DOE BSSD Metrics Progress Report Q2: 03/24/2022

SFA Laboratory Research Manager: Paul D. Adams[1], PDAdams@lbl.gov
SFA Technical Co-Manager: Adam P. Arkin[1], APArkin@lbl.gov
[1]Lawrence Berkeley National Laboratory, Berkeley CA 94720

Investigators: Paul D. Adams[1,2], Adam P. Arkin[1,2], Nitin S. Baliga[3], Romy Chakraborty[1], Adam M. Deutschbauer[1], Matthew W. Fields[4], Terry C. Hazen[5,6], Trent R. Northen[1], Michael W.W. Adams[7], Eric J. Alm[8], John-Marc Chandonia[1], Aindrila Mukhopadhyay[1], Gary E. Siuzdak[9], David A. Stahl[10], Peter J. Walian[1], Jizhong Zhou[11]

Participating Institutions: [1]Lawrence Berkeley National Laboratory, Berkeley CA 94720; [2]University of California at Berkeley, CA 94704; [3]Institute for Systems Biology, Seattle, WA 98109; [4]Montana State University, Bozeman, Mt 59717; [5]University of Tennessee, Knoxville, TN 37916; [6]Oak Ridge National Laboratory, Oak Ridge, TN 37831; [7]University of Georgia, Athens, GA 30602; [8]Massachusetts Institute of Technology, Cambridge, MA 02139; [9]Scripps Research Institute, La Jolla, CA 92037; [10]University of Washington, Seattle, WA 98105; [11]University of Oklahoma, Norman, OK 73072

BSSD 2022 Performance Metric Q2

Q2 Target: Progress on new techniques for characterizing microbial isolates in the laboratory

Introduction

LBNL ENIGMA (Ecosystems and Networks Integrated with Genes and Molecular Assemblies) SFA is a multi-disciplinary, multi-institutional research effort focused on addressing foundational knowledge gaps in groundwater and sediment microbiomes in the shallow subsurface at the contaminated Oak Ridge Field Research Site (FRC). ENIGMA is a consortium of 16 investigators at eleven institutions across the country led by Lawrence Berkeley National Laboratory. Established in 2009, ENIGMA researchers collaborate to create a multiscale, causal and predictive model of the reciprocal impacts of microbial communities on critical processes within the ecosystem (e.g., on N-cycling). Efforts focus on studying subsurface microbiomes within the contaminated Bear Creek watershed at the Oak Ridge Reservation (ORR), a site with complex gradients of contaminants, generated by research and production of nuclear materials, including nitrate, acidity, uranium, technetium and volatile organic carbon species, the fate of which is mediated in large part by the activity of subsurface microbial communities. The ENIGMA workflow uses sophisticated, increasingly model-driven, field experiments to discover the components of and measure the natural and anthropogenically perturbed dynamics of these geochemical and biological processes. From these we infer the chemical, physical and microbial interactions predictive of these dynamics and estimate the ecological forces, both stochastic and deterministic, that shape community function. We then deploy a unique array of culturing, genetic, physiological and imaging technologies to capture this diversity in the laboratory and map the genetic basis for observed behaviors. Laboratory consortia are used to map gene function and investigate material flow within and among cells in conditions that simulate relevant field processes.

Our ambition is to do so at sufficient resolution to causally predict the active biotic and abiotic mechanisms mediating key processes such as denitrification; dissect the dispersing and persistent microbial community components critical in space and time during these processes; and ultimately predict the future changes in contaminant fate from current observations and possibly arising from natural and anthropogenic perturbations. Outcomes are significant both in the fundamental science of community ecology and in gaining an applied understanding of biologically-mediated subsurface processes in contaminated sediments.

Critical to our goal of a predictive microbial ecology at the resolution of single genes is the need to rapidly advance the characterization of diverse microbial isolates in the laboratory.

The need for microbial characterization in the laboratory

The advent of next-generation DNA sequencing has led to a massive expansion in the number of sequenced bacteria [1]. However, much of what we know about these bacteria can only be predicted from their genome sequences, as almost none of these organisms have been experimentally investigated in the laboratory. This would not be a burden if we could accurately predict phenotypes and ecological niches from genome sequence alone, but unfortunately this is not yet a reality, largely due the astonishing diversity of bacteria and their proteomes, and the relatively limited availability of experimental data for bacteria. For an environmental microbiology project like ENIGMA that endeavors to link complex field observations to mechanistic insights at the single-gene and metabolite-level, existing experimental data from primarily model bacteria such as *Escherichia coli* and gene annotations derived by homology from

previously characterized proteins are insufficient to meet our aims, as the bacteria and the microbial communities we aim to understand are simply too numerous and diverse to be accurately characterized by these existing approaches. To meet this challenge, ENIGMA has developed a suite of experimental and computational tools to characterize non-model bacterial isolates in the laboratory [2]. These experimental tools are designed to be flexible, such that the techniques can be applied to a wide diversity of bacteria, and inexpensive, to enable the multi-omic characterization of hundreds to thousands of strains per year. The ultimate outcomes of this effort, which we term the ENIGMA Environmental Atlas, include the accurate mapping of genotype to phenotypes of field relevant microbes, discovery of thousands of new gene functions, and an experimental-computational framework for others to leverage for the systematic characterization of bacterial isolates and their gene products from other environments. Furthermore, the characterization data is being used to prioritize the development and investigation of synthetic microbial communities (SynComs) that recreate key phenomena observed in the field.

Overview of ENIGMA bacterial characterization

The ENIGMA pipeline for characterizing isolates is outlined in Figure 1 and begins with improved approaches for complete genome sequencing of strains isolated from the environment. Wild-type bacteria are then phenotyped under a set of field-inspired conditions, including growth on diverse substrates, sensitivity to toxic metals and low pH that are known selective pressures at the ORNL FRC, and ability to uptake and transform metabolites (exometabolomics). Strains are then assessed for genetic accessibility using established tools including random barcode transposon-site sequencing (RB-TnSeq), as well as developing capabilities for gene overexpression and CRISPR interference. For select strains, we are uncovering new mechanisms of gene regulation using DNA affinity purification sequencing (DAP-seq), and leveraging ENIGMA advances in proteomics and metabolomics to dissect key pathways in isolates and in their simplified synthetic communities (SynComs). The data from these pipelines are being fed into increasingly predictive computational tools for linking genotype-phenotype, for systematically assessing the distribution of organisms and their gene products in the environment, and to uncover the functions of previously uncharacterized proteins. Below we briefly highlight recent ENIGMA advances in each step of our characterization pipeline, specific examples of how we are using this pipeline to uncover novel biology of ORNL FRC *Rhodanobacter* isolates, and the future outlook.

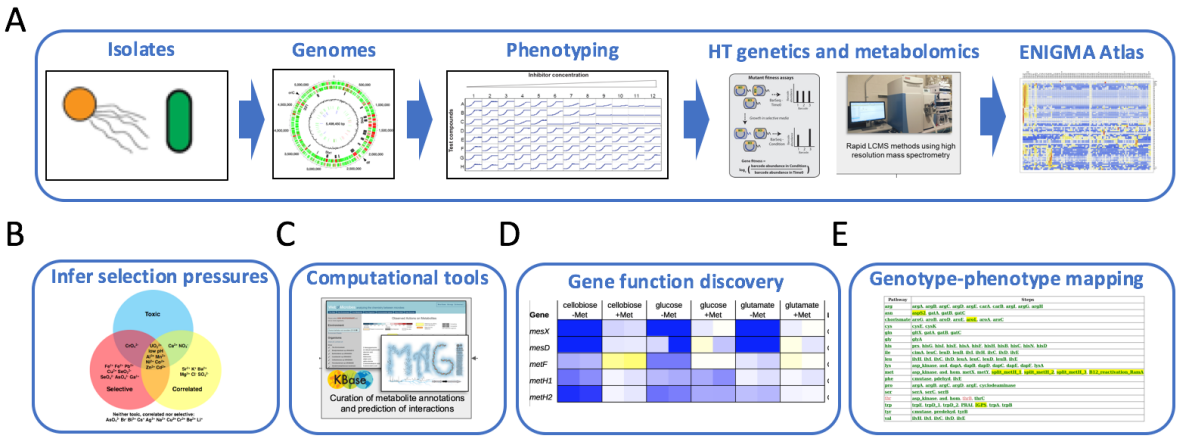


Figure 1. Overview and applications of ENIGMA characterization pipeline. A) Standard pipeline from isolates to

rich experimental data including metabolomics and genetics data feeds into ENIGMA Atlas. (B-E) Applications of ENIGMA Atlas include correlating laboratory growth measurements to field data (microbial abundance and geochemistry) to infer key selective pressures in the environment [3], new computational tools for analyzing and visualizing high-throughput phenotyping data [4], systematic discovery of new gene functions [5], and tools for improved inference of metabolic pathways from genome sequence alone [6].

Genome sequencing and annotation: Quality genome sequences are a prerequisite for any predictive mapping of genotype-phenotypes relationships. Traditionally, most bacterial genomes have been assembled solely using short-read Illumina data, which has prevented the assembly of complete genome sequences. However, complete, non-draft genomes are ideal for genotype-phenotype mapping, because they provide the most accurate representation of the best overall gene content of the organisms, can recover plasmids which are important mediators of horizontal gene transfer at the ORNL FRC [7], and can serve as reference genomes for assembling and interpreting metagenomic-assembled genomes, including those assembled using ENIGMA's Jorg tool [8]. Our current ENIGMA approach is the use of short-read Illumina sequencing to generate draft genomes in high-throughput, and for many of these bacteria we apply long-read sequencing to obtain complete genomes, using either PacBio or an in-house and optimized Oxford Nanopore approach. Using our hybrid assembly approach, we are now regularly generating extremely high-quality complete genomes (with no estimated errors) for dozens of diverse bacteria, and we anticipate scaling this strategy to the hundreds of FRC strains that currently only have draft genomes. We have found that draft genome assemblies generated from Illumina data often underestimate the actual gene content of a bacterium, when revealed through long-read sequencing [9]. All ORNL FRC genomes are currently housed in the DOE Systems Biology Knowledgebase (KBase) [10] for annotation, data sharing, and comparative genomic analysis by SFA team members.

High-throughput phenotyping of wild-type bacteria: Compared to the rapid explosion of bacterial genome sequences, the rate of phenotypic characterization of bacteria has not kept pace, despite prior advances in microbial phenotyping including phenotype microarray technology [11], panels of antibiotic sensitivity data for clinical pathogens [12], and curated databases of microbial phenotypes mined from systematics manuscripts [13]. However, there are limitations of these approaches including the lack of customization of the phenotype microarray compound panels for environments of interest, and the relatively sparse phenotypic data available from systematics papers. ENIGMA strives to mimic the field in high-throughput (as well as exercise fundamental metabolic and stress response pathways), and generates this data for strains with complete genomes, while simultaneously recording all metadata in a standardized, machine-readable format. In previous work, we have derived custom panels of field-observed carbon and nitrogen sources for high-throughput growth assays [14,15], and custom panels of metals and other inorganic ions [3] for screening the phenotypes of isolates across hundreds of conditions. For example, this high-throughput phenotyping data has been used to identify stressors that selectively impact the fitness of ENIGMA ORNL FRC isolates in the laboratory (Figure 2).

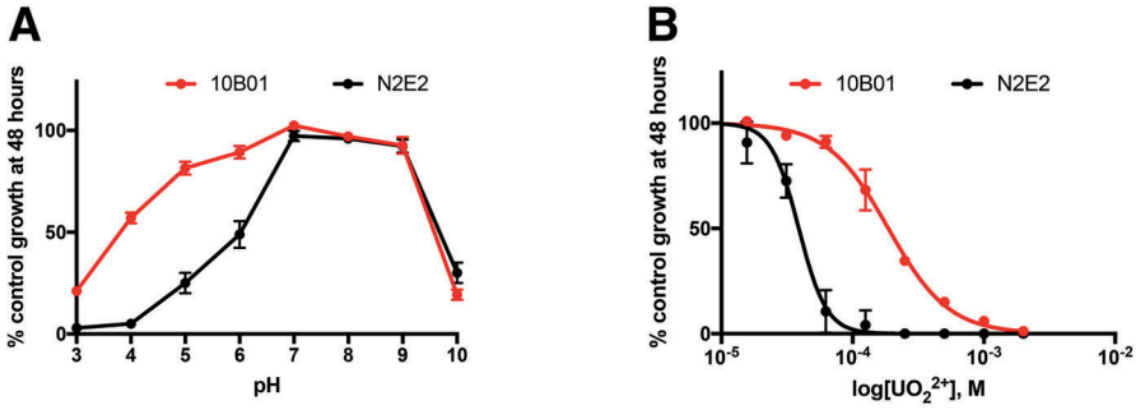


Figure 2. Dose-response data illustrating the sensitivity of *Rhodanobacter* sp. FW104-10B01 and *Pseudomonas* sp. FW300-N2E2 to pH (A) and uranyl acetate (B). Close relatives of 10B01 are predominant in ORNL FRC locations with heavy metal contamination and low pH levels, while N2E2 was isolated from a non-contaminated ORNL FRC background location.

Currently, we are developing new compound/media arrays for rapid characterization of ORNL FRC isolates to identify optimal growth conditions, determine their sensitivity to the common antibiotics ENIGMA uses for genetic engineering, and measure their responses to the major selective pressures at the ORNL FRC including toxic metals and low pH (as illustrated in Figure 2). Future efforts will include the expansion of our carbon and nitrogen compounds to additional metabolites observed at the ORNL FRC [16], expand to high-throughput anaerobic phenotyping including new automated approaches recently developed by the ENIGMA team [17], and assay the phenotypic impact of combinations of stresses to identify compounds with synergistic effects. These growth-based data are being incorporated into a database that can be mined en masse for global analyses, and also accessed interactively through a web browser.

In addition to growth-based phenotypes, we are also systematically assaying FRC isolates using exometabolomics, an approach that measures the ability of a bacterium to consume and transform the compounds in its environment [18]. ENIGMA has pioneered the development of exometabolomics as a powerful laboratory tool for characterizing, among other things, the varying nutrient preferences of our isolates and how differences in substrate preferences impact microbial communities [19]. Current efforts include the scaling of exometabolomics to a greater diversity of isolates, the development of improved metabolite mixes that better capture the chemistry of the ORNL FRC, and leveraging the data to develop improved metabolic models of isolates and synthetic communities derived from them.

High-throughput genetics: ENIGMA has a long history of developing genetic tools for non-model bacteria, including early efforts on targeted chromosomal engineering of sulfate-reducing bacteria [20]. With the maturation of next-generation sequencing, we have now developed a number of sequencing based approaches to interrogate the functions of bacterial genes and uncover mechanisms of their regulation at scale. Prominently, we have shown that one of these tools, random barcode transposon site sequencing (RB-TnSeq), can be applied en masse across multiple bacteria and many experimental conditions to systematically uncover phenotypes and functions for thousands of bacterial genes that previously were poorly understood [21]. Subsequent efforts have streamlined RB-TnSeq development in additional bacteria using libraries of barcoded transposon vectors (i.e. magic pools) [22], and developed

complementary gain-of-function approaches to examine the phenotypic impact of gene overexpression (Dub-seq) [23]. We were also involved in the development of a general loss-of-function technique called CRISPRi which our group has used to investigate host factors important for phage infection [24, 25]. To characterize gene regulation and signal transduction at scale, we have pioneered DAP-seq to identify binding motifs for transcriptional regulators in ORNL FRC isolates, and have used these tools to dissect the complex regulatory mechanisms these bacteria use to respond to environmental stresses present at the ORNL FRC [26].

To highlight the evolution of our approaches, we have recently completed a comprehensive genetic survey of one of ENIGMA's legacy bacteria, the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough (DvH). Using RB-TnSeq fitness profiles across hundreds of experiments, we identified a conditional phenotype for 1,137 non-essential protein coding genes including hundreds of genes with no previously known function [27]. Deeper exploration of this large genetic dataset resulted in a number of novel insights including the identification of novel proteins involved in the synthesis of biotin and pantothenic acid (Figure 3). In addition, we have now made all of our DvH resources (mutant libraries, individual deletion strains, individual transposon mutants) publicly available for community-wide use [28], and to date we have shared these resources with multiple external investigators. To date, we have generated RB-TnSeq libraries in 66 diverse bacteria, performed nearly 14,000 genome-wide fitness assays, generated over 50 million gene fitness measurements, and identified phenotypes for thousands of previously uncharacterized bacterial genes, many of which we have been able to propose specific functions for.

Current and future efforts include the development of novel DNA constructs and delivery methods to extend RB-TnSeq to Gram-positive ORNL FRC isolates, extend the Dub-seq approach to environmental DNA, scaling up DAP-seq (in collaboration with The Joint Genome Institute) to key isolates, and the development of new magic-pool like approaches to streamline expression system and CRISPRi development in new isolates. To enable the maximum re-use of DNA constructs, the approaches being spearheaded by different labs use universal DNA parts (including plasmids and priming regions for DNA barcode sequencing), thereby enabling the ENIGMA team to quickly share reagents and incorporate them into their own lab's workflows. Lastly, we continue to explore the physiology of key FRC isolates, increasingly using a combination of our omics approaches to derive insights that cannot be gleaned from a single approach alone. In one recent example in the ORNL FRC isolate *Pantoea* sp. MT58, we applied global mutant fitness profiling using RB-TnSeq and untargeted metabolomics to investigate the molecular mechanisms through which toxic metals inhibit bacteria [29]. Using a combination of approaches, we demonstrated that Al3+ binds intracellular arginine and that chromium ions enter the cell via a sulfate transporter.

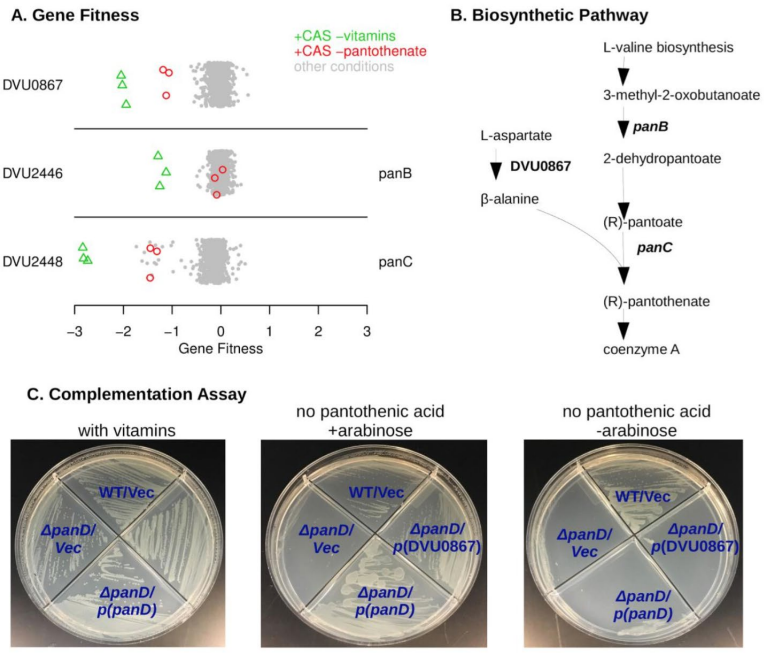


Figure 3. Identification of an atypical L-aspartate decarboxylase for pantothenic acid synthesis. (A) Gene fitness data shows that the phenotypes of DVU0867 are similar to those of two genes (*panBC*) known to be involved in the synthesis of pantothenic acid. (B) Biosynthetic pathway for pantothenic acid, including the proposed role for DVU0867. (C) Cross-species genetic complementation demonstrates that DVU0867 can functionally complement a *panD* mutation in *E. coli*, thus demonstrating that DVU0867 can perform the L-aspartate decarboxylase activity of PanD.

Computational tools for genotype-phenotype mapping: Ultimately, we strive to accurately predict all phenotypes of a bacterium from the seemingly simple (does the bacterium make all 20 amino acids?) to the very complex (why is this bacterium present at this place and time in a native environment and is it active?) from its genome sequence alone. Meeting this grand challenge would have a major impact on microbiology as it will require us to uncover new gene functions and pathways at scale, and also to make increasingly realistic measurements of bacterial phenotypes under controlled laboratory settings. Furthermore, we aim to compile a genomic and phenotypic dataset of sufficient size such that the annotation of new genomes is greatly approved by a wealth of experimental evidence for close homologs, which will support improved computational prediction of phenotype directly from genome sequence, for example by metabolic flux balance modeling [30]. In addition to the necessity for additional experimental data for a diversity of bacteria and protein families, we also recognize that the concurrent development of computational tools for data integration, exploration, storage, and display is critical for us (and others) to succeed in this massive endeavor. As such, we aim to make all of our genotypic (genome sequences) and phenotypic data FAIR (Findability, Accessibility, Interoperability, and Reusability) [31], such that our data can be maximally used by the ENIGMA team and external researchers.

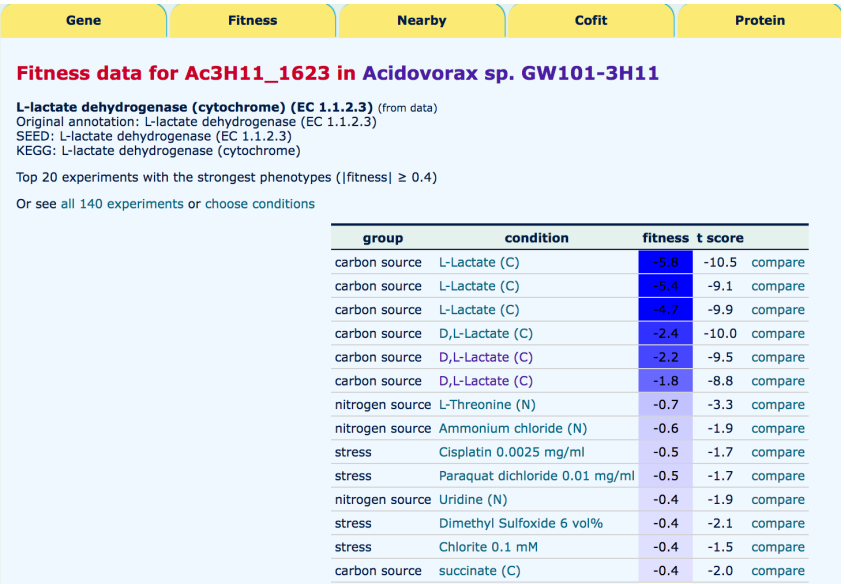
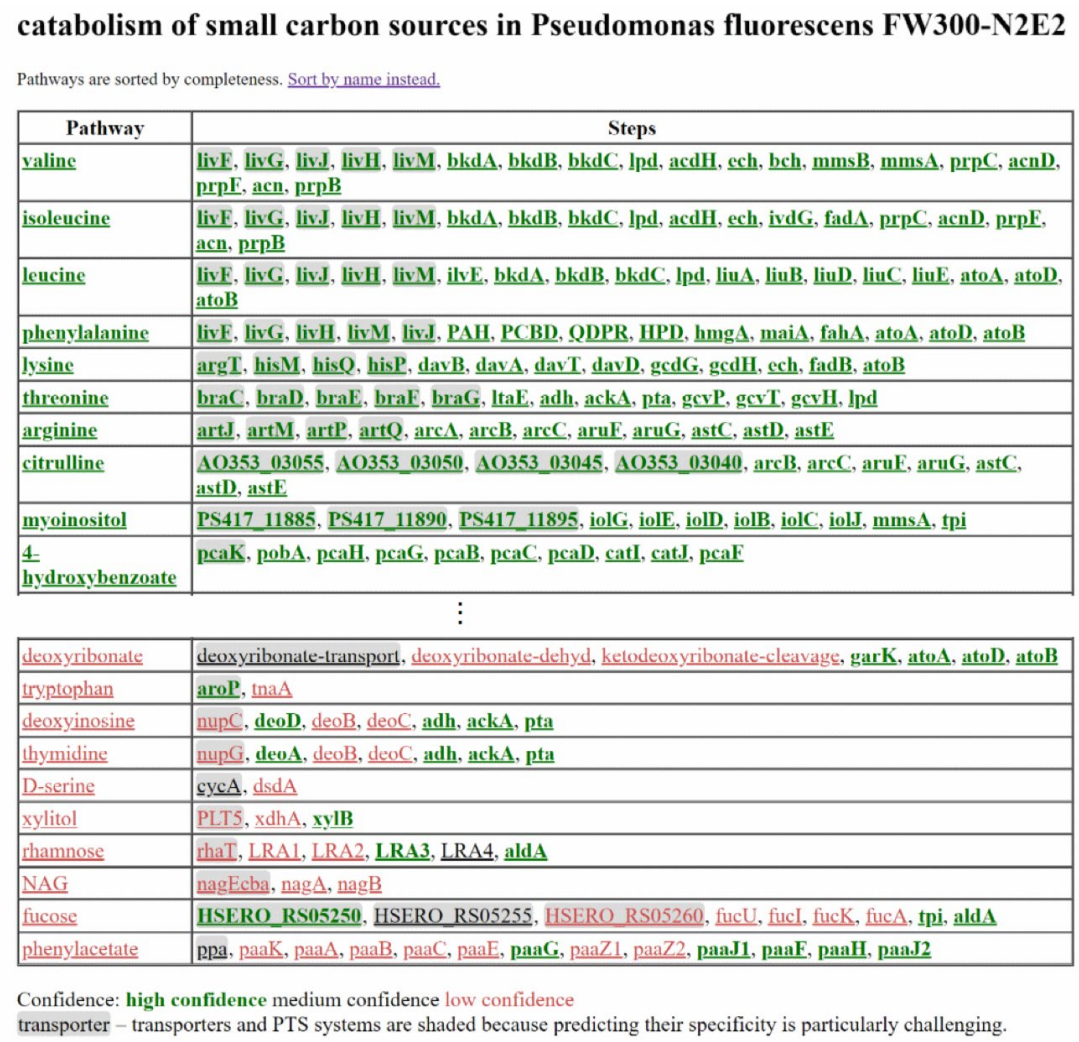


Figure 4. Fitness browser for comparative and interactive analysis of RB-TnSeq data. Screenshot of fitness browser showing the most significant phenotypes for Ac3H11_1623, a predicted L-lactate dehydrogenase from ENIGMA ORNL FRC isolate *Acidovorax* sp. GW101-3H11. As expected, this gene is only required for maximal fitness when grown on either L-lactate or D,L-lactate as the sole carbon source.

Our computational and data analysis efforts are leveraging existing resources including KBase, as well as our own internal systems including: (i) CORAL as a central platform for data storage and data integration; (ii) a flexible comparative genome browser that can be rapidly generated using only genome .gbk files as the starting point; (iii) the fitness browser for RB-TnSeq data (https://fit.genomics.lbl.gov/) (Figure 4); (iv) isolate browser for interactive analysis of growth data, (v) Web of Microbes for exometabolomics data ([4]; http://www.webofmicrobes.org/), and (vi) METLIN/XCMS for cloud-based analysis of metabolomics data (https://xcmsonline.scripps.edu) [32]. We have historically also built tools to characterize the phylogeny of organisms and genes (FASTTREE) [33], map genes in a new genome to function through linkage to experimentally characterized genes from the literature (PaperBlast)[34]. These tools make the information in these portals richer and give us and the community tools for better functional assessment of their genomes. Increasingly, we are developing new links between these data portals to enable ENIGMA science, including the development of new computational approaches to correlate and visualize our isolate genome sequences, our phenotypic data, and the distribution of these (and related) bacteria at the ORNL FRC site (from metagenomics and geochemical data). In the near term, we aim to develop a web-based computational infrastructure that enables ENIGMA researchers to quickly develop hypotheses about field phenomena that can be rapidly tested using a suite of cultured isolates and SynComs, all informed by the ENIGMA Atlas dataset.

To date, our most significant efforts towards global mapping of genotype-phenotype relationship are the accurate annotation of metabolic pathways in bacteria using a curated database of experimentally characterized proteins, including those proteins re-annotated using ENIGMA's mutant fitness data. These tools, collectively referred to as GapMind, only take as input the genome sequence of a bacterium, and return the most likely pathways by which the bacterium synthesizes amino acids or catabolizes a carbon source (Figure 5) [6,35]. GapMind can rapidly identify existing holes in our knowledge, in particular in cases where it is known that a bacterium contains a particular pathway (based on our high-throughput

phenotyping data) and yet a high-confidence metabolic pathway cannot yet be identified in the genome. Moving forward, we are expanding the number of metabolic pathways including in GapMind, and we are using this software tool as a mechanism to propagate our new protein function discoveries to all sequenced ENIGMA genomes. In addition, we are beginning to explore additional approaches to infer phenotype-genotype relationships on a global scale, including genome-wide association studies (GWAS) and tools based on machine learning and machine-learning augmented metabolic modeling.



catabolism of small carbon sources in Pseudomonas fluorescens FW300-N2E2

Pathways are sorted by completeness. Sort by name instead.

Pathway	Steps
valine	**livF**, **livG**, **livJ**, **livH**, **livM**, **bkdA**, **bkdB**, **bkdC**, **lpd**, **acdH**, **ech**, **bch**, **mmsB**, **mmsA**, **prpC**, **acnD**, **prpF**, **acn**, **prpB**
isoleucine	**livF**, **livG**, **livJ**, **livH**, **livM**, **bkdA**, **bkdB**, **bkdC**, **lpd**, **acdH**, **ech**, **ivdG**, **fadA**, **prpC**, **acnD**, **prpF**, **acn**, **prpB**
leucine	**livF**, **livG**, **livJ**, **livH**, **livM**, **ilvE**, **bkdA**, **bkdB**, **bkdC**, **lpd**, **liuA**, **liuB**, **liuD**, **liuC**, **liuE**, **atoA**, **atoD**, **atoB**
phenylalanine	**livF**, **livG**, **livH**, **livM**, **livJ**, **PAH**, **PCBD**, **QDPR**, **HPD**, **hmgA**, **maiA**, **fahA**, **atoA**, **atoD**, **atoB**
lysine	**argT**, **hisM**, **hisQ**, **hisP**, **davB**, **davA**, **davT**, **davD**, **gcdG**, **gcdH**, **ech**, **fadB**, **atoB**
threonine	**braC**, **braD**, **braE**, **braF**, **braG**, **ltaE**, **adh**, **ackA**, **pta**, **gcvP**, **gcvT**, **gcvH**, **lpd**
arginine	**artJ**, **artM**, **artP**, **artQ**, **arcA**, **arcB**, **arcC**, **aruF**, **aruG**, **astC**, **astD**, **astE**
citrulline	**AO353_03055**, **AO353_03050**, **AO353_03045**, **AO353_03040**, **arcB**, **arcC**, **aruF**, **aruG**, **astC**, **astD**, **astE**
myoinositol	**PS417_11885**, **PS417_11890**, **PS417_11895**, **iolG**, **iolE**, **iolD**, **iolB**, **iolC**, **iolJ**, **mmsA**, **tpi**
4-hydroxybenzoate	**pcaK**, **pobA**, **pcaH**, **pcaG**, **pcaB**, **pcaC**, **pcaD**, **catI**, **catJ**, **pcaF**
⋮	
deoxyribonate	deoxyribonate-transport, deoxyribonate-dehyd, ketodeoxyribonate-cleavage, **garK**, **atoA**, **atoD**, **atoB**
tryptophan	**aroP**, tnaA
deoxyinosine	nupC, **deoD**, deoB, deoC, **adh**, **ackA**, **pta**
thymidine	nupG, **deoA**, deoB, deoC, **adh**, **ackA**, **pta**
D-serine	cycA, dsdA
xylitol	PLT5, xdhA, **xylB**
rhamnose	rhaT, LRA1, LRA2, **LRA3**, LRA4, **aldA**
NAG	nagEcba, nagA, nagB
fucose	**HSERO_RS05250**, HSERO_RS05255, HSERO_RS05260, fucU, fucI, fucK, fucA, **tpi**, **aldA**
phenylacetate	ppa, paaK, paaA, paaB, paaC, paaE, **paaG**, paaZ1, paaZ2, **paaJ1**, **paaF**, **paaH**, **paaJ2**

Confidence: **high confidence** medium confidence low confidence
transporter – transporters and PTS systems are shaded because predicting their specificity is particularly challenging.

Figure 5. Screenshot of the GapMind for carbon sources tool. For the ORNL FRC bacterium *Pseudomonas fluorescens* FW300-N2E2, high-scoring carbon catabolic pathways are illustrated at the top, while low-scoring pathways are at the bottom. The genes in green are similar to experimentally characterized genes involved in the catabolism of the indicated carbon source.

Recent applications of ENIGMA characterization pipeline to *Rhodanobacter*

To highlight our suite of bacterial characterization tools, here we briefly illustrate recent studies of *Rhodanobacter*, a genus of bacteria that dominates the most contaminated (high concentrations of metals, high nitrate, and low pH) groundwater wells at the ORNL FRC [36]. Despite the availability of multiple isolates from the field site, we still do not know the molecular mechanism(s) through which these bacteria thrive in extreme environments. To address this question from multiple angles, ENIGMA has leveraged

our maturing bacterial characterization pipeline. First, we performed genome sequencing of multiple ORNL FRC *Rhodanobacter* strains and through a comparison of these genomes we found that genes related to metal resistance were both overrepresented and associated with horizontal gene transfer, suggesting that the acquisitions of these genes were important evolutionary events enabling the adaptation of these bacteria to the ORNL FRC environment [37]. Second, to validate the importance of specific genes, we developed a targeted, markerless gene deletion system for *Rhodanobacter* and used this tool to delete *narG* (encoding nitrate reductase) from *Rhodanobacter* strains FW104-R3 and FW104-R5. Next, to scale genetics in a high-throughput manner, we generated an RB-TnSeq library in *Rhodanobacter* FW104-10B01, and used this library to globally measure gene importance across a large number of different metal stress conditions. From these data, we were able to determine the metal specificity of different metal efflux systems in the FW104-10B01 genome (Figure 6). To enable single-gene follow-up studies, we archived thousands of individual RB-TnSeq FW104-10B01 mutants and are currently mapping them using a smart-pooling strategy and BarSeq. Current work is focused on the development of additional tools for *Rhodanobacter* including CRISPRi, genetic libraries in the R12 isolate which is being used in environmentally-inspired SynCom experiments, and dissection of gene regulatory networks using comparative genomics and DAP-seq.

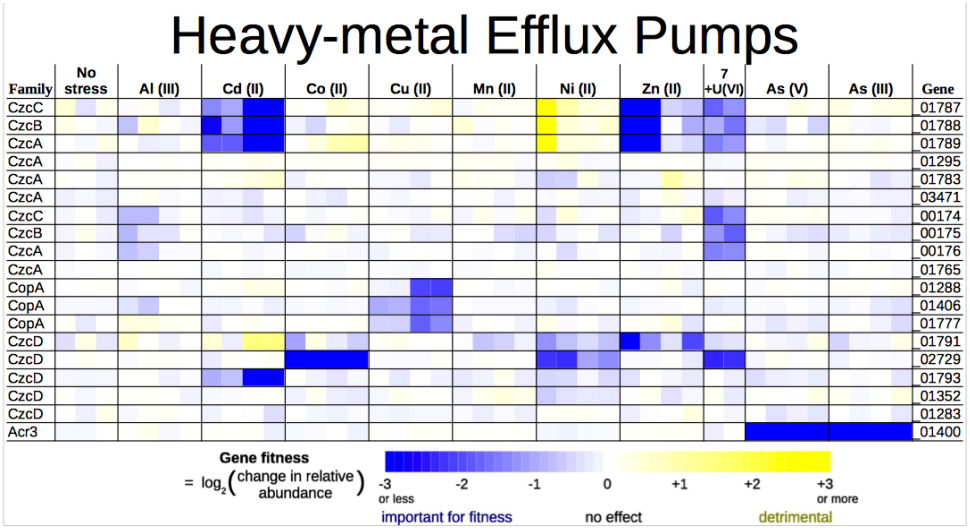


Figure 6. Phenotypes for metal efflux genes as revealed by RB-TnSeq. Heatmap of gene fitness values for predicted heavy metal resistance genes in *Rhodanobacter* FW104-10B01. Some systems are important for growth in the presence of a single metal (CopA and copper), while others are necessary for growth in the presence of multiple metals (01787-9 and Zn, U, and Cd).

Summary and future outlook

By discovering the genes, pathways, and metabolites that mediate bacterial metabolisms, stress responses, and community interactions, the ENIGMA isolate characterization pipeline provides a critical link between ORNL FRC field data and the advanced systems-biology/modeling efforts to dissect field phenomena in tractable experimental systems. In addition, our experimental and computational tools are being increasingly used by ourselves (and external researchers) to accelerate the rate of protein function discovery in diverse microorganisms. Current and future efforts are focused on the development of new scalable genetic and metablomic technologies to greatly advance microbial characterization and computation approaches to accurately predict phenotypes from genome sequence alone. However, we

note that even a massive scaling of our current approaches can only tackle a small fraction of culturable bacteria, and that a concerted, multi-team effort to characterize microbes (using diverse approaches) is necessary to advance our understanding of microorganisms to a resolution sufficient for truly predictive microbial ecology. Furthermore, we will continue to analyze our laboratory-based characterization results in the context of field measurements (metagenomic and chemical) for understanding the ecological niche and activity of these organisms. In particular, mapping field data to our isolates can suggest specific microorganisms, genetic systems, and conditions that are important for investigation in the laboratory. This data in turn can be mapped back to in situ environmental measurements, including metagenomics and chemical data, thereby providing a route to using laboratory data to interpret field data more effectively.

Cited Literature

1. Loman, N. J., & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nature Reviews. Microbiology*, *13*(12), 787–794. doi:10.1038/nrmicro3565
2. Liu, H., & Deutschbauer, A. M. (2018). Rapidly moving new bacteria to model-organism status. *Current Opinion in Biotechnology*, *51*, 116–122. doi:10.1016/j.copbio.2017.12.006
3. Carlson, H. K., Price, M. N., Callaghan, M., Aaring, A., Chakraborty, R., Liu, H., … Deutschbauer, A. M. (2019). The selective pressures on the microbial community in a metal-contaminated aquifer. *The ISME Journal*, *13*(4), 937–949. doi:10.1038/s41396-018-0328-1
4. Kosina, S. M., Greiner, A. M., Lau, R. K., Jenkins, S., Baran, R., Bowen, B. P., & Northen, T. R. (2018). Web of microbes (WoM): a curated microbial exometabolomics database for linking chemistry and microbes. *BMC Microbiology*, *18*(1), 115. doi:10.1186/s12866-018-1256-y
5. Price, M. N., Deutschbauer, A. M., & Arkin, A. P. (2021). Four families of folate-independent methionine synthases. *PLoS Genetics*, *17*(2), e1009342. doi:10.1371/journal.pgen.1009342
6. Price, M. N., Deutschbauer, A. M., & Arkin, A. P. (2020). Gapmind: automated annotation of amino acid biosynthesis. *mSystems*, *5*(3). doi:10.1128/mSystems.00291-20
7. Kothari, A., Soneja, D., Tang, A., Carlson, H. K., Deutschbauer, A. M., & Mukhopadhyay, A. (2019). Native Plasmid-Encoded Mercury Resistance Genes Are Functional and Demonstrate Natural Transformation in Environmental Bacterial Isolates. *mSystems*, *4*(6). doi:10.1128/mSystems.00588-19
8. Lui, L. M., Majumder, E. L.-W., Smith, H. J., Carlson, H. K., von Netzer, F., Fields, M. W., … Arkin, A. P. (2021). Mechanism across scales: A holistic modeling framework integrating laboratory and field studies for microbial ecology. *Frontiers in microbiology*, *12*, 642422. doi:10.3389/fmicb.2021.642422
9. Neal-McKinney, J. M., Liu, K. C., Lock, C. M., Wu, W.-H., & Hu, J. (2021). Comparison of MiSeq, MinION, and hybrid genome sequencing for analysis of Campylobacter jejuni. *Scientific Reports*, *11*(1), 5676. doi:10.1038/s41598-021-84956-6
10. Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., … Yu, D. (2018). Kbase: the united states department of energy systems biology knowledgebase. *Nature Biotechnology*, *36*(7), 566–569. doi:10.1038/nbt.4163
11. Bochner, B. R., Gadzinski, P., & Panomitros, E. (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Research*, *11*(7), 1246–1255. doi:10.1101/gr.186501
12. Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., … Stevens, R. (2020). The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Research*, *48*(D1), D606–D612. doi:10.1093/nar/gkz943
13. Barberán, A., Caceres Velazquez, H., Jones, S., & Fierer, N. (2017). Hiding in plain sight: mining bacterial species records for phenotypic trait information. *mSphere*, *2*(4). doi:10.1128/mSphere.00237-17
14. Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., … Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, *557*(7706), 503–509. doi:10.1038/s41586-018-0124-0
15. Carlson, H. K., Lui, L. M., Price, M. N., Kazakov, A. E., Carr, A. V., Kuehl, J. V., … Deutschbauer, A. M. (2020). Selective carbon sources influence the end products of microbial nitrate respiration.

The ISME Journal, *14*(8), 2034–2045. doi:10.1038/s41396-020-0666-7

16. Jenkins, S., Swenson, T. L., Lau, R., Rocha, A. M., Aaring, A., Hazen, T. C., … Northen, T. R. (2017). Construction of viable soil defined media using quantitative metabolomics analysis of soil metabolites. *Frontiers in microbiology*, *8*, 2618. doi:10.3389/fmicb.2017.02618
17. Hunt, K. A., Forbes, J., Taub, F., Elliott, N., Hardwicke, J., Petersen, R., … Stahl, D. A. (2021). An automated multiplexed turbidometric and data collection system for measuring growth kinetics of anaerobes dependent on gaseous substrates. *Journal of Microbiological Methods*, 106294. doi:10.1016/j.mimet.2021.106294
18. Silva, L. P., & Northen, T. R. (2015). Exometabolomics and MSI: deconstructing how cells interact to transform their small molecule environment. *Current Opinion in Biotechnology*, *34*, 209–216. doi:10.1016/j.copbio.2015.03.015
19. Erbilgin, O., Bowen, B. P., Kosina, S. M., Jenkins, S., Lau, R. K., & Northen, T. R. (2017). Dynamic substrate preferences predict metabolic properties of a simple microbial consortium. *BMC Bioinformatics*, *18*(1), 57. doi:10.1186/s12859-017-1478-2
20. Keller, K. L., Bender, K. S., & Wall, J. D. (2009). Development of a markerless genetic exchange system for Desulfovibrio vulgaris Hildenborough and its use in generating a strain with increased transformation efficiency. *Applied and Environmental Microbiology*, *75*(24), 7682–7691. doi:10.1128/AEM.01839-09
21. Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., … Deutschbauer, A. (2015). Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *mBio*, *6*(3), e00306-15. doi:10.1128/mBio.00306-15
22. Liu, H., Price, M. N., Waters, R. J., Ray, J., Carlson, H. K., Lamson, J. S., … Deutschbauer, A. M. (2018). Magic pools: parallel assessment of transposon delivery vectors in bacteria. *mSystems*, *3*(1). doi:10.1128/mSystems.00143-17
23. Mutalik, V. K., Novichkov, P. S., Price, M. N., Owens, T. K., Callaghan, M., Carim, S., … Arkin, A. P. (2019). Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria. *Nature Communications*, *10*(1), 308. doi:10.1038/s41467-018-08177-8
24. Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, *152*(5), 1173–1183. doi:10.1016/j.cell.2013.02.022
25. Mutalik, V. K., Adler, B. A., Rishi, H. S., Piya, D., Zhong, C., Koskella, B., … Arkin, A. P. (2020). High-throughput mapping of the phage resistance landscape in E. coli. *PLoS Biology*, *18*(10), e3000877. doi:10.1371/journal.pbio.3000877
26. Garber, M. E., Rajeev, L., Kazakov, A. E., Trinh, J., Masuno, D., Thompson, M. G., … Mukhopadhyay, A. (2018). Multiple signaling systems target a core set of transition metal homeostasis genes using similar binding motifs. *Molecular Microbiology*, *107*(6), 704–717. doi:10.1111/mmi.13909
27. Trotter, V. V., Shatsky, M., Price, M. N., Juba, T. R., Zane, G. M., De Leon, K. P., … Butland, G. P. (2021). Large-scale Genetic Characterization of a Model Sulfate Reducing Bacterium. *BioRxiv*. doi:10.1101/2021.01.13.426591
28. Wall, J. D., Zane, G. M., Juba, T. R., Kuehl, J. V., Ray, J., Chhabra, S. R., … Deutschbauer, A. M. (2021). Deletion Mutants, Archived Transposon Library, and Tagged Protein Constructs of the Model Sulfate-Reducing Bacterium Desulfovibrio vulgaris Hildenborough. *Microbiology*

Resource Announcements, *10*(11). doi:10.1128/MRA.00072-21
29. Thorgersen, M. P., Xue, J., Majumder, E. L. W., Trotter, V. V., Ge, X., Poole, F. L., … Adams, M. W. W. (2021). Deciphering Microbial Metal Toxicity Responses via Random Bar Code Transposon Site Sequencing and Activity-Based Metabolomics. *Applied and Environmental Microbiology*, *87*(21), e0103721. doi:10.1128/AEM.01037-21
30. Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, *28*(9), 977–982. doi:10.1038/nbt.1672
31. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*, 160018. doi:10.1038/sdata.2016.18
32. Majumder, E. L.-W., Billings, E. M., Benton, H. P., Martin, R. L., Palermo, A., Guijas, C., … Siuzdak, G. (2021). Cognitive analysis of metabolomics data for systems biology. *Nature Protocols*, *16*(3), 1376–1418. doi:10.1038/s41596-020-00455-4
33. Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 — approximately maximum-likelihood trees for large alignments. *Plos One*, *5*(3), e9490. doi:10.1371/journal.pone.0009490
34. Price, M. N., & Arkin, A. P. (2017). PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems*, *2*(4). doi:10.1128/mSystems.00039-17
35. Price, M. N., Deutschbauer, A. M., & Arkin, A. P. (2021). GapMind for Carbon Sources: Automated annotations of catabolic pathways. *BioRxiv*. doi:10.1101/2021.11.02.466981
36. Green, S. J., Prakash, O., Jasrotia, P., Overholt, W. A., Cardenas, E., Hubbard, D., … Kostka, J. E. (2012). Denitrifying bacteria from the genus Rhodanobacter dominate bacterial communities in the highly contaminated subsurface of a nuclear legacy waste site. *Applied and Environmental Microbiology*, *78*(4), 1039–1047. doi:10.1128/AEM.06435-11
37. Peng, M., Wang, D., Lui, L. M., Nielsen, T., Tian, R., Kempher, M. L., … Zhou, J. (2022). Genomic Features and Pervasive Negative Selection in Rhodanobacter Strains Isolated from Nitrate and Heavy Metal Contaminated Aquifer. *Microbiology spectrum*, e0259121. doi:10.1128/spectrum.02591-21