong to be biologically informative to predict function (4). To address the latter, we examined two factors that influence gene function prediction, significant fitness correlations between gene pairs across all experiments (cofitness) and the detection of a strong phenotype in at least one condition. Using stringent criteria for both parameters, cofitness with another gene greater than 0.75 and a strong phenotype ($|\text{fitness}| > 1$) in at least one experimental group, we find that 41% (651 of 1,586) of *Z. mobilis* genes are attractive targets for associative annotation using large-scale mutant fitness profiling (Fig. 3A). In contrast, 21% (691 of 3,355) of *Shewanella oneidensis* MR-1 genes meet the same two criteria across the 296 fitness experiments. The smaller percentage of *S. oneidensis* genes with high cofitness and a strong phenotype relative to *Z. mobilis* may reflect the larger genome size of *S. oneidensis* MR-1, the fact that we performed fewer experiments with this bacterium, that the *S. oneidensis* experiments were done under less-informative conditions, or the fact that we were able to interrogate essential genes in *Z. mobilis*. Of the 651 *Z. mobilis* genes with strong cofitness to another gene and a strong phenotype, 187 are predicted to be essential. Subtracting these essential genes, 464 *Z. mobilis* genes (35% of the nonessential genes that we have data for) are attractive candidates for functional annotation, which is still greater than the 21% we observed for *S. oneidensis*.

To systematically verify the biological significance of our selected cofitness threshold, we examined the capacity of cofitness to group genes into functional categories (as defined by The Institute for Genomic Research [TIGR] [now the J. Craig Venter Institute {JCVI}] subroles [38]). TIGR/JCVI subroles provide a reasonable level of functional specificity, for example the main role "amino acid biosynthesis" is further divided into seven subroles for the aromatic amino acid family, aspartate family, glutamate family, pyruvate family, serine family, histidine family, and other amino acid biosynthesis (38). For this analysis, we focused on cofitness and not strong phenotypes, as the vast majority of genes with high cofitness with another gene also have a strong phenotype ($|\text{fitness}| > 1$) in at least one group of experiments (Fig. 3A). Looking only

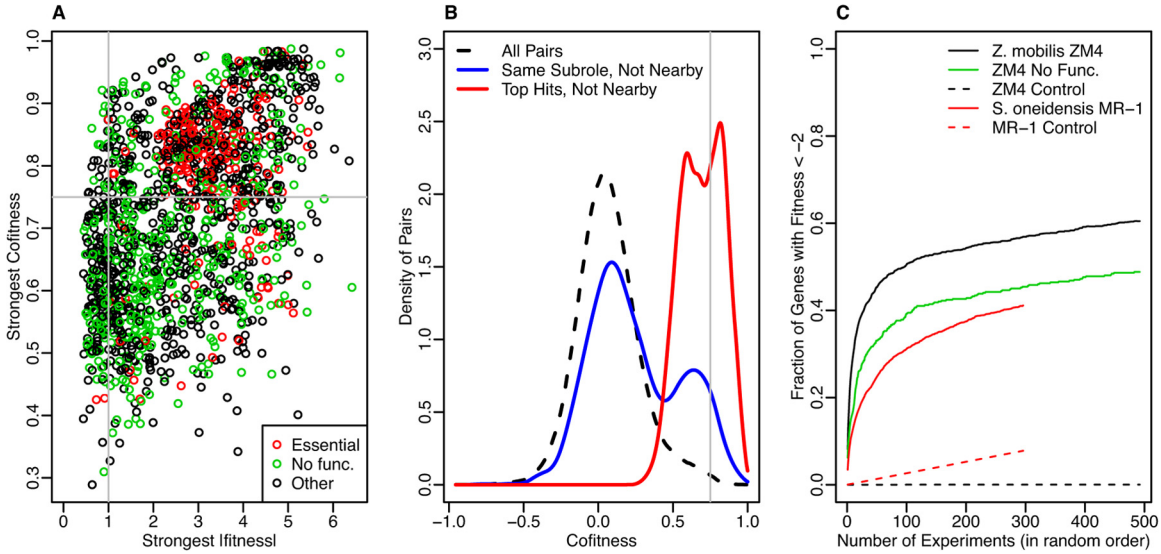

FIG 3 Utility of mutant fitness for annotating gene function in bacteria. (A) For each *Z. mobilis* gene, a scatterplot of the strongest absolute phenotype (*x* axis, either fitness reduced or enhanced) versus the strongest cofitness to another gene (*y* axis). Genes shown in red are putatively essential, and those shown in green are poorly annotated and do not have a specific annotation (no function) (see main text). The horizontal gray line marks cofitness of 0.75, and the vertical gray line marks absolute fitness of 1.0. (B) Distribution of fitness correlations (cofitness) for different classes of *Z. mobilis* gene pairs across all 492 experiments. All pairs of genes that we have data for (All Pairs), gene pairs that have the same TIGR/JCVI subrole (38) and are not within 20 kbp of each other on the chromosome (Same Subrole, Not Nearby), and genes with maximum cofitness for each gene excluding nearby hits within 20 kbp (Top Hits, Not Nearby) are shown. The distributions were estimated from the discrete data using kernel density. The vertical gray line marks cofitness of 0.75. (C) Increase in the fraction of genes with a strong reduced-fitness phenotype (fitness less than −2 [*y* axis]) in any experiment as a function of the number of mutant fitness experiments performed (*x* axis), plotted for all *Z. mobilis* genes for which we have data ($n = 1{,}586$), poorly annotated *Z. mobilis* genes ($n = 502$ [see text for criteria]), or all *S. oneidensis* MR-1 genes with fitness data ($n = 3{,}355$). Experiments are in random order. The red control (dashed) line is derived from the number of fitness values less than −2 among 17 control experiments (independent samples of start) for *S. oneidensis* MR-1. To calculate the number of *Z. mobilis* ZM4 genes expected to have fitness less than −2 by chance, we used the observed standard deviation in 17 control experiments (independent samples of start; this standard deviation was 0.40) and the theoretical probability of a normal distribution with this standard deviation and a mean of 0 giving a value below −2 (2.8×10^{-7} per gene per experiment).

at pairs of *Z. mobilis* genes not nearby in the genome (to avoid operon bias), we find that genes with the same TIGR/JCVI subrole are more likely to have significant cofitness (above 0.75) versus other gene pairs (Fig. 3B, 11% versus 2%, $P < 10^{-15}$ by Fisher exact test). This suggests that cofitness above 0.75 is a strong indicator of functional relatedness and that high cofitness may be a useful tool for inferring the function of poorly characterized genes, particularly if one or more genes with high cofitness have an informative annotation (4–6). When we look at all 1,586 *Z. mobilis* genes for which we have fitness data, we find that 39% (623 genes) have high cofitness (over 0.75) with another, nonnearby gene (Fig. 3B). These results suggest that a substantial number of bacterial genes are amenable to a cofitness-based function prediction. It is important to note that these cofitness-based gene annotations are broad (i.e., a pair of genes have shared phenotypes under a set of conditions) but nevertheless are an advance over the existing, purely computational annotations for these genes (see below for details). Furthermore, these broad annotations can lead to specific hypotheses and proposals for the biochemical and physiological roles of genes, as described below.

Annotation of poorly characterized *Z. mobilis* genes using cofitness. A key challenge in microbiology is the functional annotation of poorly annotated and hypothetical genes. To objectively identify poorly annotated genes, we made a list of *Z. mobilis* proteins that have no gene name and whose description matches "hypothetical," "family," "domain protein," "fold protein," or "related protein." This analysis identified 652 proteins without specific annotations, and we have fitness data for 502 (77%) of these 652 proteins. Of these 502 poorly annotated *Z. mobilis* genes, 35% (174 of 502) have high cofitness with another gene (cofitness > 0.75). Among these poorly annotated genes with strong cofitness to another gene, 79% (137 of 174) have cofitness above 0.75 with a well-annotated gene, demonstrating that mutant fitness-enabled gene-gene associations can be obtained for a significant number of genes with the poorest computational annotations. For a complete list of these 174 genes and the genes that they have high cofitness with, see Data set S8 in the supplemental material.

While determining the precise molecular function and biochemical activity of these poorly annotated proteins requires additional experimentation, gene-gene associations from cofitness can be used to generate more-specific annotations, including correcting misannotated genes, identifying additional evidence to support the broad cofitness-based annotation, and proposing specific physiological roles. To illustrate these points, we manually examined the gene-gene associations in Data set S8 in the supplemental material and found additional evidence, based on conserved proximity or functionally related domains, to support the functional relatedness for 57 of the poorly annotated genes and their genes with high cofitness (for details, see Text S1 in the supplemental material).

Of the 57 newly annotated genes, we proposed specific molecular functions for 33 of them (Table 1). Two of the specific annotations are for genes that were, in hindsight, annotated erroneously (*ZMO1997* and *ZMO1510*). More often, we obtained a specific prediction by comparing the gene's phenotypes with the domain content of the gene or of surrounding genes. For instance,

TABLE 1 Summary of new *Z. mobilis* gene annotations

Category and gene(s)	Brief annotation(s)[a]
Transcriptional regulators	
ZMO0100	Activates ZMO0101
ZMO0116	Regulates response to oxidative stress
ZMO0478	TF/RR with HK ZMO0480; affects the cell wall
ZMO1206	Regulates secretion-related protein ZliE (ZMO0934)
ZMO1322	TF/RR with HK ZMO1323 involved in acid stress resistance
ZMO1336	TF; activates ZMO1337
ZMO1733	OxyR (as in *Caulobacter*; see reference 39)
ZMO1738	TF/RR with HK ZMO1739; regulates essential processes
Transporters and pumps	
ZMO0285	Efflux pump component; substrate unclear
ZMO0780, *ZMO0779*	Efflux pump with ZMO0778; substrate unclear
ZMO0910	Component of polysaccharide export ABC transporter (with ZMO0911 and ZMO1467)
ZMO0964	Efflux pump component
ZMO0981	Component of ABC transporter, likely involved in the export of cell wall components
ZMO1018, *ZMO1017*, *ZMO1016*, *ZMO1015*	ABC transporter exporting component of cell envelope
ZMO1431, *ZMO1430*	Efflux pump, possibly for aromatic acids
ZMO1529, *ZMO1525*	Efflux pump components; substrate unclear
ZMO1591, *ZMO1590*	Efflux pump for aromatic compounds
ZMO1628, *ZMO1630*	Siderophore system acts as efflux system for catechol/protocatechualdehyde
Annotation correction	
ZMO1510	Misannotated as HemK family protein; actually a methyltransferase-modifying release factor
ZMO1997	Novel form of *hemJ* (as in *Acinetobacter*; see reference 47)
Other specific functions	
ZMO0112	Putative substrate of glutamine cyclotransferase
ZMO0803, *ZMO1892*	Regulate peptidoglycan recycling and attachment to outer membrane
ZMO1808	RnfH
ZMO1916	BioH (computational prediction, supported by fitness data)
Pathway-level prediction	
ZMO0055	Permease related to sulfate assimilation
ZMO0107	NDP-sugar transferase related to glycolipid synthesis
ZMO0132, *ZMO0133*	Outer membrane-associated acid tolerance proteins
ZMO0331	Peptidase related to the outer membrane
ZMO0444, *ZMO0445*, *ZMO0447*	Lipid-related enzymes affecting the cell envelope
ZMO0495	Outer membrane biogenesis protein
ZMO0767, *ZMO1319*	Outer membrane-related proteins
ZMO0934	ZliE secretion-related protein
ZMO0947, *ZMO0502*	Synthesis and export of a cell wall component
ZMO1317	Nucleotide kinase-like enzyme affecting the cell wall
ZMO1337	Hydroquinone resistance protein
ZMO1530	Capsular polysaccharide synthesis protein
ZMO1573	Peroxidase regulated by ZMO0116
ZMO1717, *ZMO1718*	Part of an outer membrane integrity system
ZMO1723	Laccase involved in oxidative stress resistance
ZMO1734	UDP glycosyltransferase-like enzyme in cell wall synthesis
ZMO1790	Heme-related transporter
ZMO1875	FeS cluster repair with *bolA* (previously published in reference 17)

[a] Abbreviations: TF, transcription factor; RR, response regulator; HK, histidine kinase; NDP, nucleotide diphosphate.

eight of the newly annotated genes are putative transcriptional regulators that have cofitness with specific genes, so we propose that they activate the expression of those genes. As a specific example, *ZMO1733* belongs to the LysR family of regulators and has cofitness with several genes that are important for resisting oxidative stresses, including an adjacent alkyl hydroperoxide reductase (*ZMO1732*), which suggests that *ZMO1733* is involved in responding to oxidative stresses. Indeed, the ortholog of *ZMO1733* in *Caulobacter crescentus* was recently shown to be the redox-sensitive regulator OxyR (39). Furthermore, three of these transcription factors contain response regulator domains, and all three of these have high cofitness with nearby histidine kinases, which presumably regulate the activity of these transcription factors.

Another 18 of the genes are components of putative ABC transporters or efflux pumps. For example, *ZMO0981* lies within a putative ABC transporter operon (*ZMO02008-ZMO0982-*

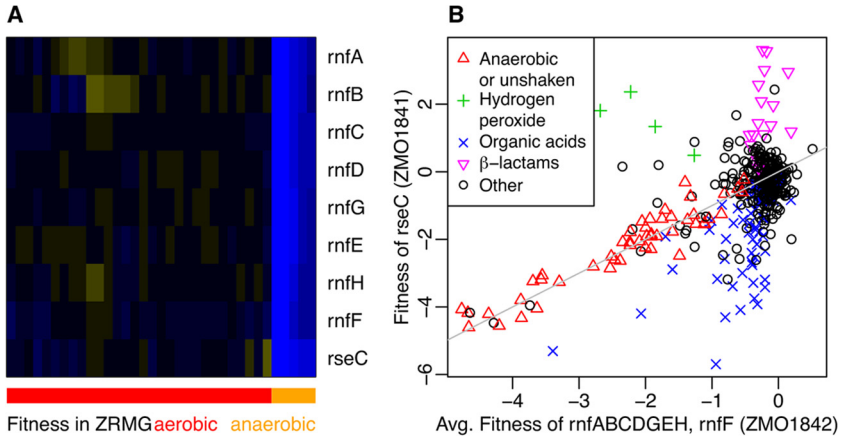

FIG 4 Function of Rnf/RseC in *Z. mobilis*. (A) Heat map of gene fitness values in rich medium in experiments for mutants in components of the Rnf complex and RseC. The experiments marked in red (*x* axis) were performed under aerobic conditions, and those marked in orange were performed under anaerobic conditions. Fitness values are color-coded as described in the legend to Fig. 1A. (B) Comparison of gene fitness values for the Rnf complex (averaged across all eight genes encoding components of the complex) versus RseC in different categories of experiments.

ZMO0981), and close homologs of this operon are sometimes annotated as dipeptide or oligopeptide transporters. However, this system was important for growth in defined medium with no added peptides, which seems inconsistent with that annotation. This operon was detrimental to fitness in the presence of beta-lactam antibiotics, and some homologous operons include putative cell wall remodeling genes or beta-lactamases, so we propose that this operon is involved in the export of a component of the cell envelope. Although the ZMO0981 protein contains recognizable ABC-like ATPase domains, we also identified phenotypes for uncharacterized proteins that do not contain recognized transporter domains but lie within a conserved operon with putative transporter components and have cofitness with them. For example, the hypothetical gene *ZMO1630* does not contain any recognizable domains (it has no InterPro hits) and has cofitness with *ZMO1628* ($r = 0.81$) and with other genes in the operon (*ZMO1631-ZMO1628*). The ZMO1631 protein is annotated as a TonB-like siderophore receptor protein, and ZMO1630 has a signal peptide and three transmembrane helices (as predicted by TMHMM) and could be a component of a transporter. In *Z. mobilis*, *ZMO1631-ZMO1628* mutants are sensitive to catechol or protocatechualdehyde, which are similar compounds (both have benzene rings with two adjacent hydroxyl groups) and are siderophores. Our prediction is that the *Z. mobilis* system naturally promotes the uptake of a ferric siderophore, while in our experiments, the proteins encoded by *ZMO1631-ZMO1628* act as an efflux pump for catechol and protocatechualdehyde. By similar logic, we predict that *ZMO1015* (which contains a Pfam domain of unknown function) (DUF330), *ZMO1591* (DUF140), and *ZMO1431* (DUF1656) encode components of transporters.

We also predicted specific functions for five other proteins. For example, we used cofitness to identify the RnfH (ZMO1808; annotated as hypothetical) and RnfF (ZMO1842; misannotated as *nosX*) components of the ion-pumping electron transport complex Rnf (40, 41). We found that this complex is required for optimal growth of *Z. mobilis* under anaerobic conditions (Fig. 4A). To our knowledge, this is the first demonstration that RnfH, which is not always present in bacterial genomes with the Rnf complex (42), is required for its activity. Furthermore, we show that *rseC* (*ZMO1841*), which is cotranscribed with *rnfF*, is functionally associated with the Rnf complex during anaerobic growth but not during organic acid or beta-lactam antibiotic stress (Fig. 4B). As another example, the hypothetical gene *ZMO1916* has cofitness with biotin synthase (*ZMO0094*; $r = 0.95$) and dethiobiotin synthase (*ZMO0095*; $r = 0.8$), which suggests that *ZMO1916* has a role in biotin synthesis. Indeed, an ortholog of *ZMO1916* in cyanobacteria was annotated as *bioK* (43) and is proposed to be a pimeloyl-acyl carrier protein methyl ester esterase. In *E. coli*, this activity is performed by BioH, but *bioH* is not present in cyanobacteria, and it is absent from *Z. mobilis* as well. As far as we know, this is the first experimental support for the involvement of these genes in biotin synthesis. In many of the cases discussed above, where we have a phenotype for an uncharacterized protein family that lies in a conserved operon, another interpretation might be that the novel protein has an unrelated function and that the phenotypes are due to polar effects. We cannot rule out this possibility, but given that these are conserved operons and that we found a moderate rate of polar effects, we think this is unlikely. Overall, we were able to make or improve specific annotations for 33 hypothetical proteins and make pathway-level predictions for 24 others (Table 1).

Seventy-nine diverse mutant fitness experiments are nearly as informative as 492 experiments. Given that technologies are rapidly advancing to the point that large-scale mutant phenotype data sets in bacteria will proliferate (12–14), we asked whether hundreds of laboratory experiments with a single bacterial species are worth the investment if the goal is to globally annotate gene function (and not to detect statistically significant but subtle phenotypes). To investigate this, we looked at the rate at which new genes with strongly reduced fitness phenotypes (fitness less than −2) appear as a function of increasing the number of experiments (selected at random) for both *Z. mobilis* and *S. oneidensis*. For this analysis, counting genes with strong phenotypes is the simplest way to show the impact of adding more experiments, because it avoids complicated issues around experiment grouping or statistical significance. We find that while each additional experiment provides an increase in the number of genes with a strong phenotype below −2, there is diminishing return after ~100 experi-

ments in both bacteria (Fig. 3C). In *Z. mobilis*, an increase from the average set of 100 random experiments to all 492 experiments only moderately increases the number of genes with a strong reduced-fitness phenotype (from 801 to 959). Similarly, 296 *S. oneidensis* experiments identify 1,379 genes with fitness below −2, compared to 1,046 genes from the average of 100 random experiments. Among the 502 genes without a specific function annotated in *Z. mobilis*, a similar trend of diminishing return is observed around 100 random experiments; moving from 100 to 492 experiments only moderately increases the number of these poorly annotated genes with fitness less than −2 from 197 to 245 (Fig. 3C).

Finally, we examined whether a rational approach for selecting the conditions would enable the same level of biological discovery while reducing cost and effort. With one experiment from each of the 79 nonredundant groups of *Z. mobilis* fitness experiments (for a list of conditions, see Data set S1 in the supplemental material), 146 of the 174 (84%) poorly characterized genes have high cofitness with one of the original genes (cofitness > 0.78). We used a higher cofitness threshold for this analysis (0.78 versus 0.75) to keep the fraction of random gene pairs with cofitness above 0.75 fixed at 0.37%, despite having fewer experiments (79 versus 492). Therefore, 79 diverse, laboratory-based mutant fitness experiments (rather than ~500) are sufficient for identifying most cofitness-based gene-gene associations in *Z. mobilis*. If we pick a random exemplar of each of the 61 nonredundant groups of *S. oneidensis* MR-1 fitness experiments (Data set S2), 1,130 genes, including 355 poorly characterized genes, have a reasonably strong phenotype (|fitness| > 0.75) and significant cofitness above 0.8. Of the 65 genes we previously annotated using mutant fitness in *S. oneidensis* (4), 46 (71%) are above these thresholds in the reduced data set.

DISCUSSION

Phenotypes for almost all genes in bacteria. To our knowledge, our finding that 89% of assayed *Z. mobilis* genes have a detectable phenotype is the highest fraction for a bacterium thus far. In *E. coli*, despite decades of extensive single-gene and genome-wide studies, a significant fraction of the genome does not have an identified phenotype (and hence function) (44). For instance, a genome-wide analysis of mutant fitness of single-gene knockout strains of *E. coli* across hundreds of conditions identified a significant phenotype for only half of the genome (8). Furthermore, deletion of approximately 10% of the genes within a single strain of *E. coli*, primarily targeting hypothetical and selfish genes, did not substantially impact the growth rate in a defined medium (45). There are a number of potential reasons why we detected such a high percentage of genes with a phenotype in *Z. mobilis*, including the sensitivity of our competitive fitness assay (11), the small genome size of *Z. mobilis*, and our grouping of similar experiments to increase statistical power for detecting subtle phenotypes. Furthermore, we confirmed that these subtle phenotypes are genuine by showing that detrimental mutations continued to decrease in abundance when we continued the experiment for more generations (Fig. 1E), by showing that virtually all of the genes with phenotypes in rich medium are expressed in rich medium (Fig. 1D), and by showing that even operons that have only subtle phenotypes tend to have high cofitness (Fig. 1C).

Utility of more phenotypes. Given that 79 fitness experiments suffice to find phenotypes for most genes in *Z. mobilis*, it is not surprising that doing hundreds of additional assays failed to find many additional significant phenotypes. However, measuring fitness in additional conditions did not make the phenotypes more interpretable, which surprised us. Intuitively, more fitness experiments allow many genes to show more-complex fitness patterns that contain more information about gene function, but we were not able to take advantage of this. We believe that this is partly because most of these additional experiments were stresses by small molecules, which are difficult to interpret, so we relied on cofitness. Also, genes with significant but only weak phenotypes tended to have lower cofitness (Fig. 3A), which limited our ability to predict the functions of these genes. It is possible that better statistical methods or complementary data of other types (e.g., protein-protein interactions or double mutants) would increase the utility of the additional conditions. A related issue is that doing more-similar experiments (e.g., different concentrations of an inhibitor) allows for increased confidence in subtle phenotypes (reducing fitness by just 3% per generation), but we do not see how to use subtle phenotypes for annotation. In an organism with a broader range of metabolic or respiratory capabilities, such as *S. oneidensis*, many genes have specific phenotypes relating to metabolism or respiration that are readily interpretable (4). For example, doing an additional fitness experiment with a new carbon source might yield a specific phenotype for one or two operons involved in the transport or catabolism of that compound. In terms of the genome-wide numbers of genes with phenotypes, this is not impressive, but it does lead to specific annotations.

Implications for annotating gene function in bacteria. Given the ease of bacterial genome sequencing, it is imperative that high-throughput approaches for elucidating gene function are developed to determine gene function in a wide range of bacteria. In addition to demonstrating that nearly all genes in bacteria have an identifiable phenotype, our results and methods suggest that scaling mutant fitness-based gene annotation to many bacteria is feasible. First, the majority of our *Z. mobilis* experiments (316 of 492) were performed in microplates, demonstrating that bacterial mutant fitness assays can be performed in a miniaturized, high-throughput growth format. Second, switching from microarrays to sequencing DNA bar codes will enable greater throughput and lower cost (13, 46).

Our genome-wide fitness results and analysis do not mean that we have completely validated the function(s) of the genes discussed. Rather, our work flow provides a data-driven, high-throughput approach to generate many gene function predictions of different specificities using gene-gene fitness correlations. In fact, we view our genome-wide data sets as a starting point for generating specific hypotheses on the functions of poorly characterized genes, which could be followed up with more-traditional, single-gene investigations. However, given the sheer number of uncharacterized proteins, this will be possible only for the most interesting genes. We hope that other high-throughput approaches will provide complementary information so that we can make reliable claims about the functions of most of the other uncharacterized proteins.

Although this study focused on poorly annotated genes, fitness data could also be used to test the more-specific functional annotations. We noted several erroneous annotations during the analysis of hypothetical proteins (e.g., ZMO1842 was misannotated as *nosX* instead of as *rnfF*, and a *soxR*-like regulator was omitted from the annotation, see Text S1 in the supplemental material). To

illustrate this issue more broadly, we considered the protein-coding genes of *Z. mobilis* that have specific annotations and significant beneficial phenotypes and are strongly beneficial in at least one group of experiments (average fitness under −1). There are 491 such proteins, and we examined a random subset of 20 of them (Text S2). We confirmed the annotations for nine of these genes and found two erroneous annotations. For the remaining nine genes, we could not make a clear determination; this included four genes that were important for fitness in most conditions, which confirms the gene's importance but does not link it to a biological process. If scaled up to the 491 candidate proteins, this approach could probably be used to confirm hundreds of annotations and identify dozens of erroneous annotations.

Last, our results show that less than 100 experiments, instead of 492, suffice to find phenotypes and informative cofitness for many genes. Although these experiments were selected in hindsight, we expect that most of the redundancy of the fitness experiments could be avoided in future studies with other organisms. Most of the experiments grouped into clusters that comprised replicate experiments, near-replicate experiments, such as different concentrations of the same inhibitor, or experiments that involved structurally similar compounds, such as antibiotics of the same class.

What fitness experiments should be conducted for another bacterium? We recommend selecting conditions based on the organism's energetic or metabolic capabilities, i.e., different sources of carbon and nitrogen, or combinations of electron donors and acceptors. We recommend that a few dozen dissimilar stresses be performed as well; our clustering should help to select these conditions (see Data sets S1 and S2 in the supplemental material). In conclusion, this work provides a general approach to discover the functions of many genes in diverse bacteria by using mutant fitness.

ACKNOWLEDGMENTS

This work was initially funded by the Energy Biosciences Institute grant OO7G02 and completed with funding from ENIGMA. The work conducted by ENIGMA was supported by the Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under contract DE-AC02-05CH11231.

REFERENCES

1. **Galperin MY, Koonin EV.** 2010. From complete genome sequence to 'complete' understanding? Trends Biotechnol. **28:**398–406. http://dx.doi.org/10.1016/j.tibtech.2010.05.006.
2. **Raskin DM, Seshadri R, Pukatzki SU, Mekalanos JJ.** 2006. Bacterial genomics and pathogen evolution. Cell **124:**703–714. http://dx.doi.org/10.1016/j.cell.2006.02.002.
3. **Roberts RJ, Chang YC, Hu Z, Rachlin JN, Anton BP, Pokrzywa RM, Choi HP, Faller LL, Guleria J, Housman G, Klitgord N, Mazumdar V, McGettrick MG, Osmani L, Swaminathan R, Tao KR, Letovsky S, Vitkup D, Segre D, Salzberg SL, Delisi C, Steffen M, Kasif S.** 2011. COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. Nucleic Acids Res. **39:**D11–D14. http://dx.doi.org/10.1093/nar/gkq1168.
4. **Deutschbauer A, Price MN, Wetmore KM, Shao W, Baumohl JK, Xu Z, Nguyen M, Tamse R, Davis RW, Arkin AP.** 2011. Evidence-based annotation of gene function in *Shewanella oneidensis* MR-1 using genome-wide fitness profiling across 121 conditions. PLoS Genet. **7:**e1002385. http://dx.doi.org/10.1371/journal.pgen.1002385.
5. **Dudley AM, Janse DM, Tanay A, Shamir R, Church GM.** 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. Mol. Syst. Biol. **1:**2005.0001. http://dx.doi.org/10.1038/msb4100004.
6. **Hillenmeyer ME, Ericson E, Davis RW, Nislow C, Koller D, Giaever G.** 2010. Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. Genome Biol. **11:**R30. http://dx.doi.org/10.1186/gb-2010-11-3-r30.
7. **Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St. Onge RP, Tyers M, Koller D, Altman RB, Davis RW, Nislow C, Giaever G.** 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science **320:**362–365. http://dx.doi.org/10.1126/science.1150021.
8. **Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA.** 2011. Phenotypic landscape of a bacterial cell. Cell **144:**143–156. http://dx.doi.org/10.1016/j.cell.2010.11.052.
9. **Oh J, Fung E, Schlecht U, Davis RW, Giaever G, St. Onge RP, Deutschbauer A, Nislow C.** 2010. Gene annotation and drug target discovery in *Candida albicans* with a tagged transposon mutant collection. PLoS Pathog. **6:**e1001140. http://dx.doi.org/10.1371/journal.ppat.1001140.
10. **Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, et al.** 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature **418:**387–391. http://dx.doi.org/10.1038/nature00935.
11. **Oh J, Fung E, Price MN, Dehal PS, Davis RW, Giaever G, Nislow C, Arkin AP, Deutschbauer A.** 2010. A universal TagModule collection for parallel genetic analysis of microorganisms. Nucleic Acids Res. **38:**e146. http://dx.doi.org/10.1093/nar/gkq419.
12. **Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK.** 2009. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. Genome Res. **19:**2308–2316. http://dx.doi.org/10.1101/gr.097097.109.
13. **Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, Roth FP, Giaever G, Nislow C.** 2009. Quantitative phenotyping via deep barcode sequencing. Genome Res. **19:**1836–1842. http://dx.doi.org/10.1101/gr.093955.109.
14. **van Opijnen T, Bodi KL, Camilli A.** 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods **6:**767–772. http://dx.doi.org/10.1038/nmeth.1377.
15. **Rogers PL, Jeon YJ, Lee KJ, Lawford HG.** 2007. *Zymomonas mobilis* for fuel ethanol and higher value products. Adv. Biochem. Eng. Biotechnol. **108:**263–288. http://dx.doi.org/10.1007/10_2007_060.
16. **Bochner B, Gomez V, Ziman M, Yang S, Brown SD.** 2010. Phenotype microarray profiling of *Zymomonas mobilis* ZM4. Appl. Biochem. Biotechnol. **161:**116–123. http://dx.doi.org/10.1007/s12010-009-8842-2.
17. **Skerker JM, Leon D, Price MN, Mar JS, Tarjan DR, Wetmore KM, Deutschbauer AM, Baumohl JK, Bauer S, Ibanez AB, Mitchell VD, Wu CH, Hu P, Hazen T, Arkin AP.** 2013. Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates. Mol. Syst. Biol. **9:**674. http://dx.doi.org/10.1038/msb.2013.30.
18. **Pierce SE, Davis RW, Nislow C, Giaever G.** 2007. Genome-wide analysis of barcoded *Saccharomyces cerevisiae* gene-deletion mutants in pooled cultures. Nat. Protoc. **2:**2958–2974. http://dx.doi.org/10.1038/nprot.2007.427.
19. **Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, Dubchak IL, Alm EJ, Arkin AP.** 2010. MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res. **38:**D396–D400. http://dx.doi.org/10.1093/nar/gkp919.
20. **Price MN, Deutschbauer AM, Kuehl JV, Liu H, Witkowska HE, Arkin AP.** 2011. Evidence-based annotation of transcripts and proteins in the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. J. Bacteriol. **193:**5716–5727. http://dx.doi.org/10.1128/JB.05563-11.
21. **Yang S, Pappas KM, Hauser LJ, Land ML, Chen GL, Hurst GB, Pan C, Kouvelis VN, Typas MA, Pelletier DA, Klingeman DM, Chang YJ, Samatova NF, Brown SD.** 2009. Improved genome annotation for *Zy-*

momonas mobilis. Nat. Biotechnol. **27:**893–894. http://dx.doi.org/10.1038/nbt1009-893.
22. **Price MN, Deutschbauer AM, Skerker JM, Wetmore KM, Ruths T, Mar JS, Kuehl JV, Shao W, Arkin AP.** 2013. Indirect and suboptimal control of gene expression is widespread in bacteria. Mol. Syst. Biol. **9:**660. http://dx.doi.org/10.1038/msb.2013.16.
23. **Truglio JJ, Croteau DL, Van Houten B, Kisker C.** 2006. Prokaryotic nucleotide excision repair: the UvrABC system. Chem. Rev. **106:**233–252. http://dx.doi.org/10.1021/cr040471u.
24. **Kuzminov A.** 1999. Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. Microbiol. Mol. Biol. Rev. **63:**751–813.
25. **Zdraveski ZZ, Mello JA, Marinus MG, Essigmann JM.** 2000. Multiple pathways of recombination define cellular responses to cisplatin. Chem. Biol. **7:**39–50.
26. **Cox MM.** 2007. Motoring along with the bacterial RecA protein. Nat. Rev. Mol. Cell Biol. **8:**127–138. http://dx.doi.org/10.1038/nrm2099.
27. **Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Coller JA, Fero MJ, McAdams HH, Shapiro L.** 2011. The essential genome of a bacterium. Mol. Syst. Biol. **7:**528. http://dx.doi.org/10.1038/msb.2011.58.
28. **Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP, Davis RW.** 2004. Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. Proc. Natl. Acad. Sci. U. S. A. **101:**793–798. http://dx.doi.org/10.1073/pnas.0307490100.
29. **de Daruvar A, Collado-Vides J, Valencia A.** 2002. Analysis of the cellular functions of *Escherichia coli* operons and their conservation in Bacillus subtilis. J. Mol. Evol. **55:**211–221. http://dx.doi.org/10.1007/s00239-002-2317-1.
30. **Price MN, Arkin AP, Alm EJ.** 2006. The life-cycle of operons. PLoS Genet. **2:**e96. http://dx.doi.org/10.1371/journal.pgen.0020096.
31. **Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV.** 2002. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res. **30:**2212–2223. http://dx.doi.org/10.1093/nar/30.10.2212.
32. **Baran R, Bowen BP, Price MN, Arkin AP, Deutschbauer AM, Northen TR.** 2013. Metabolic footprinting of mutant libraries to map metabolite utilization to genotype. ACS Chem. Biol. **8:**189–199. http://dx.doi.org/10.1021/cb300477w.
33. **Lynch M, Conery JS.** 2003. The origins of genome complexity. Science **302:**1401–1404. http://dx.doi.org/10.1126/science.1089370.
34. **Koskiniemi S, Sun S, Berg OG, Andersson DI.** 2012. Selection-driven gene loss in bacteria. PLoS Genet. **8:**e1002787. http://dx.doi.org/10.1371/journal.pgen.1002787.
35. **Lee MC, Marx CJ.** 2012. Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genet. **8:**e1002651. http://dx.doi.org/10.1371/journal.pgen.1002651.
36. **Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC, Tavazoie S.** 2013. Bacterial adaptation through loss of function. PLoS Genet. **9:**e1003617. http://dx.doi.org/10.1371/journal.pgen.1003617.
37. **Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA.** 2008. DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. **36:**D88–D92. http://dx.doi.org/10.1093/nar/gkm964.
38. **Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O.** 2001. The Comprehensive Microbial Resource. Nucleic Acids Res. **29:**123–125. http://dx.doi.org/10.1093/nar/29.1.123.
39. **Italiani VC, da Silva Neto JF, Braz VS, Marques MV.** 2011. Regulation of catalase-peroxidase KatG is OxyR dependent and Fur independent in *Caulobacter crescentus*. J. Bacteriol. **193:**1734–1744. http://dx.doi.org/10.1128/JB.01339-10.
40. **Biegel E, Muller V.** 2010. Bacterial Na+-translocating ferredoxin:NAD+ oxidoreductase. Proc. Natl. Acad. Sci. U. S. A. **107:**18138–18142. http://dx.doi.org/10.1073/pnas.1010318107.
41. **Schmehl M, Jahn A, Meyer zu Vilsendorf A, Hennecke S, Masepohl B, Schuppler M, Marxer M, Oelze J, Klipp W.** 1993. Identification of a new class of nitrogen fixation genes in *Rhodobacter capsulatus*: a putative membrane complex involved in electron transport to nitrogenase. Mol. Gen. Genet. **241:**602–615.
42. **Biegel E, Schmidt S, Gonzalez JM, Muller V.** 2011. Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell. Mol. Life Sci. **68:**613–634. http://dx.doi.org/10.1007/s00018-010-0555-8.
43. **Rodionov DA, Mironov AA, Gelfand MS.** 2002. Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. Genome Res. **12:**1507–1516. http://dx.doi.org/10.1101/gr.314502.
44. **Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muniz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP, Karp PD.** 2011. EcoCyc: a comprehensive database of *Escherichia coli* biology. Nucleic Acids Res. **39:**D583–D590. http://dx.doi.org/10.1093/nar/gkq1143.
45. **Kolisnychenko V, Plunkett G, III, Herring CD, Feher T, Posfai J, Blattner FR, Posfai G.** 2002. Engineering a reduced *Escherichia coli* genome. Genome Res. **12:**640–647. http://dx.doi.org/10.1101/gr.217202.
46. **Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, Nislow C.** 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res. **38:**e142. http://dx.doi.org/10.1093/nar/gkq368.
47. **Boynton TO, Gerdes S, Craven SH, Neidle EL, Phillips JD, Dailey HA.** 2011. Discovery of a gene involved in a third bacterial protoporphyrinogen oxidase activity through comparative genomic analysis and functional complementation. Appl. Environ. Microbiol. **77:**4795–4801. http://dx.doi.org/10.1128/AEM.00171-11.