Pacific Northwest
NATIONAL LABORATORY

PNNL- 36738

Developing a pipeline to expand the genetic code of diverse bacteria for microbial engineering

September 2024

Joshua Elmore
Elise van Fossen

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL- 36738

Developing a pipeline to expand the genetic code of diverse bacteria for microbial engineering

September 2024

Joshua Elmore
Elise van Fossen

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Microbial biotechnologies are key to addressing grand challenges to promote human health, reverse carbon emissions, recycle mixed plastic waste, remediate contaminated soils, and achieve sustainable economies. Synthetic biology has enabled design of diverse microbes and their proteins for useful purposes, but the narrowness of the natural genetic code limits functional diversity (e.g., biosynthesis) of engineered microbes. The natural genetic code defines the fundamental rules of translating genetic information into proteins comprised of 22 ‘canonical’ amino acids. However, using a technique called genetic code expansion (GCE), the chemical properties and therefore functions of proteins can be transformed by incorporation of one or more of ~200 chemically diverse ‘non-canonical’ amino acids. The effective application of genetic code expansion in diverse microbes has the potential to revolutionize biotechnology. However, despite over 50 years of research and its transformative potential, the application of genetic code expansion has been limited to a handful of bacterial species. In this project, we will perform three tasks to both overcome the barriers that prevent wide spread adoption of GCE as molecular tool and demonstrate its potential for biotechnological applications. Specifically, we will (1) develop a genetic engineering methodology that will enable use of GCE in a broad range of bacterial hosts, and (2) use high-throughput functional genomics methods to identify physiological responses to both genetic code expansion and exposure to non-canonical amino acids in three different bacteria

Summary

Microbial biotechnologies will be central to addressing grand challenges to promote human health, reverse carbon emissions, recycle mixed plastic waste, remediate contaminated soils, and achieve sustainable economies. Synthetic biology has enabled researchers to both redesign diverse microbes for useful purposes, but is currently bound to following the natural genetic code. The genetic code defines the fundamental rules of translating genetic information into proteins comprised of 22 ‘canonical’ amino acids. As the complexity of desired engineered bacterial functions increases, limiting proteins that drive the functions to just the chemistries enabled by these 22 canonical amino acids severely restricts the potential functions of engineered organisms. Genetic code expansion is a recently developed approach that expands the ruleset to enable proteins that can include one or more of approximately 200 ‘non-canonical’ amino acids. *The effective application of genetic code expansion in diverse microbes has the potential to revolutionize essentially all fields that employ biotechnology - whether for application or for fundamental biology*.

Current barriers in wide-spread adoption of genetic code expansion include: (1) challenges in efficiently transferring the requisite biochemical machinery into both diverse model and biotechnologically-relevant microbes, (2) optimizing performance of that machinery in the new host, and (3) large gaps in fundamental knowledge of how genetic code expansion and the presence of non-canonical amino acids impact cellular physiology. These technical limitations and knowledge gaps have limited its application in diverse organisms to only a few niche use cases. We will combine existing and develop new high-throughput functional genomics and synthetic biology tools and approaches to nearly double the number of bacteria where genetic code expansion has been applied (to 17 total bacteria) and evaluate the impact of expanding the code upon the physiology of three related organisms with distinct application spaces (e.g., human health, biofuels, etc.). We expect that this work will uncover generalizable design rules for expanding the genetic code in new microbes, as well as uncovering fundamental knowledge of which cellular functions are negatively impacted by this expansion.

Acknowledgments

This research was supported by the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

GCE – genetic code expansion
GFP – green fluorescent protein
trGFP – truncated green fluorescence protein
SAGE – Serine recombinase-Assisted Genome Engineering
ncAA – non-canonical amino acid
tRNA – transfer RNA
RB-TnSeq – randomly barcoded transposon mutant sequencing
AzF – L-*para*-azidophenylalanine
sfGFP – super folder green fluorescent protein
Phe – L-phenylalanine
Tyr – L-tyrosine
3-Chloro-Y – 3-chlorotyrosine
4-Chloro-F – 4-chlorophenylalanine
4-Azido-F - L-*para*-azidophenylalanine
Ab – antibiotic
AprR – apramycin resistance marker
aaRS – aminoacyl-tRNA synthetase

1.0 Introduction

Microbial biotechnologies are key to addressing grand challenges to promote human health, reverse carbon emissions, recycle mixed plastic waste, remediate soils, and achieve sustainable economies (Figure 1). Synthetic biology has enabled design of diverse microbes and their proteins for useful purposes, but the narrowness of the natural genetic code limits functional diversity (e.g., biosynthesis) of engineered microbes. The natural genetic code defines the fundamental rules of translating genetic information into proteins comprised of 22 'canonical' amino acids. However, using a technique called genetic code expansion (GCE), the chemical properties and therefore functions of proteins can be transformed by incorporation of one or more of ~200 chemically diverse 'non-canonical' amino acids (Figure 2). *The effective application of genetic code expansion in diverse microbes has the potential to revolutionize biotechnology. However, despite over 50 years of research and its transformative potential, the application of genetic code expansion has been limited to a small group of bacterial species.* Herein, we will develop a pipeline to enable deployment of GCE in essentially any bacterial species and will identify general bacterial physiology impacts of GCE.

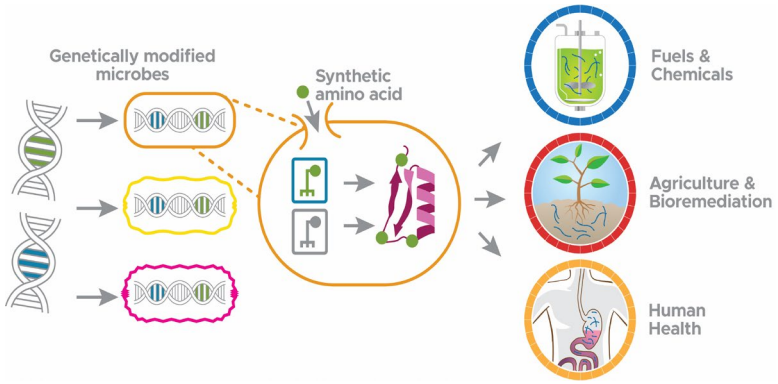

Figure 1. This project aims to remove barriers that prevent the use of genetic code expansion (GCE) in non-model organisms that can perform functions to help address societally important challenges.

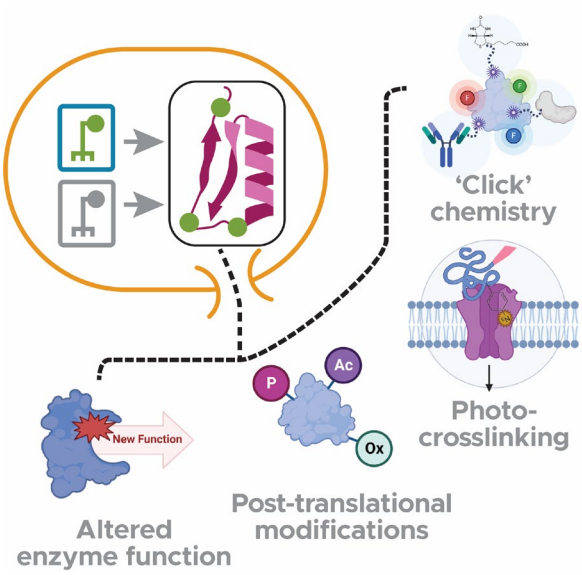

Figure 2. Examples of the types of protein modifications that are possible when using genetic code expansion.

Current barriers to wide-spread adoption of GCE by the research community include: (1) challenges in efficiently transferring the requisite biochemical machinery into diverse model and biotechnologically relevant microbes, (2) optimizing performance of that machinery in the new host, and (3) large gaps in fundamental understanding of the physiological impacts of genetic code expansion and the presence of non-canonical amino acids. These technical limitations and knowledge gaps have limited the use of GCE outside of the model organism *Escherichia coli* to a few niche cases (Stork, Squyres et al. 2021). We will combine our synthetic biology approaches developed at PNNL to enable rapid and facile application of GCE in a collection of 8 bacteria that represent distinct research areas (e.g., biofuels, human health, etc.), and this will be a huge step forward for the field by nearly doubling the number of bacteria where GCE has been applied. Existing functional genomics methods will be applied and modified to evaluate the impact of GCE upon the

physiology of three related organisms. We expect this new knowledge to enable generalizable design rules for performing GCE with minimal impact on the host, as well as uncovering fundamental knowledge of cellular functions affected by this expansion.

2.0 Results

Our goal was to develop a host-agnostic Genetic Code Expansion (GCE) platform that addresses the limitations of most current GCE systems, specifically their inadequacy in complex environments due to factors such as the use of unstable plasmids that necessitate antibiotics for selection and compromise host fitness. For this, the platform needed to be robust, capable of incorporating multiple non-canonical amino acids into one or more proteins simultaneously, suitable for genetic library-based optimization strategies, genetically stable without the need for selection, free from the requirement of exogenous inducer chemicals, and easily transferable among phylogenetically diverse bacteria.

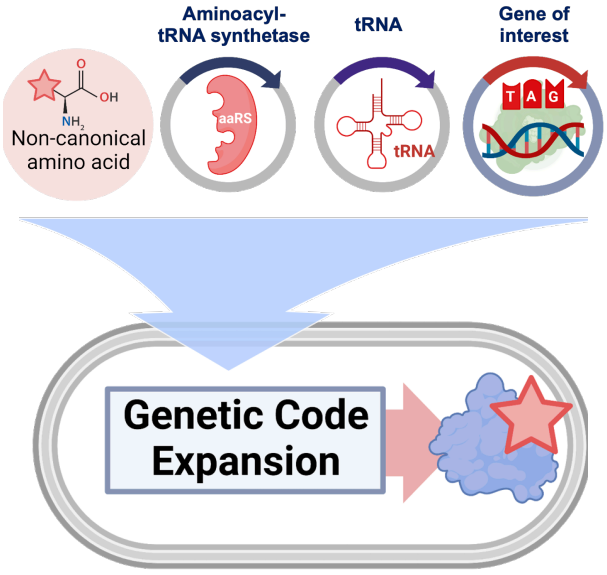

Figure 3. Genetic code expansion requires the non-canonical amino and a minimum of three different genetic components.

We employed serine recombinase-assisted genome engineering (SAGE) (Elmore, Dexter et al. 2023), a host-agnostic genetic tool, to integrate the entire GCE machinery into the host organisms. Implementing genetic code expansion typically requires at least three genetic components: an orthogonal aminoacyl-tRNA synthetase (o-aaRS), a corresponding orthogonal tRNA (o-tRNA), and a target protein with a stop codon at the site of non-canonical amino acid (ncAA) incorporation (Figure 3). SAGE uses transiently expressed serine recombinases to recombine attP sequences on non-replicating cargo plasmids with a specific attB sequence previously integrated into the host chromosome. A newer variant of this system incorporates distinct antibiotic selection markers on each cargo plasmid, enabling the integration of each GCE component in a single step (Figure 4).

For the initial development of a SAGE-based GCE toolkit, we focused on incorporating the non-canonical amino acid (ncAA) *para*-azidophenylalanine (AzF) into proteins. The proof-of-concept GCE system utilized is based on the well-studied *Methanococcus jannaschii* tyrosyl-tRNA synthetase/tRNA-Tyr(CUA) (*Mj* TyrRS/tRNA) system. We opted to incorporate AzF using the pCNFRS system due to its high efficiency in incorporating AzF into proteins across various organisms, its low incidence of mis-incorporation of canonical amino acids, and its ability to incorporate at least 18 other ncAAs. This versatility allows for the use of amino acids with a wide array of chemical functionalities. AzF contains an azide functional group, which is

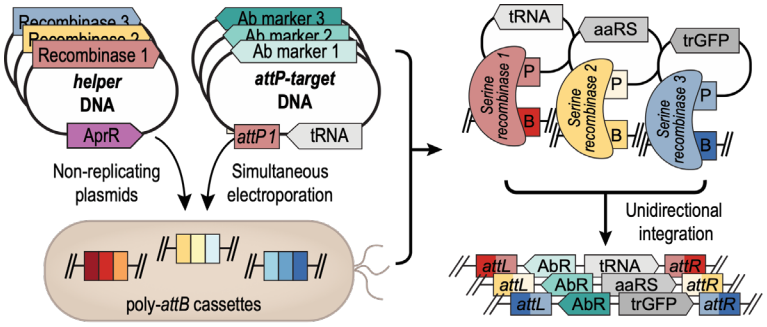

Figure 4. Multiplexed SAGE enables facile incorporation multiple GCE machinery components with the need to recycle selection markers.

widely used in protein labeling through bioorthogonal reactions, crosslinking, and structural studies.

We selected *Pseudomonas putida* KT2440 as the initial host organism for developing the SAGE-based GCE system. This choice was made because *P. putida* is a non-model host increasingly utilized for both fundamental and industrial applications (Martínez-García and de Lorenzo, 2024) and is commonly engineered using SAGE. Furthermore, *P. putida* is closely related to many plant-growth-promoting bacteria, making the development of this system in *P. putida* likely to enable immediate applications in bacteria used across diverse environments, from soil to bioreactors.

To maintain the expression of GCE machinery at physiologically relevant levels, we employed a strategy using native regulatory elements to control both the tRNA and synthetase expression. This approach also retains any sequences in the unprocessed tRNA transcript that may be recognized by the native tRNA processing machinery. We identified the most abundant codon in Pseudomonas putida KT2440 (CUG – Leucine) and placed the orthogonal tRNA and aminoacyl-tRNA synthetase under the control of the native tRNA-Leu(CUG) and Leu-tRNA synthetase promoters, respectively. The terminators for these native components were also utilized for their respective GCE machinery.

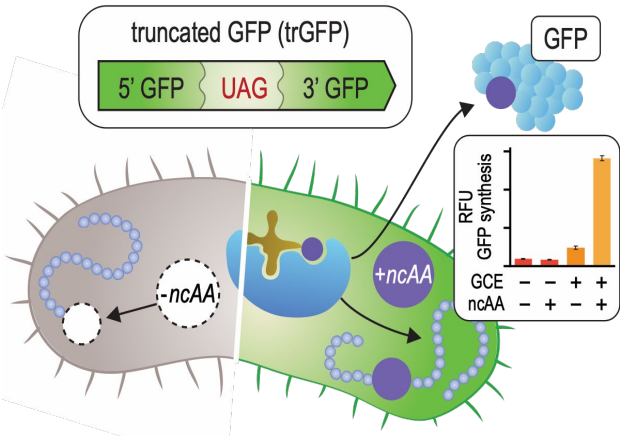

Figure 5. Graphical description of truncated GFP assay. Cell fluorescence is directly correlated with production of full length ncAA-containing GFP.

The efficacy of our GCE system was evaluated using a truncated GFP (trGFP) assay (Figure 5). We integrated a copy of superfolder GFP (sfGFP) containing an amber stop codon at the 150th codon into the genome using SAGE and assessed the expression of full-length GFP in the presence and absence of both AzF and the GCE machinery. This specific codon was chosen because it does not affect GFP stability and truncation at this location does not result in the synthesis of a fluorescent peptide fragment. To ensure robust expression of sfGFP, we utilized the JEa3 promoter, known for its high transcriptional activity (Elmore, Peabody et al., 2022).

As previously reported, we observed that the pCNFRS system in Pseudomonas putida enabled efficient incorporation of AzF, as indicated by the expression of full-length sfGFP only in the presence of the ncAA (Ozer, Yaniv et al., 2021). There were low but detectable levels of full-length sfGFP produced in the absence of ncAAs, suggesting mischarging of the tRNA-Tyr(CUA) with canonical amino acids. This observation aligns with previous findings that polyspecific MjTyrRS-derived aminoacyl-tRNA synthetase variants, like the one used in our system, recognize canonical amino acids such as phenylalanine and tyrosine in the absence of ncAAs. However, in the presence of ncAA, the system exhibited high fidelity, with minimal incorporation of canonical amino acids. The ratio of full-length protein with ncAA incorporated versus protein with unintended incorporation of other amino acids is referred to as "relative fidelity" and is a critical measure of protein production and accuracy (Porter and Mehl, 2018).

While our SAGE-based GCE system functioned as expected, the efficiency was low, resulting in only a moderate amount of full-length GFP being produced. To enhance performance, we incorporated an orthogonal elongation factor Tu (Ef-Tu) protein into our design (Deley Cox, Cole et al., 2019). The orthogonal Ef-Tu assists in transferring the AzF-charged aminoacyl-tRNA into the ribosome, thereby enabling more efficient use of the charged tRNA for protein synthesis (Figure 6). Additionally, we found that the efficiency of full-length GFP production was significantly

dependent on the presence of an amino acid mixture in the growth medium. This finding is suboptimal for many intended GCE applications, such as in bioreactors or soils, where a consistent and abundant supply of amino acids may not be readily available. (e.g., bioreactors, soils, etc.) are not likely to have a consistent and abundant supply of amino acids.

We hypothesized that the presence of additional amino acids, particularly aromatic amino acids, enables individual cells to produce more full-length GFP due to their influence on the expression of enzymes that import AzF into the cell. The presumption is that in the presence of aromatic amino acids, the expression of such transporters is significantly upregulated compared to when these amino acids are absent. If this hypothesis is correct, then the constitutive expression of these transporters in the absence of additional amino acids might enhance the effectiveness of GCE and potentially improve performance even when ncAAs are at lower concentrations in the environment. However, to test this, we first needed to identify potential aromatic amino acid transporters.

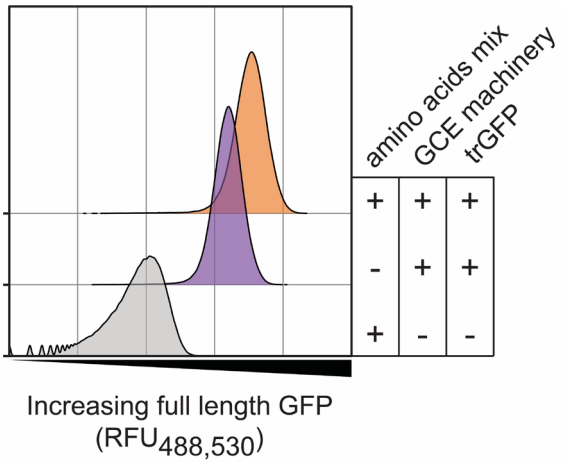

Figure 6. Flow cytometry data showing the relative fluorescence in the bacterial population. The further the curve is to the right, the more full length GFP is being produced in each cell.

Many ncAAs contain functional groups commonly associated with toxicity, and aromatic compounds, in general, can be toxic at high concentrations. Bacteria protect themselves from excessive concentrations of native aromatic compounds and xenobiotic compounds, such as ncAAs, through transport mechanisms. If these transporters are constitutively expressed, they may aid in importing ncAAs into the cell. We hypothesized that disrupting the expression of these transporters would cause growth defects in the presence of high concentrations of xenobiotic compounds without affecting growth in their absence. Additionally, transporters involved in the bacteria's growth on aromatic amino acids as the sole carbon sources might also facilitate the import of AzF or other aromatic ncAAs.

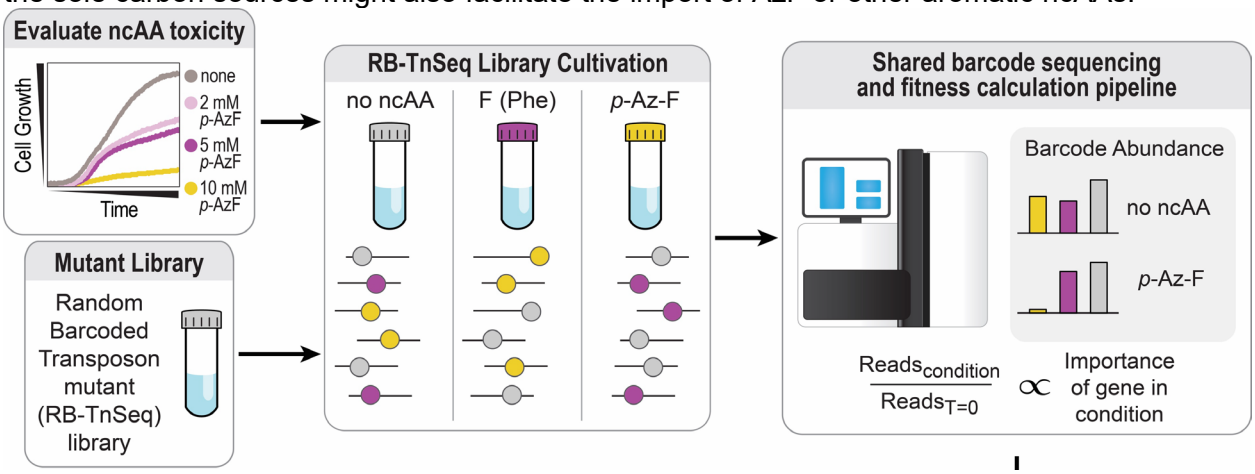

Figure 7. Graphical diagram of RB-TnSeq workflow. The top left contains a series of *Pseudomonas putida* growth curves in media containing citrate as the growth substrate and varying levels of putatively non-metabolized AzF as a stressor.

Organism	Gene Locus Tag	Annotation (computationally predicted)	Predicted Transport Directionality	Method Used to ID candidate transporters
Pseudomonas facilor TBS28	PFR28_00536 to PFR28_00541	high-affinity branched-chain amino acid ABC transporter	import	RB-TnSeq with cultivation in toxic levels of ncAAs
	PFR28_00629 to PFR28_00632	Probable efflux pump protein complex	import	
	PFR28_00737	drug/metabolite transporter	export	
	PFR28_03621	Threonine efflux protein	export	
	PFR28_03686 to PFR28_03688	PqiABC family transporter	import	
	PFR28_03826	putative inner membrane transporter YedA	export	
Pseudomonas putida KT2440	PP_0505 to PP_0508	high-affinity branched-chain amino acid ABC transporter	export	
	PP_0589	Bcr/CflA family multidrug resistance transporter	export	
	PP_1112	major facilitator superfamily transporter	import	
	PP_1137 to PP_1142	IlvKHMGF branched chain amino acid ABC transporter	import	
	PP_1384 to PP_1386	multidrug resistance transporter	export	
	PP_1778 to PP_1779	Lipopolysaccharide ABC export system	import	
	PP_4604	putative permease of the drug/metabolite transporter (DMT) superfamily	import	
	PP_4767	phenylalanine/threonine exporter	export	

Table 1. List of putative transporter-encoding genes whose negative or positive fitness scores in conditions containing toxic levels of ncAAs suggest a role in ncAA transport.

We employed random barcode transposon mutant sequencing (RB-TnSeq) (Wetmore, Price et al., 2015) to identify transporters that may facilitate the import of AzF and other aromatic ncAAs. RB-TnSeq is a powerful tool that enables phenotypic screening of all non-essential genes in a single experiment. For this purpose, we used two previously generated transposon mutant libraries derived from *Pseudomonas putida* and *Pseudomonas facilor*. These libraries contain over 300,000 individually barcoded transposon insertion mutants.

Using high-throughput barcode sequencing, we tracked the relative abundance of transposon insertion mutants to identify those whose abundance decreases following growth under specific conditions. A fitness value was assigned to each gene based on changes in the relative abundance of gene disruption mutants. A negative fitness value indicates that mutants with transposon insertions in that gene have decreased abundance in the population during growth under the tested condition, suggesting that the protein encoded by that gene is important for growth under those conditions. Conversely, a positive fitness value indicates that the mutant

becomes more abundant, suggesting that removing the function of this protein offers a growth advantage under the tested condition.

We first screened *P. putida* and *P. facilor* for resistance to high concentrations of aromatic ncAAs in mineral medium containing citrate and ammonium as the primary growth substrates. In general, each strain was able to grow in media containing up to 10 mM of each ncAA, but a substantial growth defect was observed in these conditions. Representative data for such assays is shown in **Figure 7**. We performed RB-TnSeq assays by cultivating the mutant library in media containing citrate alone and citrate with 5 or 10 mM of each ncAA. We also performed similar assays using L-tyrosine or L-phenylalanine as either the sole carbon source or in media containing both citrate and the canonical aromatic amino acids. Transporter encoding genes with positive fitness values in the presence of toxic levels of ncAAs encode potential ncAA import enzymes and those with negative fitness values encode potential ncAA export enzymes. Fitness values for a subset of genes in these experiments is shown in **Figure 8** and select transporters of interest are listed in **Table 1**. We often observed that transporters that appeared to be involved in canonical aromatic amino acid transport (those with negative fitness during growth using the amino acid as the sole carbon source) did not have either a positive or negative fitness value in samples with toxic levels of ncAAs. This suggests that the transporters that actively import those canonical amino acids that may not be responsible for ncAA transport. However, this does not rule out that the canonical aromatic amino acids induce expression of the enzymes that transport the ncAAs.

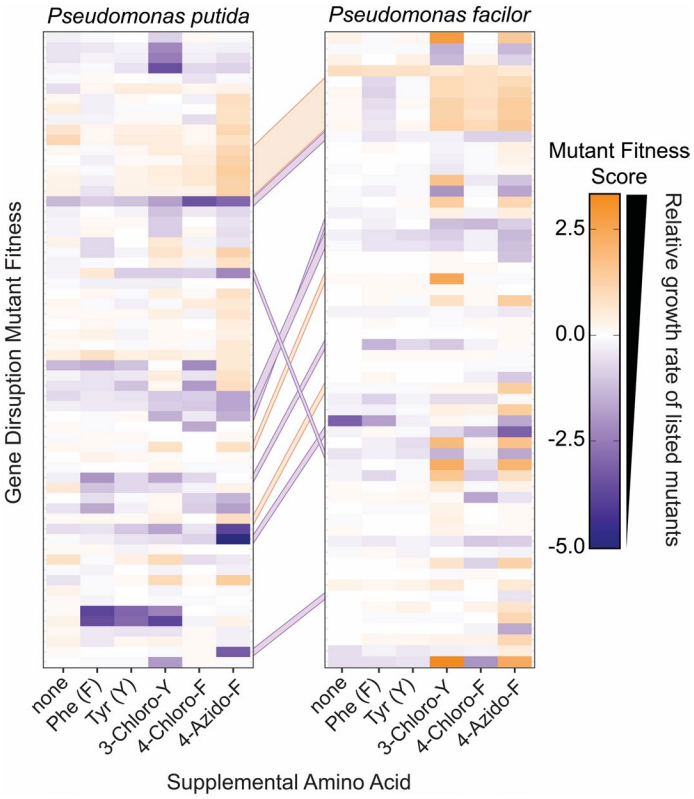

Figure 8. Heat map of fitness values for a select set of genes that have condition-specific fitness values in the presence of either canonical amino acids or high concentrations of AzF or chlorinated aromatic amino acids. Colored bars that span between the *P. facilor* and *P. putida* heat maps represent conserved proteins that have similar fitness patterns in both organisms.

Our next goal was to constitutively express the identified transporters and evaluate their impact on GCE efficiency. Specifically, we aimed to determine whether full-length GFP production increases, especially when ncAA concentrations in the media are significantly reduced. To achieve this, we incorporated four sets of the putative transporter-encoding genes under the control of a constitutive promoter into the previously described strain (Figure 6). We then compared the ability of strains with each of these transporters, as well as a control strain with no transporters, to produce full-length GFP in the presence of AzF concentrations ranging from 2 µM to 200 µM (Figure 9).

Each strain was tested in both a mineral medium, as described above, and a rich Lennox Broth-based medium, which is commonly used for recombinant protein expression—the current

primary application of genetic code expansion. Expression of the putative PqiABC transporter, which is related to the paraquat-inducible transporter from *E. coli*, greatly reduced the ability to generate full length GFP – suggesting that it exports rather than imports AzF. One transporter, which is predicted to encode a branched chain amino acid transporter, greatly increased increased production of full length GFP in a dose dependent manner in both mineral medium and Lennox Broth, suggesting that it can import AzF and is likely responsible for at least some of the native AzF import. Interestingly, the magnitude of improvement was highest in Lennox Broth where the GFP production was increased nearly 10-fold versus the control strain. Indeed, the production of full length GFP produced by this strain appears to be higher when cultivated in medium containing 20 μM than GFP produced by the other strains when cultivated in 10-fold higher AzF concentration. This supports the supposition that these gene encode a transporter than efficiently imports the ncAA into the cell. Finally, the other two transporters also showed improved full lenfth GFP production over the control strain when cultivated in in Lennox Broth, but there was no statistical increase in production between the strains during growth in mineral media. The physiological basis of the different results in the media types is not currently clear, but it is an interesting direction for future research. Ultimately this work shows that leveraging substrate toxicity and high-throughput fitness assays is an effective method to identify how compounds are transported into a cell and that this approach can be leveraged for biotechnological purposes.

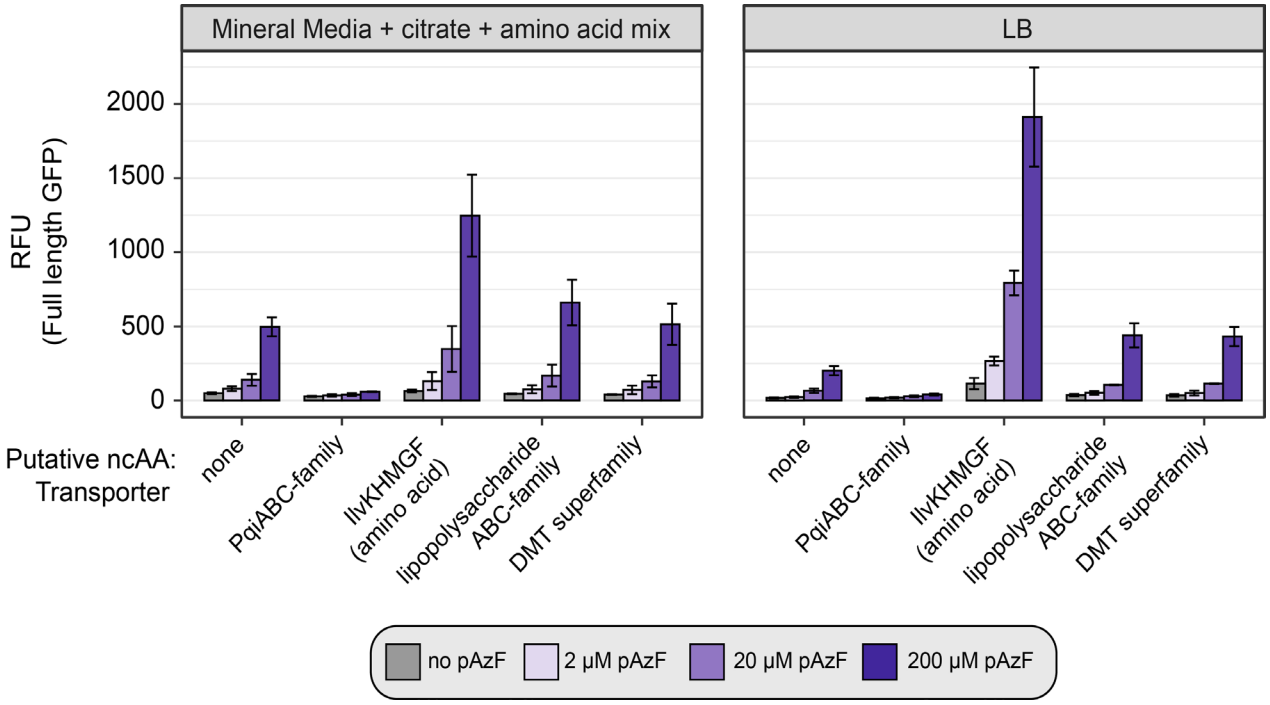

Figure 9. Production of full length GFP by *P. putida* strains that do or do not constitutively express genes encoding the putative AzF transporters identified in the fitness screens in the presence of a range of AzF concentrations. The y-axis represents the median relative fluorescence of 200,000 cells measured using flow cytometry. The expressed transporter is indicated below the x-axis. Coloring of the bar indicates the concentration of AzF in the cultivation medium. RFU values have had the autofluorescence of wild-type *P. putida* subtracted, and thus represent the fluorescence that is associated with full length GFP. Error bars represent the standard error in 3 biological replicates.

3.0 Accomplishments

Data generated in this work has supported the successful renewal of the Persistence Control of Engineered Functions Science Focus Area. It has also supported a poster presentation at the Society of Industrial Microbiology and Biotechnology Annual Meeting in 2023. We are also utilizing data generated from this work to support two manuscripts that are currently in preparation. Finally, this data will support future proposals for funding from multiple different external funding sources.

4.0 References

DeLey Cox, V. E., M. F. Cole and E. A. Gaucher (2019). "Incorporation of Modified Amino Acids by Engineered Elongation Factors with Expanded Substrate Capabilities." ACS Synthetic Biology 8(2): 287-296.

Elmore, J. R., G. N. Dexter, H. Baldino, J. D. Huenemann, R. Francis, G. L. t. Peabody, J. Martinez-Baird, L. A. Riley, T. Simmons, D. Coleman-Derr, A. M. Guss and R. G. Egbert (2023). "High-throughput genetic engineering of nonmodel and undomesticated bacteria via iterative site-specific genome integration." Science Advances **9**(10): eade1285.

Elmore, J. R., G. Peabody, R. K. Jha, G. N. Dexter, T. Dale and A. Guss (2022). "Dynamic control systems that mimic natural regulation of catabolic pathways enable rapid production of lignocellulose-derived bioproducts." bioRxiv: 2022.2001.2012.475730.

Martínez-García, E. and V. de Lorenzo (2024). "*Pseudomonas putida* KT2440 as a synthetic biology chassis and a metabolic engineering platform." Current Opinion in Biotechnology **85**.

Ozer, E., K. Yaniv, E. Chetrit, A. Boyarski, M. M. Meijler, R. Berkovich, A. Kushmaro and L. Alfonta (2021). "An inside look at a biofilm and flagella biotracking." Science Advances **7**(24).

Porter, J. J. and R. A. Mehl (2018). "Genetic Code Expansion: A Powerful Tool for Understanding the Physiological Consequences of Oxidative Stress Protein Modifications." Oxidative Medicine and Cellular Longevity 2018: 7607463.

Stork, D. A., G. R. Squyres, E. Kuru, K. A. Gromek, J. Rittichier, A. Jog, B. M. Burton, G. M. Church, E. C. Garner and A. M. Kunjapur (2021). "Designing efficient genetic code expansion in Bacillus subtilis to gain biological insights." Nature Communications **12**(1).

Wetmore, K. M., M. N. Price, R. J. Waters, J. S. Lamson, J. He, C. A. Hoover, M. J. Blow, J. Bristow, G. Butland, A. P. Arkin and A. Deutschbauer (2015). "Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons." Mbio **6**(3).

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov