PLOS GENETICS

RESEARCH ARTICLE

Filling gaps in bacterial amino acid biosynthesis pathways with high-throughput genetics

Morgan N. Price[1]*, Grant M. Zane[2], Jennifer V. Kuehl[1], Ryan A. Melnyk[1], Judy D. Wall[2], Adam M. Deutschbauer[1]*, Adam P. Arkin[1]*

1 Environmental Genomics & Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **2** Department of Biochemistry, University of Missouri, Columbia, Missouri, United States of America

* morgannprice@yahoo.com (MNP); amdeutschbauer@lbl.gov (AMD); aparkin@lbl.gov (APA)

OPEN ACCESS

Citation: Price MN, Zane GM, Kuehl JV, Melnyk RA, Wall JD, Deutschbauer AM, et al. (2018) Filling gaps in bacterial amino acid biosynthesis pathways with high-throughput genetics. PLoS Genet 14(1): e1007147. https://doi.org/10.1371/journal.pgen.1007147

Editor: Josep Casadesús, Universidad de Sevilla, SPAIN

Received: October 2, 2017

Accepted: December 10, 2017

Published: January 11, 2018

Data Availability Statement: Mutant fitness data and Sanger sequencing data are available for download at http://genomics.lbl.gov/supplemental/auxo. The fitness data can also be viewed in the Fitness Browser (http://fit.genomics.lbl.gov/) or downloaded from figshare (https://doi.org/10.6084/m9.figshare.5134840.v1). The predicted essential genes are available from http://genomics.lbl.gov/supplemental/bigfit/ or from the R image in https://doi.org/10.6084/m9.figshare.5134837.v1. The Sanger sequencing data has been submitted to Genbank (accession KY549926).

Abstract

For many bacteria with sequenced genomes, we do not understand how they synthesize some amino acids. This makes it challenging to reconstruct their metabolism, and has led to speculation that bacteria might be cross-feeding amino acids. We studied heterotrophic bacteria from 10 different genera that grow without added amino acids even though an automated tool predicts that the bacteria have gaps in their amino acid synthesis pathways. Across these bacteria, there were 11 gaps in their amino acid biosynthesis pathways that we could not fill using current knowledge. Using genome-wide mutant fitness data, we identified novel enzymes that fill 9 of the 11 gaps and hence explain the biosynthesis of methionine, threonine, serine, or histidine by bacteria from six genera. We also found that the sulfate-reducing bacterium *Desulfovibrio vulgaris* synthesizes homocysteine (which is a precursor to methionine) by using DUF39, NIL/ferredoxin, and COG2122 proteins, and that homoserine is not an intermediate in this pathway. Our results suggest that most free-living bacteria can likely make all 20 amino acids and illustrate how high-throughput genetics can uncover previously-unknown amino acid biosynthesis genes.

Author summary

For a few bacteria, it is well known how they can make all 20 of the standard amino acids (the building blocks of proteins). For many other bacteria, their genome sequence implies that there are gaps in these biosynthetic pathways, so that the bacteria cannot make all of the amino acids and would need to take up some of them from their environment instead. But many bacteria can grow in minimal media (without any amino acids) despite these apparent gaps. We studied 10 bacteria with predicted gaps in amino acid biosynthesis that nevertheless grow in minimal media. Most of these gaps were spurious, but 11 of the gaps were genuine and could not be explained by current knowledge. Using high-throughput genetics, we systematically identified genes that were required for growth in minimal media and identified the biosynthetic genes that fill 9 of the 11 gaps. We hope that this

Funding: This material by ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231. JDW, AMD, and APA received the funding. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

approach can be applied to many more bacteria and will eventually allow us to accurately predict the nutritional requirements of a bacterium from its genome sequence.

Introduction

Although it has been known for decades how the model bacterium *Escherichia coli* synthesizes all 20 of the standard amino acids, novel pathways for amino acid biosynthesis continue to be discovered in other bacteria and in archaea. For example, unlike *E. coli*, most bacteria use tRNA-dependent pathways for the biosynthesis of some of these amino acids [1]. More recent discoveries include a novel pathway for methionine biosynthesis in methanogenic archaea and sulfate-reducing bacteria [2–4] and a novel route by which *Pelagibacter ubique*, which is perhaps the most abundant bacterium on earth, synthesizes glycine from a waste product of photosynthetic organisms [5].

Our incomplete knowledge of these pathways makes it difficult to accurately predict a bacterium's minimal growth requirements from its genome sequence. Besides saving experimental effort, such predictions would make it possible to understand the ecological role of bacteria that are difficult to cultivate. For example, Mee and colleagues [6] and D'Souza and colleagues [7] both used comparative genomics to predict that most bacteria cannot synthesize all 20 of the standard amino acids. Both groups suggested that most free-living bacteria are reliant on cross-feeding of amino acids that are released by other bacteria. However, neither group tested this experimentally by measuring bacterial growth requirements. In contrast, we recently studied 24 heterotrophic bacteria from 15 different genera [8] and found that all but one of these bacteria grew in defined minimal media with no amino acids present. If we used the automated method that Mee and colleagues relied on [9] and excluded *Escherichia coli*, then we had predictions for bacteria from 12 genera that grow in defined media. All of these bacteria were incorrectly predicted to be auxotrophic for multiple amino acids. In other words, there were many putative gaps in the biosynthetic pathways—enzymatic steps that are required for biosynthesis and were missing from the genome annotation—but these gaps were misleading.

To understand why, we manually examined the predictions for bacteria from 10 different genera. Although the automated method identified a total of 173 gaps, we argue that just 11 of these represent genuine gaps in our biological knowledge. Another gap relates to a recently-discovered and poorly-understood pathway for the biosynthesis of homocysteine (which is a precursor of methionine) in sulfate-reducing bacteria and archaea [2–4]. To identify the genes that encode these reactions, we used genome-wide mutant fitness assays, in which a pool of ~40,000–500,000 randomly-barcoded transposon mutants are grown together and DNA sequencing is used to measure how each mutant's abundance changes during growth (RB-TnSeq) [8,10]. Given the genetic data, we looked for genes that were important for fitness in minimal media, but not in rich media, and whose mutants were rescued by the addition of a specific amino acid. Of the 11 genuine gaps, we identified genes to fill nine of them; these novel enzymes explain the biosynthesis of four different amino acids by bacteria from six different genera. And we found that homocysteine synthesis in *Desulfovibrio vulgaris* Miyazaki F required DUF39, NIL/ferredoxin, and COG2122 proteins, as expected from studies of this pathway in other organisms [2,3,11]. Our genetic data imply that, in contrast to all other known pathways, homoserine is not an intermediate in homocysteine synthesis in *D. vulgaris*.

Results

A high rate of error in automated predictions of auxotrophy

Mee and colleagues [6] relied on predicted phenotypes from the Integrated Microbial Genomes web site (https://img.jgi.doe.gov/; [9]), while D'Souza and colleagues [7] made their own predictions. We focus our analysis on the IMG predictions because they are publicly available, but the predictions by D'Souza and colleagues have similar issues (see S1 Text). Among 24 heterotrophic bacteria that grow in defined media without added amino acids and for which we have mutant fitness data, we selected one representative of each genus whose genome is present in IMG. We also excluded the traditional model bacterium *E. coli*. This left us with 12 bacteria, and for each bacterium, IMG incorrectly predicted auxotrophy for at least two amino acids.

We examined 10 of these bacteria in more detail. On average, IMG predicts that these bacteria are prototrophic for 9.6 amino acids, auxotrophic for 6.2 amino acids, and makes no prediction either way for 4.2 amino acids (Fig 1). To verify that these bacteria grow in the absence of any externally provided amino acids, we performed multi-transfer growth experiments for seven of them (S2 Text). We also tested the vitamin requirements of these seven bacteria (S2 Text). Six of the bacteria (*Burkholderia phytofirmans* PsJN, *Desulfovibrio vulgaris* Miyazaki F, *Herbaspirillum seropedicae* SmR1, *Marinobacter adhaerens* HP15, *Phaeobacter inhibens* BS107, and *Pseudomonas stutzeri* RCH2) did not require the addition of any amino acids or vitamins for growth. *Sinorhizobium meliloti* 1021 did not grow without added vitamins, which is consistent with a previous report that it requires biotin (but not amino acids) for growth [12].

For each of the 104 cases in which IMG failed to predict that the bacterium could synthesize the amino acid, we used the IMG website to identify the gaps—the reactions that were necessary for biosynthesis of the amino acid but no gene was predicted (in IMG) to encode them. In many cases, a pathway had more than one gap (more than one reaction that is required to make the amino acid was not associated with a gene). Overall, there were a total of 173 gaps in amino acid biosynthesis among the 10 bacteria.

Of 173 putative gaps in amino acid biosynthesis pathways, 11 are gaps in biological knowledge

To identify candidate genes for the gaps, we began with three standard annotation resources: TIGRFam [13], KEGG [14], and SEED/RAST [15]. If a protein was annotated with the putatively missing enzymatic capability by at least two out of three of these resources, then we considered it to be a clear candidate. We found that 140 of the 173 gaps (81%) had clear candidates (see overview in Table 1). Five of the clear candidates are fused to other biosynthetic enzymes, and these fusions might cause the genes to be missed when searching for bidirectional best hits [9]. In some other cases, the presence of a potential ortholog is noted on the IMG website, but the gene's annotation was not deemed high-confidence enough to predict that the pathway is present. These ambiguous cases are labeled as "auxotroph" on the IMG website (as of August 2017) and were so considered in the analyses of Mee and colleagues [6].

To test if these 140 clear candidates were actually involved in amino acid biosynthesis, we examined genome-wide mutant fitness data for each of the 10 bacteria across dozens of growth conditions [8]. In general, a gene that is involved in the biosynthesis of amino acids should be important for fitness in most defined minimal media experiments but not during growth in media that contains yeast extract (which contains all of the standard amino acids) or casamino acids (which contains all of the standard amino acids except tryptophan). For nine of the gaps, more than one clear candidate was identified in the genome and no strong phenotype was

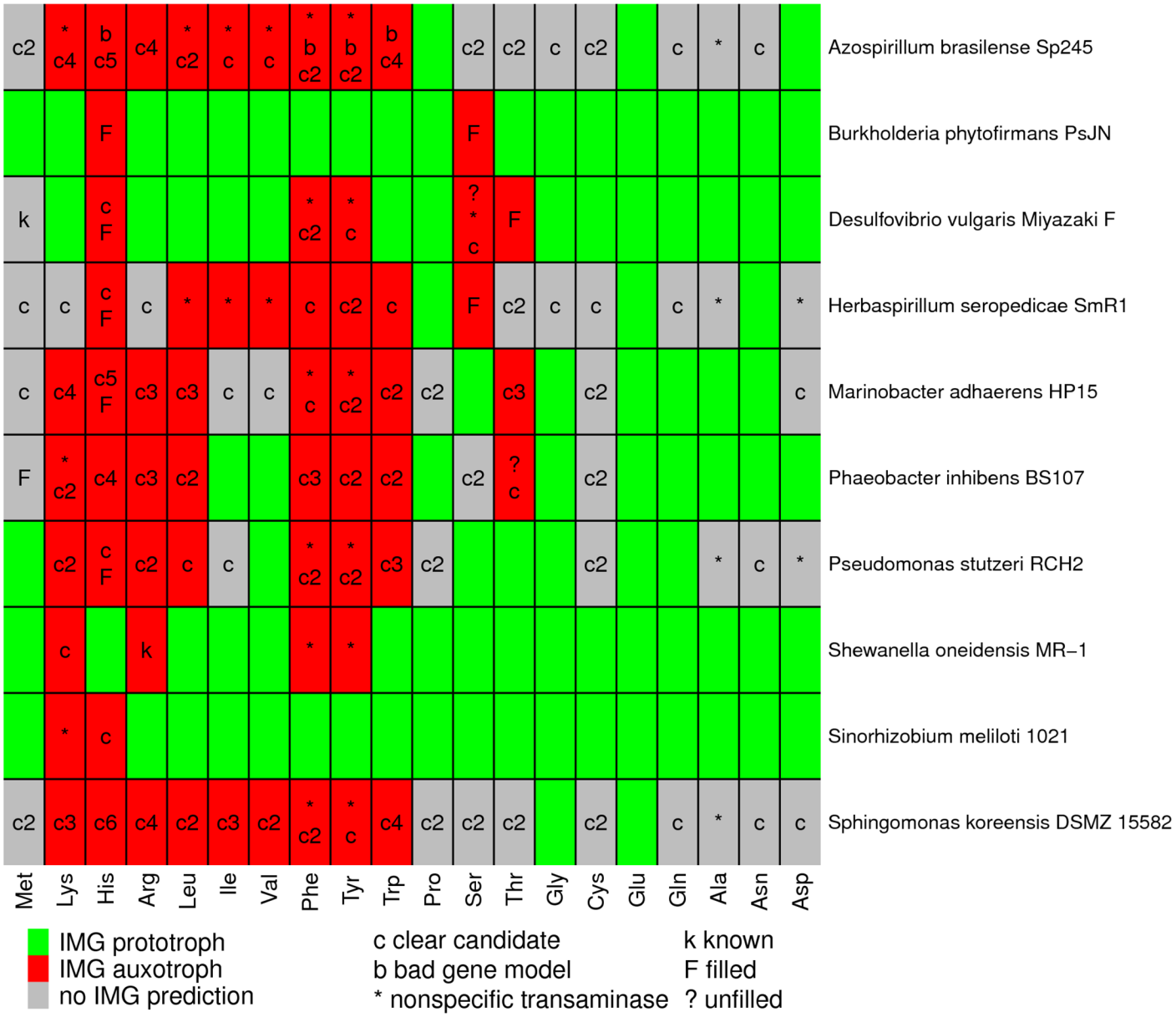

Fig 1. Gaps in amino acid biosynthesis in 10 bacteria. For each amino acid, we identified the missing reactions or gaps in the IMG predictions, and we show a single-letter code with the classification of each gap. A cell may have multiple codes, one for each gap. "Clear candidate" means that at least two annotation resources identified a gene for the gap. If there was more than one gap with a clear candidate, this is shown with a number (i.e., "c2" for two gaps with clear candidates). "Known" means that the step is described in the literature but not in the databases. Some reactions are involved in the biosynthesis of multiple amino acids, so some gaps are shown multiple times.

https://doi.org/10.1371/journal.pgen.1007147.g001

found in the mutant fitness data (S1 Table), which may indicate genetic redundancy. For the remaining 131 steps, we classified 61 genes as auxotrophs because they were important for fitness in most defined media conditions. Some examples are shown in Fig 2. Note that we define gene fitness in a condition as the $\log_2$ change in the abundance of mutants in that gene after growth from an optical density of 0.02 to saturation (usually 4–8 generations) [10]. Genes that are not important for fitness will have fitness values near zero, and fitness values of under -2 indicate a strong defect in growth. As shown in Fig 2, these auxotrophs had strong fitness

Table 1. Classification of the 173 gaps in amino acid biosynthesis pathways in 10 bacteria.

Group	Subgroup	Cases
Clear candidate	multiple candidates	9
	auxotrophic	61
	essential	54
	no fitness data	9
	other (redundant)	5
	other	2
Gene model error		2
Transaminase		18
Known gap		2
Genuine gap	filled (this study)	9
	unfilled	2

https://doi.org/10.1371/journal.pgen.1007147.t001

defects if amino acids are absent and had little phenotype (fitness near zero) if amino acids were added. Furthermore, in some cases, the gene is expected to be required for the synthesis of just one amino acid, and we measured gene fitness with that amino acid as the sole source of carbon or nitrogen. In these cases, the amino acid rescues the auxotroph, as expected. For example, the bottom right panel in Fig 2 shows that mutants in *hisF* from *Sphingomonas koreensis* (Ga0059261_1048) were rescued when L-histidine was present. Overall, the fitness data shows that these 61 gaps were filled correctly.

Another 54 of the clear candidates were identified as putatively essential for growth in rich media with added amino acids because few transposon mutants in those genes were recovered [8]. Amino acid biosynthesis genes may be essential in rich media because the bacterium cannot take up the amino acid or because the biosynthetic pathway overlaps with another essential process. For example, in *E. coli*, *dapABDEF* are required for both lysine synthesis and peptidoglycan synthesis and are essential for growth in standard rich media such as LB [8,16]. In the 10 bacteria under consideration, the clear candidates are much more likely than other proteins to be essential (41% versus 9%; [8]), which suggests that most of the essential candidates are truly involved in amino acid biosynthesis.

Among the remaining clear candidate genes, we had insufficient coverage to quantify mutant fitness [10] for nine (non-essential) genes. (Some genes have few insertions because they are short, or because of random factors in library generation; also, if mutants in a gene grow slowly in rich media, then it may be difficult to study them with a pooled approach.) Another five of the clear candidates may be genetically redundant with other genes (S1 Table). There were just two clear candidates for which the lack of an auxotrophic phenotype was surprising (Psest_1986 and DvMF_1902), and we identified a potential explanation for Psest_1986 (S3 Text). Overall, we confirmed 115 of 140 of the clear candidate genes (82%) as being either auxotrophic or essential.

Of the remaining 33 gaps, two gaps were due to an error in the genome sequence or in the identification of a protein-coding gene. First, a missing step in *Azospirillum brasilense* Sp245 is due to a sequencing error that created a frameshift in histidinol dehydrogenase. Nucleotides 1,148,979 to 1,150,284 of the main chromosome (NC_016594.1; [17]) are very similar (over 80% amino acid identity) to *hisD* (AZL_d03600) from *Azospirillum sp*. B510, but the reading frame is interrupted by a frameshift. This region lies between AZOBR_RS19500 and _RS19510 and is not currently annotated with any genes. Transposon mutants in this region have reduced fitness in defined media (data of [8]), which suggested that this region of the genome is functional. We sequenced the region on both strands using Sanger sequencing and both

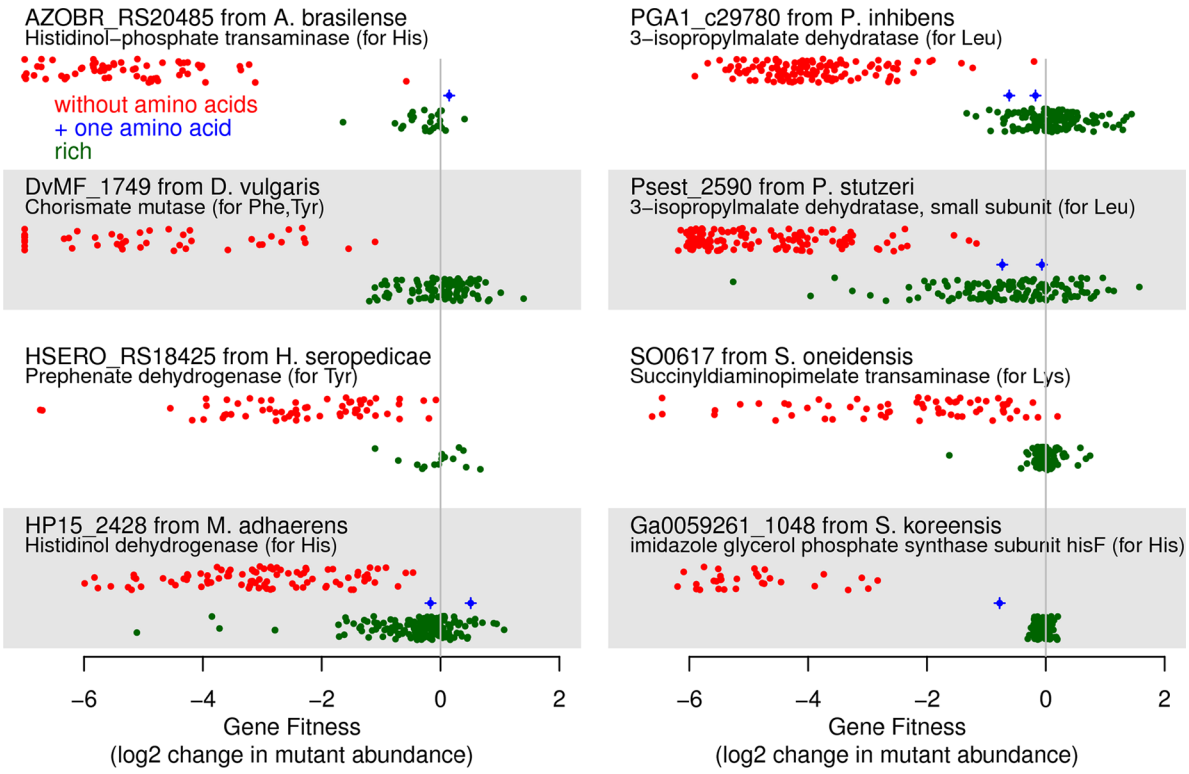

Fig 2. Mutant fitness data for clear candidates that have auxotrophic phenotypes. We selected one gene at random from each organism that has such a gene. In each panel, the *x* axis shows gene fitness (values below -7 are shown at -7), and the *y* axis separates experiments by whether most amino acids were available (green points) or not (red points). In between, we show experiments in which only the relevant amino acid was provided (blue points, if any). Within each category, the *y* axis is random.

https://doi.org/10.1371/journal.pgen.1007147.g002

reads identified a single nucleotide insertion error at nucleotide 1,149,827 of the published sequence. Once this error is corrected, there is an open reading frame for the complete *hisD* gene (it aligns over its full length and without gaps to AZL_d03600). Second, in the original annotation of *A. brasilense*, which is used in IMG, there is a pseudogene that is annotated as two parts of shikimate kinase (AZOBR_40120, AZOBR_40121). (Shikimate kinase is required for the biosynthesis of aromatic amino acids.) In the updated annotation for the same genome sequence in NCBI's RefSeq, there is instead a protein-coding gene (AZOBR_RS03225) that is annotated as shikimate kinase. The newly predicted protein does not contain any frameshifts (the error was purely due to gene calling). In a mutant library with 105,000 different transposon insertions in *A. brasilense*, there are no insertions within the 636 nucleotides of AZOBR_RS03225 [8]. This suggests that AZOBR_RS03225 is a genuine and essential protein.

Of the remaining 31 gaps, 18 were transaminase reactions, which may be nonspecific [18]. For example, IMG lists aromatic amino-acid transaminase as the necessary gene for the final step in the biosynthesis of phenylalanine and tyrosine, but *E. coli* contains multiple transaminases with overlapping substrate specificities that can perform these steps. *E. coli tyrB* and *aspC* contribute to the synthesis of tyrosine and phenylalanine; *ilvE* contributes to phenylalanine synthesis; and all three of these genes contribute to the synthesis of other amino acids as well, namely aspartate, isoleucine, and valine [19]. IMG predicted that aromatic amino-acid transaminase is missing in eight of the ten bacteria we examined, but all eight of these bacteria contain multiple amino acid transaminase genes. IMG also lists N-succinyldiaminopimelate aminotransferase as required for the synthesis of lysine via succinylated intermediates. In *E.*

coli, this activity is provided by both *argD* and *serC*, which also catalyze transamination reactions in the biosynthesis of arginine and serine [18] (also reviewed in EcoCyc, [20]). Similarly, in *Azospirillum brasilense* Sp245 and in *Phaeobacter inhibens* BS107, a putative ornithine transaminase (*argD*; AZOBR_RS19025 or PGA1_c24230) may provide the missing N-succinyldiaminopimelate aminotransferase activity. This gene is essential in both organisms [8], which is consistent with our proposal because this step is also required for peptidoglycan synthesis. IMG also failed to predict that two of the bacteria could synthesize alanine because of a missing alanine aminotransferase. *E. coli* encodes many different transaminases that can form alanine, and many of these also carry out other transamination reactions [21]. For example, one of the major alanine transaminases in *E. coli* is AvtA, which is also believed to use valine as a substrate [22]. Because many amino acid transaminases have multiple physiological substrates and because their specificity is currently difficult to annotate, the absence of a transaminase should not be used to predict auxotrophy.

After removing the transaminase reactions, 13 gaps remained, but two of these gaps had already been filled by experimental studies. First, in *Shewanella oneidensis* MR-1, SO3749 is the missing acetylornithine deacetylase for arginine synthesis [23]. Second, the sulfate-reducing bacterium *Desulfovibrio vulgaris* Miyazaki F contains recently-discovered genes for the biosynthesis of homocysteine, which is a precursor to methionine (DUF39, NIL/ferredoxin, and COG2122; [2,3,11]). Unfortunately, the information from these studies has not made its way into the annotation databases.

Of the 173 gaps in amino acid biosynthesis from the automated tool, just 11 represented genuine gaps in biological knowledge. For 9 of these 11 gaps, we provide genetic evidence for the genes that provide the missing enzymatic capabilities. In addition, our genetic data for *Desulfovibrio vulgaris* Miyazaki F provides insights into the recently-discovered pathway for methionine synthesis. We will first describe methionine synthesis in *D. vulgaris* in more detail and then each of the 9 gaps that we filled using high-throughput genetics data.

Methionine synthesis in *Desulfovibrio vulgaris*

We recently found that a DUF39 protein (DUF is short for domain of unknown function) is required for homocysteine formation in *Desulfovibrio alaskensis* [3]. A genetic study in the methanogen *Methanosarcina acetivorans* [2] also found that a DUF39 protein (MA1821) is involved in homocysteine synthesis, along with a protein containing NIL and ferredoxin domains (MA1822). (The NIL domain is named after a conserved subsequence and the PFam curators suggest that it might be a substrate binding domain.) Biochemical studies of cell extracts suggest that in methanogens, homocysteine is formed by reductive sulfur transfer to aspartate semialdehyde, and this process requires the DUF39 and/or the NIL/ferredoxin proteins [4].

In contrast, in the well-characterized pathways for methionine synthesis, aspartate semialdehyde is converted to homocysteine via multiple steps: aspartate semialdehyde is first reduced to homoserine by homoserine dehydrogenase; then the alcohol group is activated by acetylation, succinylation, or phosphorylation; and finally the sulfide is transferred to form homocysteine. The molecular functions of the DUF39 and NIL/ferredoxin proteins are not known, but a conserved cysteine in the DUF39 protein MA1821 is modified to a persulfide *in vivo*, which suggests that it is involved in the transfer of a sulfur atom to aspartate semialdehyde [24]. Rauch and colleagues [24] also propose that DUF39 might produce a thioaldehyde intermediate which could be reduced by the NIL/ferredoxin protein to yield homocysteine.

The genome of *Desulfovibrio vulgaris* Miyazaki encodes DUF39 and NIL/ferredoxin proteins and does not appear to encode any of the well-characterized pathways for methionine

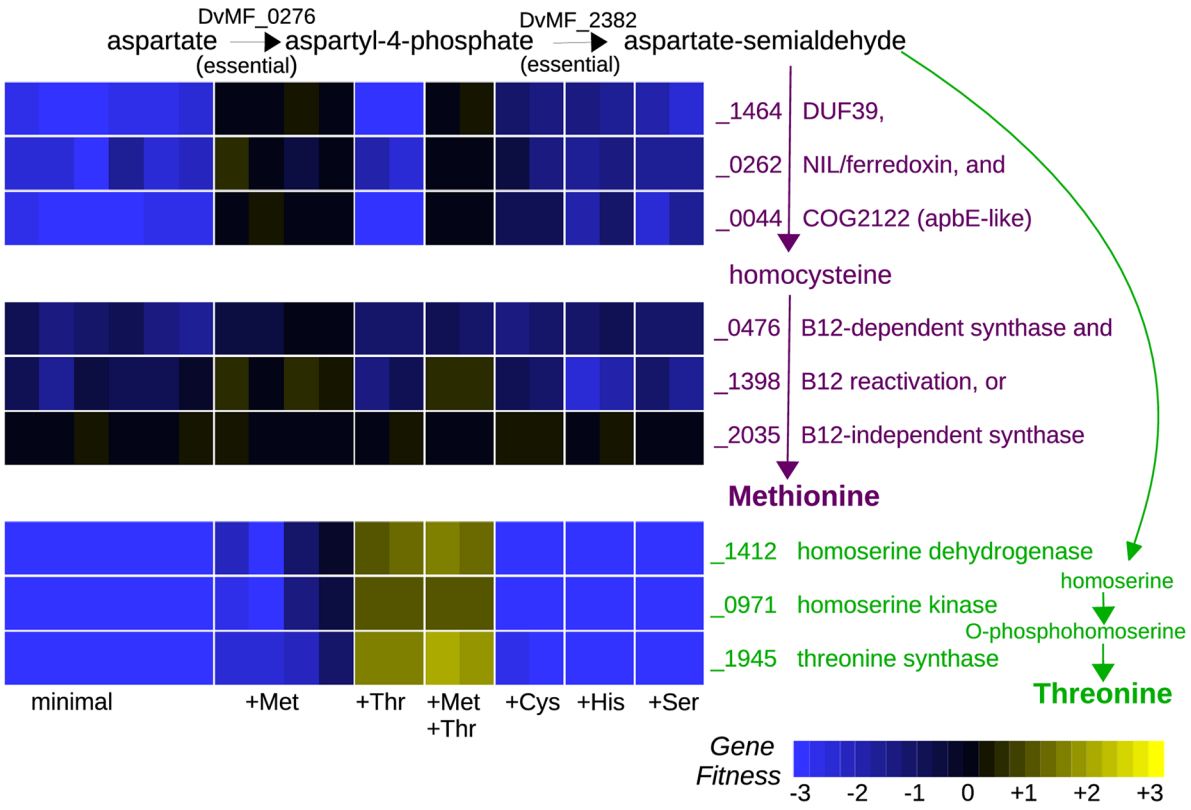

Fig 3. Synthesis of methionine and threonine in Desulfovibrio vulgaris Miyazaki F. We grew a pool of transposon mutants of *D. vulgaris* in a defined lactate/sulfate medium with or without added amino acids. L-methionine or L-threonine were supplemented at either 1 mM or 10 mM. D,L-cysteine, D,L-histidine, or L-serine were supplemented at 1 mM.

https://doi.org/10.1371/journal.pgen.1007147.g003

biosynthesis, so we expected that it would synthesize methionine by the recently discovered pathway. Indeed, we found that the DUF39 protein (DvMF_1464) and the NIL/ferredoxin protein (DvMF_0262) were important for growth in minimal media and that mutants in these genes were rescued by added methionine (Fig 3). As far as we know, this is the first experimental evidence that the NIL/ferredoxin protein is required for methionine synthesis in *Desulfovibrio*.

We also found that homoserine dehydrogenase (DvMF_1412), which is required for homoserine synthesis, was important for fitness in minimal media (Fig 3). This was expected because homoserine is an intermediate in the biosynthesis of threonine as well as in the standard pathway of homocysteine biosynthesis. Furthermore, supplementing the medium with threonine alone was sufficient to rescue the homoserine dehydrogenase mutants (Fig 3), which implies that homoserine dehydrogenase is not required for methionine synthesis. This confirms that homoserine is not an intermediate in the DUF39 pathway for homocysteine synthesis, as previously suggested based on a biochemical study of cell extracts from methanogens [4].

Comparative genomics analyses had also suggested that COG2122 might be involved in the this pathway [2,3]. (COG is short for clusters of orthologous groups.) COG2122 is distantly related to the flavin transferase ApbE [25,26], but the homology does not seem to extend to the flavin binding region of ApbE [11]. Rauch and colleagues [24] instead suggested that COG2122 might be involved in the persulfide modifications of DUF39 and also of O-phosphoseryl-tRNA-cysteinyl-tRNA synthase (SepCysS), which is required for cysteine synthesis in *M. acetivorans* (but is not present in *D. vulgaris*). We found that the COG2122 protein in *D.*

vulgaris (DvMF_0044) was important for growth in defined media and that mutants were rescued by added methionine (Fig 3). We also found that, across a variety of growth conditions, the fitness of the COG2122 protein was virtually identical to that of the DUF39 protein (r = 0.98 across 170 fitness experiments). This strongly suggests that the COG2122 protein is also required for homocysteine formation. In contrast, Rauch and colleagues [11] found that the orthologous protein from *Methanosarcina acetivorans (*MA1715) was important for growth in defined media at low sulfide concentrations, but that mutants could be rescued by either higher sulfide concentrations or by added cysteine or homocysteine. Because we grew *D. vulgaris* with sulfate as the electron acceptor (which is reduced to sulfide), and because we added 1 mM sulfide to the media as a reductant, it seemed implausible that COG2122 would be important in *D. vulgaris* because of low sulfide concentrations. Nevertheless, we performed a fitness experiment with added D,L-cysteine. We found that in *D. vulgaris*, mutants of all three homocysteine biosynthesis genes were partially rescued by added cysteine (Fig 3; mean fitness = -1.0). This suggests that *D. vulgaris* might have a minor alternate route to homocysteine, perhaps via a putative cystathionine β-lyase (DvMF_1822). In any case, our data confirm that COG2122 is involved in the conversion of aspartate semi-aldehyde and sulfide to homocysteine; its apparent dispensability for homocysteine formation in *M. acetivorans* under some conditions might be due to genetic redundancy.

In summary, we identified three genes in *D. vulgaris* that are required for homocysteine synthesis, as expected from previous studies of another species of *Desulfovibrio* and of methanogens. We provided genetic evidence that homoserine is not an intermediate in this pathway and that COG2122 is required.

A homoserine kinase for threonine synthesis in *Desulfovibrio vulgaris* that is related to shikimate kinase

IMG predicts that *D. vulgaris* is a threonine auxotroph because of a missing homoserine kinase. We identified mutants in three genes as being rescued by added threonine: threonine synthase (DvMF_1945), homoserine dehydrogenase (DvMF_1412), and DvMF_0971, which was originally annotated as a shikimate kinase (Fig 3). This observation suggests that DvMF_0971 might instead be a homoserine kinase, as both reactions involve the phosphorylation of an alcohol group. Indeed, the genome encodes another shikimate kinase (DvMF_1410) which appears to be essential, as are other genes in the chorismate synthesis pathway (DvMF_1750, DvMF_0373, DvMF_0962, DvMF_1748, and DvMF_1408). Furthermore, DvMF_1410 is similar to the shikimate kinase II from *E. coli* (47% identical), while DvMF_0971 is more distantly related (27% identical). To test our prediction that DvMF_0971 is a homoserine kinase, we cloned it and transformed it into a *thrB*- strain of *E. coli* from the Keio deletion collection [27]. This strain of *E. coli* does not grow in minimal media due to a lack of homoserine kinase activity, and its growth was rescued by the expression of DvMF_0971. Thus, DvMF_0971 is the homoserine kinase of *D. vulgaris*.

A split methionine synthase in *Phaeobacter inhibens*

The IMG website makes no prediction as to whether *Phaeobacter inhibens* DSM 17395 (formerly *P. gallaeciensis*; also known as strain BS107) can synthesize methionine because it was unable to identify a gene for methionine synthase. We propose that *P. inhibens* contains a vitamin B12-dependent methionine synthase that is split into three genes (Fig 4A). For comparison, in *E. coli*, the vitamin B12-dependent methionine synthase (MetH) contains a methyltransferase domain (PF02574), a pterin binding domain (PF00809), a vitamin B12 binding cap (PF02607), and a vitamin B12 binding domain (PF02310). In *P. inhibens*, PGA1_c13370 has the methyltransferase

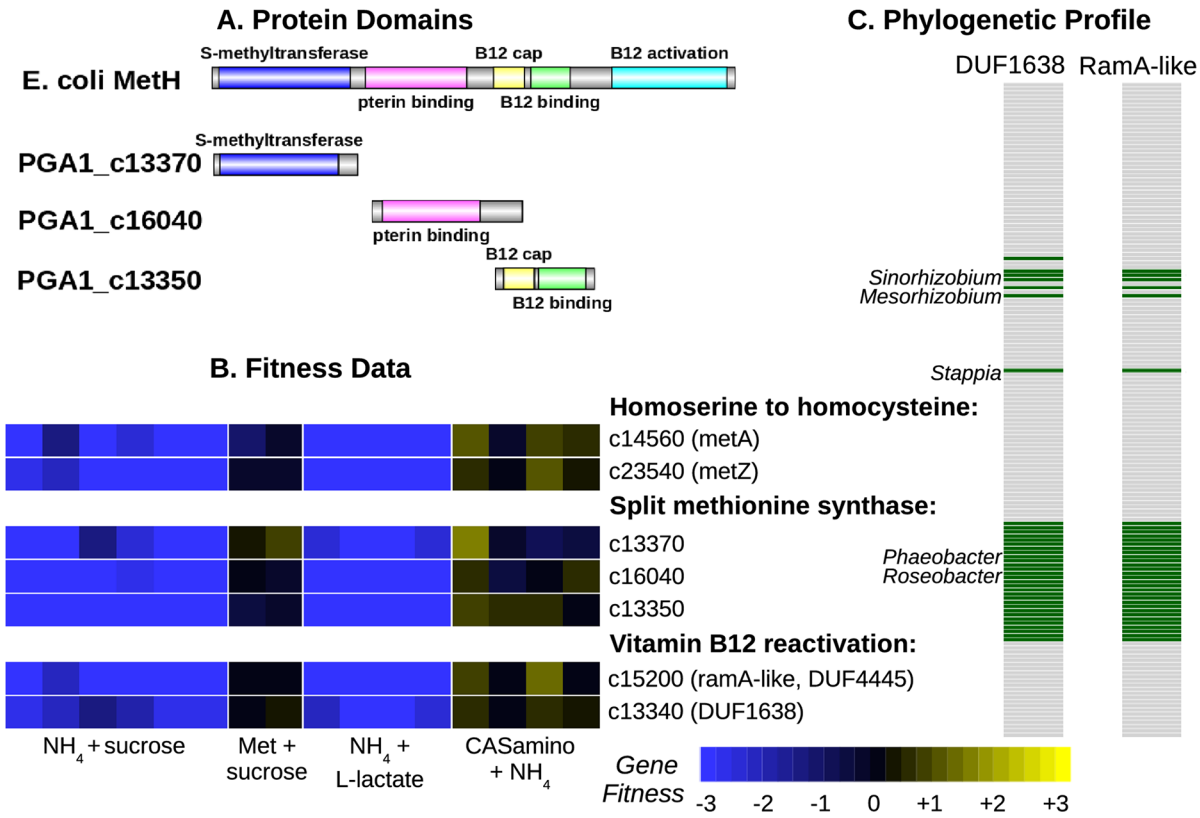

Fig 4. Methionine synthesis in *Phaeobacter inhibens* by a three-part methionine synthase and two vitamin B12 reactivation proteins. (A) Domain content of MetH from *E. coli* and of the three-part methionine synthase of *P. inhibens*. (B) Fitness data of the methionine synthesis genes. We grew pools of mutants of *P. inhibens* aerobically in defined media with a variety of carbon and nitrogen sources. (C) Phylogenetic profile of the presence or absence of the vitamin B12 reactivation proteins across 158 α-Proteobacterial genomes from MicrobesOnline [29]. The bacteria are ordered by evolutionary relationships and some of the genera that contain these proteins are labeled.

https://doi.org/10.1371/journal.pgen.1007147.g004

domain, PGA1_c16040 has the pterin-binding domain, and MtbC (PGA1_c13350) was originally annotated as "putative dimethylamine corrinoid protein" and contains the vitamin B12-binding cap and vitamin B12-binding domains. Thole and colleagues [28] previously suggested that two of these proteins might be involved in methionine synthesis. We found that all three of these genes were important for growth in defined media, and their mutants were rescued by the addition of methionine or of casamino acids (Fig 4B). Furthermore, these genes had similar phenotypes across 270 diverse fitness experiments (all $r > 0.7$, Pearson correlations). This confirms that they work together to provide the missing methionine synthase activity.

A RamA-like protein for vitamin B12 activation in *Phaeobacter* and *Desulfovibrio*

The activity of vitamin B12-dependent methionine synthase also requires the "reactivation" of vitamin B12 to reduce co(II)balamin, which can form as a side reaction of this enzyme, to co(I)balamin. In *E. coli*, the reactivation of B12 is provided by yet another domain (PF02965) at the C terminus of the MetH protein, but in other bacteria this can be a separate protein. However no member of PF02965 was found in *P. inhibens* or in related bacteria such as *Dinoroseobacter shibae*. Thole and colleagues [28] proposed that PGA1_c13360, which contains a radical SAM domain, might be involved in B12 activation, but we found that this gene was not

important for growth in minimal media (all fitness values were within -0.5 to +0.5). Instead, we identified two other genes that had correlated fitness with the other methionine synthase genes and are likely to be involved in B12 reactivation: a protein with ferredoxin and DUF 4445 domains (PGA1_c15200) and a DUF1638 protein (PGA1_c13340). As shown in Fig 4B, mutants in these genes are rescued by added methionine.

The DUF4445 protein is distantly related to RamA, which uses ATP to drive the reductive activation of corrinoids in methanogens [30]. Indeed, Ferguson and colleagues [30] predicted that bacterial homologs of RamA would be involved in vitamin B12 reactivation, and we previously proposed that in *Desulfovibrio alaskensis*, a RamA-like protein (Dde_2711) would be involved in B12 reactivation because it is cofit with MetH (r = 0.90; [3]). We also found evidence that DUF4445 is involved in the reactivation of B12 in *D. vulgaris*, which encodes a B12-dependent methionine synthase (DvMF_0476) that lacks the standard B12 activation domain. This methionine synthase has a very similar fitness pattern as DvMF_1398, which contains two DUF4445 domains (r = 0.92 across 170 experiments; also see Fig 3). We infer that DUF4445 proteins perform the reactivation of vitamin B12 in diverse bacteria.

We do not have a specific proposal for the function of the DUF1638 protein. Although it is adjacent to the gene that encodes the vitamin B12-binding cap and vitamin B12-binding domains, the DUF1638 protein is downstream and at the end of the operon, so its phenotype is unlikely to be due to polar effects. We thought that DUF1638 could be involved in the synthesis of vitamin B12 rather than in B12 reactivation *per se*, but unlike the DUF1638 protein, the genes in the vitamin B12 synthesis pathway have few insertions and are probably essential in *P. inhibens*. (The CobIGJMKFLHBNSTQDPV proteins are all essential, as is one of two CobO-like proteins. Vitamin B12 synthesis may be essential, even when methionine is provided, because of a vitamin B12-dependent ribonucleotide reductase.) Across the α-Proteobacteria, the presence or absence of DUF1638 is nearly identical to that of the RamA-like protein (Fig 4C), which is consistent with a close functional relationship. One unexplained aspect of this distribution is that some of the bacteria with the RamA-like and DUF1638 proteins also encode MetH with a standard reactivation domain (*i.e.*, *Sinorhizobium meliloti* 1021). Overall, we identified five genes that are involved in methionine synthesis in *P. inhibens*: three pieces of methionine synthase and two proteins for the reactivation of vitamin B12, including a RamA-like protein that is also involved in vitamin B12 reactivation in *Desulfovibrio*.

An "FAD-linked oxidase" is involved in serine synthesis

In *Burkholderia phytofirmans* and in *Herbaspirillum seropedicae*, we were unable to find the standard D-phosphoglycerate dehydrogenase (*serA*), which catalyzes the first step in serine biosynthesis. We propose that another oxidase provides this missing activity: BPHYT_RS 03150 in *B. phytofirmans* or HSERO_RS19500 in *H. seropedicae*. These two proteins are very similar (76% amino acid identity) and both were originally annotated as "FAD-linked oxidase." They contain an N-terminal DUF3683 domain, FAD-binding and FAD oxidase domains, a 4Fe-4S dicluster domain, two cysteine-rich CCG domains (which are often associated with redox proteins), and a C-terminal DUF3400 domain.

In *B. phytofirmans*, this oxidase was important for fitness in most defined media, but not in rich media (LB), or when casamino acids were added, or when L-serine was the nitrogen source (Fig 5A). If we supplemented our standard glucose/ammonia minimal media with L-serine, then mutants in this gene were partially rescued, as were mutants in the gene for the next step in serine synthesis (phosphoserine transaminase or *serC*; Fig 5B). *B. phytofirmans* may preferentially uptake ammonia instead of serine, which could explain why the two serine synthesis genes are only partially rescued by added serine if ammonia is present. The gene for

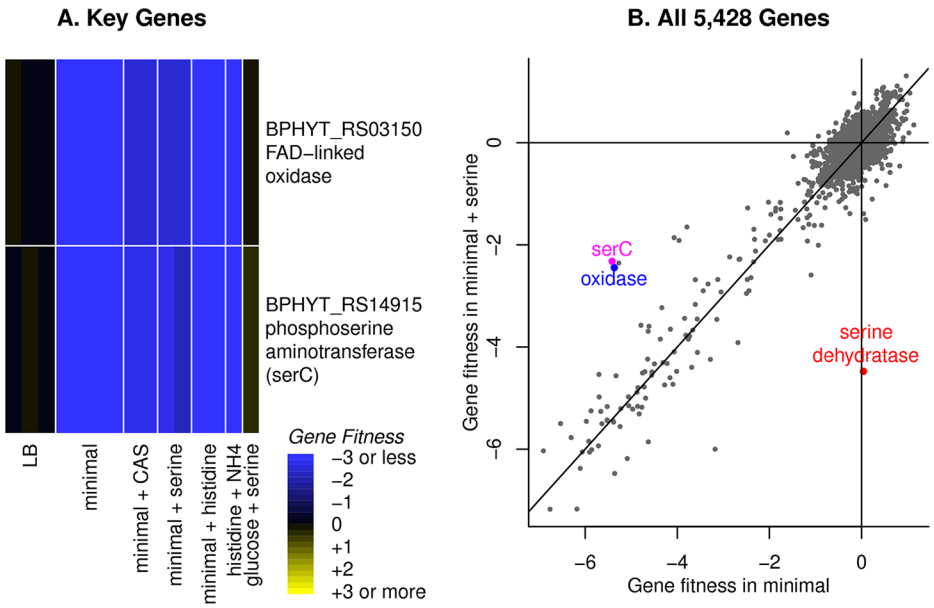

Fig 5. Serine biosynthesis in *Burkholderia phytofirmans*. (A) A heatmap of gene fitness of the putative phosphoglycerate dehydrogenase ("FAD-linked oxidase") and another serine biosynthesis gene (the aminotransferase *serC*). Our standard minimal media for this organism contains glucose and ammonia, and CAS is short for casamino acids, which contains both L-serine and L-histidine. **(B)** A comparison of gene fitness during growth in minimal media (*x* axis) or in media that was supplemented with 1 mM L-serine (*y* axis). The lines show $x = 0$, $y = 0$, or $x = y$. We highlight the genes from part (A) as well as the catabolic serine dehydratase. The other point near *serC* and the dehydrogenase is a putative cell wall synthesis gene (BPHYT_RS14855, ADP-L-glycero-D-manno-heptose-6-epimerase). Each point shows the average of two replicate experiments.

https://doi.org/10.1371/journal.pgen.1007147.g005

the final step in serine synthesis (phosphoserine phosphatase or *serC*, BPHYT_RS09200) may be important for fitness even in rich media, as mutant strains were at low abundance in our pool of mutants.

In *H. seropedicae*, this oxidase appears to be essential in rich media with amino acids. Mutants in the phosphoserine transaminase (HSERO_RS18435) and the phosphoserine phosphatase (HSERO_RS03150) were also at low abundance in our mutant pool, so the poor viability of mutants in HSERO_RS19500 is consistent with a role in serine synthesis. To look for other candidates for this step, we collected fitness data for *H. seropedicae* in minimal media with and without added L-serine, but we did not identify any genes whose mutants were rescued by added L-serine. (Averaging across two replicate experiments, there were no genes with fitness under -2 in minimal glucose media and fitness above -1 in minimal glucose media that was supplemented with 1 mM L-serine.)

We predict that in both bacteria, the phosphoglycerate dehydrogenase activity is provided by the FAD-linked oxidase that has additional DUF3683, CCG, and DUF3400 domains. However, neither BPHYT_RS03150 nor HSERO_RS19500 complemented the growth deficiency of a *serA*- strain of *E. coli* from the Keio collection in minimal media [27]. The FAD-linked oxidase might require another cofactor or protein for activity, or it might have some other unexpected role in serine synthesis. Both organisms contain genes for both of the other steps in serine synthesis (the phosphoserine transaminase *serC* and the phosphoserine phosphatase *serB*), and these genes are either essential or their mutants are auxotrophic, so we do not expect the FAD-linked oxidase to be involved in these other steps.

Histidinol-phosphate phosphatases for histidine synthesis that are similar to phosphoserine phosphatases

We propose that in four of the 10 bacteria, genes that were originally annotated as phosphoserine phosphatases provide the missing histidinol-phosphate phosphatase activities. Phosphoserine and histidinol phosphate are both of the form R-C(NH_3^+)-$CH_2OPO_3^{2-}$, so this is biochemically plausible. These genes are: BPHYT_RS03625 from *Burkholderia phytofirmans*; Psest_3864 from *Pseudomonas stutzeri* RCH2; HP15_461 from *Marinobacter adhaerens* HP15; and HSERO_RS03150 from *Herbaspirillum seropedicae* SmR1. In three of the four bacteria, this gene was important for fitness in minimal media but not in minimal media that was supplemented with histidine (Fig 6A, 6B and 6C). In *H. seropedicae*, mutants in HSERO_RS03150 are at low abundance in our pools, so we do not have fitness data for it. The genes in *H. seropedicae* whose mutants were rescued by added histidine are all annotated as performing other steps in histidine biosynthesis (Fig 6D), so our data does not suggest another candidate for this step. The four putative histidinol-phosphate phosphatases are all similar to each other ($\geq$45% amino acid identity), so the results from the various bacteria corroborate each other.

We also note that each of these bacteria contain another gene that is annotated as phosphoserine phosphatase (BPHYT_RS09200, Psest_0489, HP15_2518, and HSERO_RS15175). We believe that these genes are correctly annotated, but we only have fitness data for one of them. We found that Psest_0489 from *P. stutzeri* is important for fitness in most but not all defined media conditions. It is possible that another enzyme in this organism also has phosphoserine phosphatase activity: Psest_2327 is 89% identical to ThrH from *P. aeruginosa*, which has this activity [31]. If so, this would explain why Psest_0489 is not important for fitness in some defined media conditions with no serine added.

A phosphoribosyl-ATP diphosphatase from the MazG family

The mutant fitness data for *D. vulgaris* did not identify a candidate for phosphoribosyl-ATP diphosphatase, which is required for histidine biosynthesis. By sequence analysis, we identified DvMF_3078 as a candidate, but we did not have fitness data for this gene. DvMF_3078 is related to MazG (nucleotide pyrophosphatase), which performs a similar reaction. (In both reactions, a nucleotide 5'-triphosphate is converted to a nucleotide 5'-monophosphate.) The ortholog of DvMF_3078 in *D. alaskensis* G20 (Dde_2453) is important for fitness in minimal media and its fitness pattern is most similar to that of *hisI* (r = 0.96; data of [3,32]). These observations suggested that DvMF_3078 encodes the missing phosphoribosyl-ATP diphosphatase.

To test this hypothesis, we studied a transposon mutant of the orthologous gene from *D. vulgaris* Hildenborough (DVU1186). We found that a transposon mutant of DVU1186 (strain GZ8414) grew little if at all in minimal media and that its growth was rescued by the addition of 0.1 mM L-histidine (Fig 7). As a control, we also tested a transposon mutant in DVU2938 (a DUF39 protein which is involved in methionine synthesis) and found that it was unable to grow in defined media even if histidine was added (Fig 7). Thus, in the genus *Desulfovibrio*, a MazG family protein is required for histidine biosynthesis and is probably the missing gene for phosphoribosyl-ATP diphosphatase. It is interesting to note that the MazG family is distantly related to HisE, which provides the phosphoribosyl-ATP diphosphatase activity in most bacteria [33].

Two unfilled gaps

Two of the 11 genuine gaps remain unfilled. First, we did not identify the phosphoserine phosphatase in *D. vulgaris*. We thought that this activity might be provided by DvMF_0940, which

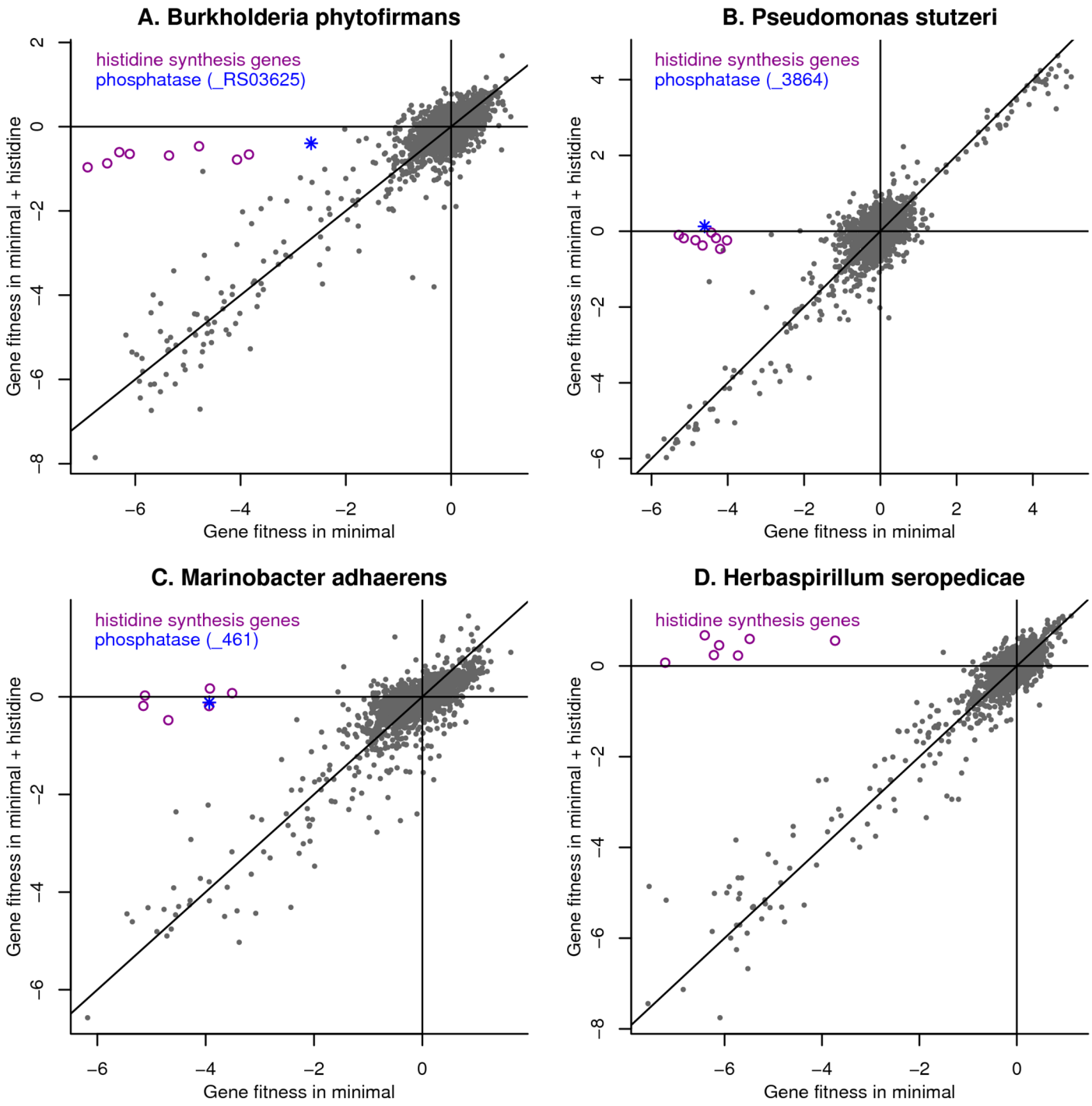

Fig 6. Identification of novel histidinol-phosphate phosphatases. We compared gene fitness in minimal media (*x* axis) or in minimal media supplemented with 1 mM L-histidine (*y* axis) for (A) *Burkholderia phytofirmans*, (B) *Pseudomonas stutzeri*, (C) *Marinobacter adhaerens*, and (D) *Herbaspirillum seropedicae*. In each panel, we highlight the putative phosphatase as well as genes for other steps in histidine biosynthesis. For *H. seropedicae*, the phosphatase is not shown because of a lack of data. The lines show $x = 0$, $y = 0$, and $x = y$. Each point shows the average gene fitness from two replicate experiments.

https://doi.org/10.1371/journal.pgen.1007147.g006

is annotated as histidinol-phosphate phosphatase and could be bifunctional. We also thought that a putative phosphatase (DvMF_1903) that lies downstream of the phosphoglycerate

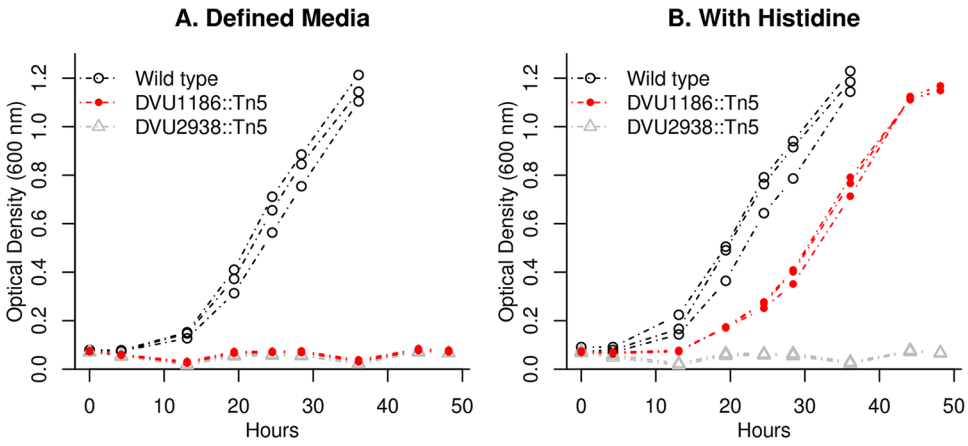

Fig 7. DVU1186 is required for histidine synthesis. We grew transposon mutants of DVU1186 (strain GZ8414) and of DVU2938 (strain GZ9865), as well as wild type *D. vulgaris* Hildenborough, in defined lactate-sulfate medium (panel A) or in the same medium with 0.1 mM L-histidine added (panel B). For each strain and condition, we show three replicates.

https://doi.org/10.1371/journal.pgen.1007147.g007

dehydrogenase (which is the previous step in serine synthesis) was a plausible candidate. Unfortunately, we do not have fitness data for either of these genes. We cloned each of these proteins into an *E. coli serB*- strain from the Keio collection, which lacks phosphoserine phosphatase, and neither of the proteins from *D. vulgaris* were able to rescue its growth in minimal media. As another test, we constructed a deletion of DVU0338 from *D. vulgaris* Hildenborough, which is similar to the putative phosphatase DvMF_1903 (77% amino acid identity). We found that the DVU0338- strain could grow in defined media. We suspect that some other protein provides the phosphoserine phosphatase activity in *D. vulgaris*.

Second, the fitness data did not identify a candidate for homoserine kinase in *Phaeobacter inhibens*. We thought that the shikimate kinase (PGA1_c14090), which is essential, might be bifunctional and act on homoserine as well. However, when we cloned this protein into an *E. coli thrB*- strain from the Keio collection, which lacks homoserine kinase, it did not rescue growth in minimal media.

Discussion

Many gaps in our understanding of amino acid biosynthesis

Although amino acid biosynthesis is well understood in model organisms such as *Escherichia coli*, our results imply that there are many steps that remain unknown, even in the relatively well-studied Proteobacteria. Once genome-wide fitness data from more diverse bacteria is available, we hope to explain many more mysteries. For example, at least one more pathway of homocysteine synthesis remains to be discovered: thermophilic autotrophs from several divisions of prokaryotes (i.e., *Aquifex aeolicus*, *Pyrolobus fumarii* 1A, and *Acidimicrobium ferrooxidans* DSM 10331) contain neither the traditional nor the DUF39 pathways of homocysteine biosynthesis, nor the protein thiocarboxylate pathway [34]. It also appears that novel homoserine kinases remain be discovered, as some bacteria that grow in minimal media encode threonine synthase but not any of the known homoserine kinases (i.e., *Bacteroides thetaiotaomicron* VPI-5482 [35] and *Dehalococcoides ethenogenes* 915 [36]).

Given our limited knowledge, it may be premature to try to predict whether bacteria can synthesize amino acids from their genome sequences. We did identify some recurring types of gaps that should not be used to predict auxotrophy. If we exclude the gaps from IMG that had

clear candidates from other annotation resources or were due to errors in gene models, then there were 31 gaps, of which 18 were transaminases, 5 were phosphatases, and 3 were kinases. We cannot determine if this information would lead to better predictions of auxotrophies, as we only studied prototrophic bacteria. If the growth requirements were known for large numbers of bacteria along with their genome sequences, it should become possible to make useful predictions. It would also help if there were a complete database of proteins that have already been experimentally characterized, as this could eliminate the "known" gaps. A related issue is that variant pathways are often not represented in pathway databases [37]. For example, as of November 2017, the DUF39 and the protein thiocarboxylate pathways of methionine synthesis are absent from MetaCyc and KEGG, and the proteins that are known to be involved in these pathways are not annotated correctly in UniProt. Finally, because there are so many proteins whose functions cannot be predicted accurately, we need new approaches to obtain large-scale data about proteins' functions.

Most free-living bacteria can synthesize all 20 amino acids

The heterotrophic bacteria that we studied are not a random sample of all heterotrophic bacteria. Nevertheless, of the heterotrophic bacteria that we studied previously, 23 of 24 grew in minimal media [8], and we selected these on the basis of their genetic tractability rather than their growth requirements. Furthermore, as far as we know, these bacteria were isolated and propagated in media that was supplemented with yeast extract, such as LB, R2A, or marine broth. We do not see why these media would select for prototrophic bacteria. So we predict that most free-living bacteria can synthesize all 20 amino acids, even though we do not know how.

By free living, we mean bacteria that live in water, soil, or sediment as opposed to living primarily within or on a larger organism (i.e., parasites, pathogens, endosymbionts, or dedicated commensal bacteria). Our impression is that auxotrophies are widespread in dedicated pathogens and in endosymbionts, while they are uncommon in plant-associated commensal bacteria. We are not sure about bacteria that are commensal with animals: it would not be surprising if bacteria that evolve in protein-rich environments would become auxotrophic. On the other hand, some animal-associated bacteria (for example *B. thetaiotaomicron*, which is abundant in the human colon), can make all of the amino acids. Also, although lactic acid bacteria are adapted to high nutrient levels, some of them are prototrophic for amino acids [38,39], and those that are auxotrophic have lost the capability to make amino acids recently [40]. Lactic acid bacteria from dairy products are more likely to be auxotrophic, which may reflect their evolution in particularly protein-rich environments [40]. We also note that even free-living bacteria with reduced genomes can synthesize all 20 amino acids. For example, consider the abundant ocean bacterium *Pelagibacter ubique*, which has a streamlined genome and has just 1,354 protein-coding genes [41]. *P. ubique* has unusual nutritional requirements for reduced sulfur compounds and for glycolate, but given these compounds, it can make all 20 amino acids [5]. These compounds are released by photosynthetic organisms in the ocean, so it appears that *P. ubique* synthesizes all 20 amino acids in nature.

A limitation of our argument is that all of the bacteria that we studied belong to the cultured minority and all are Proteobacteria. Because isolations are usually performed with media that contain peptides or amino acids, we expect that as-yet uncultured bacteria are about as likely to be auxotrophs as cultured bacteria. Similarly, because the Proteobacteria are the best-studied phylum of bacteria, we expect that there are even more gaps in the amino acid biosynthesis pathways of other bacteria. Anecdotally, we know of many gaps in amino acid biosynthesis in non-Proteobacteria. We have already mentioned gaps in *Aquifex*, *Acidimicrobium*,

Dehalococcoides, and *Bacteroides*, and S1 Text discusses erroneous predictions of auxotrophy in *Clostridium* and *Trichodesmium*. Also, some of the amino acid biosynthesis genes that we studied in Proteobacteria have likely orthologs in other phyla of bacteria. First, *Dehaloccoides ethenogenes* 915 grows in minimal media and appears to use the DUF39 pathway of homocysteine synthesis (encoded by DET0921:DET0919). Second, *Thermodesulfatator indicus* CIR29812 DSM 15286 grows in minimal media [42] and encodes a positional ortholog (Thein-DRAFT_1819) of the phosphoribosyl-ATP diphosphatase DVU1186, and its genome does not contain any other obvious candidate gene for this activity.

Under the black queen hypothesis, dependencies between organisms can be selected for if the capability is "leaky" and benefits other organisms nearby [43]. For example, an organism that degrades a toxic compound will also reduce the concentration of that compound that is experienced by its neighbors. It has been suggested that this mechanism could favor the loss of amino acid synthesis genes [7], but we argue that amino acid synthesis is not so leaky. Even if small amounts of amino acid or protein leak out of nearby cells, it seems questionable that this would provide adequate amino acids for growth, given that about half of the dry weight of bacteria is protein (BioNumbers 101955; [44]). And although some mutant strains of *E. coli* will secrete amino acids in sufficient quantities to maintain the growth of auxotrophic strains [45], we do not know of any evidence that this is occurs in nature, except for endosymbionts.

Although we are skeptical about the idea that bacteria cross-feed each other amino acids, there is some evidence for the cross feeding of vitamins [46]. Because bacteria need vitamins at far lower concentrations than they need amino acids, it seems more plausible that the black queen mechanism could apply to vitamins. Alternatively, because vitamins are present at low concentrations, they might be preferentially recycled from lysed cells rather than broken down for energy. If a subset of bacteria synthesize vitamins rather than taking them up and release them when they die, then many other bacteria would not need to synthesize vitamins. (Even if vitamins are available, some bacteria might be selected to synthesize them if they require relatively high amounts of the vitamin for their metabolism or if vitamin receptors are targeted by phage.) Nevertheless, of the seven bacteria we tested, six grew without added vitamins. Most free-living bacteria may not require exogenous vitamins for growth either.

Materials and methods

Comparison to IMG predictions

We began with 24 heterotrophic bacteria from 15 genera that we had collected large-scale mutant fitness data for [8]. We had previously found that 23 of these bacteria grew in defined media without added amino acids [8]. Since that study was conducted, we also generated a mutant library in *Herbaspirillum seropedicae* SmR1 (see below), which is a plant-associated (endophytic) and nitrogen-fixing bacterium. Although all of these bacteria have been sequenced, not all of them are available in the IMG website, so not all of them have auxotrophy predictions. Also, we arbitrarily selected one representative of each genus. This left us with 13 bacteria. One of these is *Escherichia coli*, which is a traditional model organism and, not surprisingly, IMG correctly predicts that it can make all 20 amino acids. Also, we did not do a detailed analysis for *Dinoroseobacter shibae* DFL-12 or *Dechlorosoma suillum* PS (also known as *Azospira suillum* PS). These had 5 and 6 spurious auxotrophies, respectively, which is similar to numbers for the other bacteria that we did a detailed analysis of. Predictions of amino acid synthesis capabilities were taken from the IMG web site (https://img.jgi.doe.gov/) on May 20, 2016.

Strains and media

Except for *H. seropedicae*, mutant libraries were described in [8]. The original sources of the strains are given in Table 2 along with the standard minimal media that we used for each organism and the standard carbon source that we used. These media all contain ammonium chloride as the standard nitrogen source, but this was omitted for some nitrogen source experiments. The minimal media also contain inorganic salts, buffer, and either Wolfe's vitamins or Thauer's vitamins. Media components are given in the supplementary material of [8] and are available for each experiment in the Fitness Browser (http://fit.genomics.lbl.gov/).

We also studied individual transposon mutants of DVU1186 and DVU2938 from *D. vulgaris* Hildenborough (ATCC 29579). These were obtained by using barcoded variants of the mini-Tn*5* transposon delivery vector pRL27 [47], which were delivered by conjugation. Transformants were selected on agar plates (1.5 g/L) with the antibiotic G418 (400 μg/ml) and a rich lactate-sulfate medium. Individual colonies were picked into 96 well plates and characterized by arbitrary PCR and Sanger sequencing [47].

We also studied a deletion mutant of DVU0338 from *D. vulgaris* Hildenborough (strain JW9475). This was constructed in a *upp*- background, with JW710 as the parent strain [48].

All bacteria were cultured aerobically with shaking, except that *D. vulgaris* Miyazaki (which is strictly anaerobic) was grown without shaking 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. Media for *D. vulgaris* Miyazaki was prepared in a Coy anaerobic chamber with an atmosphere of about 2% H_2, 5% CO_2, and 93% N_2. Wild type and mutant strains of *D. vulgaris* Hildenborough were grown in a similar way as *D. vulgaris* Miyazaki, but with these differences: the culture volume was only 5 mL; media was degassed for 5 min with 100% nitrogen prior to being autoclaved; and 1.4 mM thioglycolate was used as the reductant in the medium (instead of 1 mM sulfide).

Complementation assays were performed using deletion strains of *thrB*, *serA*, or *serB* from the Keio collection [27]. Genes of interest were cloned under the control of the *bla* promoter from pUC19 into the vector pBBR1-MCS5 [49]. After sequence verification, we introduced the complementation plasmids into the corresponding knockout strains by transformation. The growth of these strains was tested on M9 minimal media agar plates.

Mutagenesis of *Herbaspirillum seropedicae* SmR1

The mutant library of *H. seropedicae* will be described in more detail elsewhere. It was generated using an *E. coli* conjugation donor that contains the plasmid pKMW7 (which delivers a

Table 2. Sources of wild-type strains and their minimal media.

Strain	Source	Minimal media
Azospirillum brasilense Sp245	Stephen Fairclough, JGI	RCH2_defined_noCarbon + glucose
Burkholderia phytofirmans PsJN	DSM 17436	RCH2_defined_noCarbon + glucose
Desulfovibrio vulgaris Miyazaki F	Terry Hazen, ORNL	MoLS4_no_lactate + D,L-lactate
Herbaspirillum seropedicae SmR1	Gary Stacey, U. Missouri	RCH2_defined_noCarbon
Marinobacter adhaerens HP15	DSM 23420	DinoMM_noCarbon + L-lactate
Phaeobacter inhibens BS107	DSM 17395	DinoMM_noCarbon + L-lactate
Pseudomonas stutzeri RCH2	Romy Chakraborty, LBNL	RCH2_defined_noCarbon + glucose
Shewanella oneidensis MR-1	ATCC 700550	ShewMM_noCarbon + D,L-lactate
Sinorhizobium meliloti 1021	ATCC 51124	RCH2_defined_noCarbon + glucose
Sphingomonas koreensis DSMZ 15582	DSM 15582	RCH2_defined_noCarbon + glucose

https://doi.org/10.1371/journal.pgen.1007147.t002

Tn5 transposon with 20-nucleotide random barcodes) and is an auxotroph for diaminopimelate [10]. The transposon insertion sites were amplified as described previously [10] and sequenced using Illumina HiSeq2500 in rapid run mode. Each read links a 20 nucleotide barcode to a location in the genome. We identified insertions (supported by at least two reads) at 82,441 different locations in the 5.5 MB genome (NC_014323). We associated 98,021 different barcodes with insertions in the genome (with at least 10 reads for each of these barcodes) and estimated fitness values for 3,878 of the 4,243 non-essential proteins.

Identification of essential proteins

We identified essential proteins in *H. seropedicae* as described previously for the other bacteria [8]. This approach was validated previously [8,50]. Briefly, we limited the analysis to protein-coding genes that were long enough such that the absence of a transposon insertion in the central 10–90% of the gene would be surprising. Protein-coding genes that were long enough (at least 400 nt for *H. seropedicae*) were considered essential if the density of insertion locations (which was normalized by GC content) and the total reads (summed across all insertions) divided by the gene's length were both less than 20% of the typical protein's value. Using these thresholds, we identified 472 essential proteins (S2 Table). These proteins might not be entirely essential but they should be required for good growth in LB.

We also manually classified three steps in biosynthetic pathways as being essential. These involved genes that were not considered in the automated analysis but lack any insertions. In *A. brasilense*, shikimate kinase (AZOBR_RS03225) was originally annotated as a pseudogene, so it was not considered during the automated analysis, but it appears to be essential. Also in *A. brasilense*, aspartyl/glutamyl-tRNA amidotransferase contains three subunits, of which two were automatically identified as essential and one (AZOBR_RS20640) was too short for the automated approach but had no insertions. We classified this step as being essential. Similarly, we classified this step as essential in *P. stutzeri* despite the short length of Psest_3328, which also has no insertions.

Mutant fitness assays

Most of the mutant fitness assays that we analyzed for this study were described previously [8,10]. The compendium of mutant fitness assays for *H. seropedicae* will be described elsewhere: it includes growth in minimal media with 26 different carbon sources; growth in minimal media with 1 alternative nitrogen source; and growth in rich media with 12 different inhibitory compounds added.

For this study, we conducted additional mutant fitness assays with amino acids as additional nutrients. These assays were performed and analyzed as described previously [8]. Briefly, a pool of transposon mutants is recovered from the freezer in rich media and grown until it reaches log phase. It is then inoculated at OD_{600} = 0.02 into 5 mL of media in a glass tube and allowed to reach saturation. Gene fitness values are computed by comparing the sample after growth to the sample before growth (i.e., at the time of transfer) via genomic DNA extraction, PCR amplification of barcodes, and sequencing on Illumina HiSeq. For each fitness experiment, metrics of internal consistency and biological consistency were computed and experiments with low quality scores were discarded, as described previously [10].

Some of the samples were sequenced with a staggered "BarSeq2" primer rather than the primer we used previously. The staggered primer contains 2–5 random nucleotides just downstream of the Illumina adapter. This increases the diversity of the sequence and allows BarSeq to be conducted with the HiSeq 4000.

For the mutant libraries that we published previously [8,10], we used the same strains to estimate gene fitness, so that gene fitness values would match for the previously-published results.

Resequencing of *hisD* in *A. brasilense* Sp245

We amplified the *hisD* region from *A. brasilense* by PCR using the primers TCTCCCAGGAG GAGGTGGAC and ATCGCCTTCACGCTGTCCGCATCG. The same primers were used for Sanger sequencing.

Sequence analysis

To assign genes to TIGRFams [13], we used HMMer 3.1b1 44 [51] and the trusted score cutoff for each family in TIGRFam 15.0. TIGRFam assigns enzyme commision numbers to some of its families.

To assign genes to enzyme commission numbers via KEGG [14], we downloaded the last public release of KEGG (from 2011) and we searched for a best hit with over 30% identity and above 80% coverage using RapSearch v2.22 [52]. If the best hit was assigned an enzyme commission number by KEGG, then we transferred that annotation to the gene.

To assign genes to enzyme commision numbers via SEED and RAST [15], we used the SEED server, based on code from http://servers.nmpdr.org/sapling/server.cgi?code=server_paper_example6.pl. These results were viewed using the Fitness Browser (http://fit.genomics.lbl.gov/).

To compute the phylogenetic profile of DUF1638 and the RamA-like protein, we used MicrobesOnline [29]. For DUF1638, we used the presence or absence of PF07796. For the RamA-like protein, we used the presence of an ortholog of the RamA-like protein (VIMSS 5050244). However in Roseobacter sp. SK209-2-6, the RamA-like protein is split into two proteins (RSK20926_19262 and RSK20926_19267), and we manually classified the RamA-like protein as present in this bacterium.

Source code

Code for analyzing fitness data and for the Fitness Browser is available at https://bitbucket.org/berkeleylab/feba.

Supporting information

S1 Table. Gaps in amino acid synthesis from the IMG web site.
(XLS)

S2 Table. Essential proteins in Herbaspirillum seropedicae SmR1.
(XLS)

S1 Text. Testing the auxotroph predictions of D'Souza and colleagues.
(PDF)

S2 Text. Tests of growth without added amino acids or vitamins.
(PDF)

S3 Text. Clear candidates whose mutants were not consistently auxotrophic.
(PDF)

Author Contributions

Conceptualization: Morgan N. Price, Adam M. Deutschbauer, Adam P. Arkin.

Data curation: Adam M. Deutschbauer.

Formal analysis: Morgan N. Price.

Funding acquisition: Judy D. Wall, Adam M. Deutschbauer, Adam P. Arkin.

Investigation: Morgan N. Price, Grant M. Zane, Jennifer V. Kuehl, Adam M. Deutschbauer.

Resources: Grant M. Zane, Ryan A. Melnyk, Judy D. Wall, Adam M. Deutschbauer.

Software: Morgan N. Price.

Supervision: Judy D. Wall, Adam M. Deutschbauer, Adam P. Arkin.

Writing – original draft: Morgan N. Price.

Writing – review & editing: Morgan N. Price, Grant M. Zane, Jennifer V. Kuehl, Adam M. Deutschbauer, Adam P. Arkin.

References

1. Sheppard K, Yuan J, Hohn MJ, Jester B, Devine KM, Söll D. From one amino acid to another: tRNA-dependent amino acid biosynthesis. Nucleic Acids Res. 2008 Apr; 36(6):1813–25. https://doi.org/10.1093/nar/gkn015 PMID: 18252769
2. Rauch BJ, Gustafson A, Perona JJ. Novel proteins for homocysteine biosynthesis in anaerobic microorganisms. Mol Microbiol. 2014 Dec; 94(6):1330–42. https://doi.org/10.1111/mmi.12832 PMID: 25315403
3. Kuehl JV, Price MN, Ray J, Wetmore KM, Esquivel Z, Kazakov AE, et al. Functional genomics with a comprehensive library of transposon mutants for the sulfate-reducing bacterium Desulfovibrio alaskensis G20. MBio. 2014 May 27; 5(3):e01041–14. https://doi.org/10.1128/mBio.01041-14 PMID: 24865553
4. Allen KD, Miller DV, Rauch BJ, Perona JJ, White RH. Homocysteine is biosynthesized from aspartate semialdehyde and hydrogen sulfide in methanogenic archaea. Biochemistry. 2015 May 26; 54 (20):3129–32. https://doi.org/10.1021/acs.biochem.5b00118 PMID: 25938369
5. Carini P, Steindler L, Beszteri S, Giovannoni SJ. Nutrient requirements for growth of the extreme oligotroph "Candidatus Pelagibacter ubique" HTCC1062 on a defined medium. ISME J. 2013 Mar; 7(3):592–602. https://doi.org/10.1038/ismej.2012.122 PMID: 23096402
6. Mee MT, Collins JJ, Church GM, Wang HH. Syntrophic exchange in synthetic microbial communities. Proc Natl Acad Sci U S A. 2014 May 20; 111(20):E2149–56. https://doi.org/10.1073/pnas.1405641111 PMID: 24778240
7. D'Souza G, Waschina S, Pande S, Bohl K, Kaleta C, Kost C. Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. Evolution. 2014 Jul 9; 68(9):2559–70. https://doi.org/10.1111/evo.12468 PMID: 24910088
8. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Kuehl JV, et al. Deep Annotation of Protein Function across Diverse Bacteria from Mutant Phenotypes. bioRxiv [Internet]. 2016; Available from: http://dx.doi.org/10.1101/072470
9. Chen I-MA, Markowitz VM, Chu K, Anderson I, Mavromatis K, Kyrpides NC, et al. Improving microbial genome annotations in an integrated database context. PLoS ONE. 2013 Feb 12; 8(2):e54859. https://doi.org/10.1371/journal.pone.0054859 PMID: 23424620
10. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, et al. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. MBio. 2015 May 12; 6(3): e00306–15. https://doi.org/10.1128/mBio.00306-15 PMID: 25968644
11. Rauch BJ, Perona JJ. Efficient Sulfide Assimilation in Methanosarcina acetivorans Is Mediated by the MA1715 Protein. J Bacteriol. 2016 Jul 15; 198(14):1974–83. https://doi.org/10.1128/JB.00141-16 PMID: 27137504
12. Watson RJ, Heys R, Martin T, Savard M. Sinorhizobium meliloti cells require biotin and either cobalt or methionine for growth. Appl Environ Microbiol. 2001 Aug; 67(8):3767–70. https://doi.org/10.1128/AEM.67.8.3767-3770.2001 PMID: 11472965

13. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. Tigrfams and genome properties in 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D387–95. https://doi.org/10.1093/nar/gks1234 PMID: 23197656

14. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1; 28(1):27–30. PMID: 10592173

15. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014 Jan; 42 (Database issue):D206–14. https://doi.org/10.1093/nar/gkt1226 PMID: 24293654

16. Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, et al. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol. 2003 Oct; 185(19):5673–84. https://doi.org/10.1128/JB.185.19.5673-5684.2003 PMID: 13129938

17. Wisniewski-Dyé F, Borziak K, Khalsa-Moyers G, Alexandre G, Sukharnikov LO, Wuichet K, et al. Azospirillum genomes reveal transition of bacteria from aquatic to terrestrial environments. PLoS Genet. 2011 Dec 22; 7(12):e1002430. https://doi.org/10.1371/journal.pgen.1002430 PMID: 22216014

18. Lal PB, Schneider BL, Vu K, Reitzer L. The redundant aminotransferases in lysine and arginine synthesis and the extent of aminotransferase redundancy in Escherichia coli. Mol Microbiol. 2014 Nov; 94 (4):843–56. https://doi.org/10.1111/mmi.12801 PMID: 25243376

19. Pittard J, Yang J. Biosynthesis of the Aromatic Amino Acids. EcoSal Plus [Internet]. 2008; Available from: http://www.asmscience.org/content/journal/ecosalplus/10.1128/ecosalplus.3.6.1.8

20. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, et al. EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res. 2005 Jan 1; 33(Database issue): D334–7. https://doi.org/10.1093/nar/gki108 PMID: 15608210

21. Kim SH, Schneider BL, Reitzer L. Genetics and regulation of the major enzymes of alanine synthesis in Escherichia coli. J Bacteriol. 2010 Oct; 192(20):5304–11. https://doi.org/10.1128/JB.00738-10 PMID: 20729367

22. Whalen WA, Berg CM. Analysis of an avtA::Mu d1(Ap lac) mutant: metabolic role of transaminase C. J Bacteriol. 1982 May; 150(2):739–46. PMID: 7040341

23. Deutschbauer A, Price MN, Wetmore KM, Shao W, Baumohl JK, Xu Z, et al. Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. PLoS Genet. 2011 Nov 17; 7(11):e1002385. https://doi.org/10.1371/journal.pgen.1002385 PMID: 22125499

24. Rauch BJ, Klimek J, David L, Perona JJ. Persulfide Formation Mediates Cysteine and Homocysteine Biosynthesis in Methanosarcina acetivorans. Biochemistry. 2017 Feb 28; 56(8):1051–61. https://doi.org/10.1021/acs.biochem.6b00931 PMID: 28165724

25. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015 Jan; 43(Database issue): D261–9. https://doi.org/10.1093/nar/gku1223 PMID: 25428365

26. Bertsova YV, Fadeeva MS, Kostyrko VA, Serebryakova MV, Baykov AA, Bogachev AV. Alternative pyrimidine biosynthesis protein ApbE is a flavin transferase catalyzing covalent attachment of FMN to a threonine residue in bacterial flavoproteins. J Biol Chem. 2013 May 17; 288(20):14276–86. https://doi.org/10.1074/jbc.M113.455402 PMID: 23558683

27. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006 Feb 21; 2:2006.0008.

28. Thole S, Kalhoefer D, Voget S, Berger M, Engelhardt T, Liesegang H, et al. Phaeobacter gallaeciensis genomes from globally opposite locations reveal high similarity of adaptation to surface life. ISME J. 2012 Dec; 6(12):2229–44. https://doi.org/10.1038/ismej.2012.62 PMID: 22717884

29. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, et al. MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res. 2010 Jan; 38(Database issue):D396–400. https://doi.org/10.1093/nar/gkp919 PMID: 19906701

30. Ferguson T, Soares JA, Lienard T, Gottschalk G, Krzycki JA. RamA, a protein required for reductive activation of corrinoid-dependent methylamine methyltransferase reactions in methanogenic archaea. J Biol Chem. 2009 Jan 23; 284(4):2285–95. https://doi.org/10.1074/jbc.M807392200 PMID: 19043046

31. Singh SK, Yang K, Karthikeyan S, Huynh T, Zhang X, Phillips MA, et al. The thrH gene product of Pseudomonas aeruginosa is a dual activity enzyme with a novel phosphoserine:homoserine phosphotransferase activity. J Biol Chem. 2004 Mar 26; 279(13):13166–73. https://doi.org/10.1074/jbc.M311393200 PMID: 14699121

32. Price MN, Ray J, Wetmore KM, Kuehl JV, Bauer S, Deutschbauer AM, et al. The genetic basis of energy conservation in the sulfate-reducing bacterium Desulfovibrio alaskensis G20. Front Microbiol. 2014 Oct 31; 5:577. https://doi.org/10.3389/fmicb.2014.00577 PMID: 25400629

33. Moroz OV, Murzin AG, Makarova KS, Koonin EV, Wilson KS, Galperin MY. Dimeric dUTPases, HisE, and MazG belong to a new superfamily of all-alpha NTP pyrophosphohydrolases with potential "housecleaning" functions. J Mol Biol. 2005 Mar 25; 347(2):243–55. https://doi.org/10.1016/j.jmb.2005.01.030 PMID: 15740738

34. Krishnamoorthy K, Begley TP. Protein thiocarboxylate-dependent methionine biosynthesis in Wolinella succinogenes. J Am Chem Soc. 2011 Jan 19; 133(2):379–86. https://doi.org/10.1021/ja107424t PMID: 21162571

35. Varel VH, Bryant MP. Nutritional features of Bacteroides fragilis subsp. fragilis. Appl Microbiol. 1974 Aug; 28(2):251–7. PMID: 4853401

36. He J, Holmes VF, Lee PKH, Alvarez-Cohen L. Influence of vitamin B12 and cocultures on the growth of Dehalococcoides isolates in defined medium. Appl Environ Microbiol. 2007 May; 73(9):2847–53. https://doi.org/10.1128/AEM.02574-06 PMID: 17337553

37. de Crécy-Lagard V. Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway. Comput Struct Biotechnol J. 2014; 10(16):41–50. https://doi.org/10.1016/j.csbj.2014.05.008 PMID: 25210598

38. Cocaign-Bousquet M, Garrigues C, Novak L, Lindley ND, Loublere P. Rational development of a simple synthetic medium for the sustained growth of Lactococcus lactis. Journal of Applied Microbiology. 1995; 79(1):108–16.

39. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, et al. Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci U S A. 2006 Oct 17; 103(42):15611–6. https://doi.org/10.1073/pnas.0607117103 PMID: 17030793

40. Bringel F, Hubert J-C. Extent of genetic lesions of the arginine and pyrimidine biosynthetic pathways in Lactobacillus plantarum, L. paraplantarum, L. pentosus, and L. casei: prevalence of CO(2)-dependent auxotrophs and characterization of deficient arg genes in L. plantarum. Appl Environ Microbiol. 2003 May; 69(5):2674–83. https://doi.org/10.1128/AEM.69.5.2674-2683.2003 PMID: 12732536

41. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, et al. Genome streamlining in a cosmopolitan oceanic bacterium. Science. 2005 Aug 19; 309(5738):1242–5. https://doi.org/10.1126/science.1114057 PMID: 16109880

42. Oberhardt MA, Zarecki R, Gronow S, Lang E, Klenk H-P, Gophna U, et al. Harnessing the landscape of microbial culture media to predict new organism-media pairings. Nat Commun. 2015 Oct 13; 6:8493. https://doi.org/10.1038/ncomms9493 PMID: 26460590

43. Morris JJ, Lenski RE, Zinser ER. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio. 2012 May 2; 3(2).

44. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers—the database of key numbers in molecular and cell biology. Nucleic Acids Res. 2010 Jan; 38(Database issue):D750–3. https://doi.org/10.1093/nar/gkp889 PMID: 19854939

45. Pande S, Merker H, Bohl K, Reichelt M, Schuster S, de Figueiredo LF, et al. Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. ISME J. 2014 May; 8(5):953–62. https://doi.org/10.1038/ismej.2013.211 PMID: 24285359

46. Seth EC, Taga ME. Nutrient cross-feeding in the microbial world. Front Microbiol. 2014 Jul 8; 5:350. https://doi.org/10.3389/fmicb.2014.00350 PMID: 25071756

47. Oh J, Fung E, Price MN, Dehal PS, Davis RW, Giaever G, et al. A universal TagModule collection for parallel genetic analysis of microorganisms. Nucleic Acids Res. 2010 Aug; 38(14):e146. https://doi.org/10.1093/nar/gkq419 PMID: 20494978

48. Keller KL, Bender KS, Wall JD. Development of a markerless genetic exchange system for Desulfovibrio vulgaris Hildenborough and its use in generating a strain with increased transformation efficiency. Appl Environ Microbiol. 2009 Dec; 75(24):7682–91. https://doi.org/10.1128/AEM.01839-09 PMID: 19837844

49. Kovach ME, Elzer PH, Hill DS, Robertson GT, Farris MA, Roop RM, et al. Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes. Gene. 1995 Dec 1; 166(1):175–6. PMID: 8529885

50. Rubin BE, Wetmore KM, Price MN, Diamond S, Shultzaberger RK, Lowe LC, et al. The essential gene set of a photosynthetic organism. Proc Natl Acad Sci U S A. 2015 Dec 1; 112(48):E6634–43. https://doi.org/10.1073/pnas.1519220112 PMID: 26508635

51. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011 Oct 20; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

52. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics. 2012 Jan 1; 28(1):125–6. https://doi.org/10.1093/bioinformatics/btr595 PMID: 22039206