Distribution Agreement

Signature:

________________________________ _______________

Kristen Hall Blanchard Date

Geobacillus stearothermophilus NUB3621 as a vector for metabolic engineering

By

Kristen Hall Blanchard
Doctor of Philosophy

Genetics and Molecular Biology

Ichiro Matsumura
Advisor

Paul Doetsch
Committee Member

Judith Fridovich-Keil
Committee Member

Charles Moran
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Geobacillus stearothermophilus NUB3621 as a vector for metabolic engineering

By

Kristen Hall Blanchard

B.S., Emory University, 2010

Advisor: Ichiro Matsumura, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Genetics and Molecular Biology

2016

Abstract

Microbial cells can be coopted to produce desired compounds. Utilizing cells to synthesize products avoids creating racemic mixtures, which can be undesirable when a desired compound behaves differently than its enantiomer. Cellular product synthesis generally requires metabolic engineering of the host strain, or manipulating the cells' metabolic environment to improve production of the target molecule.

Metabolic engineering requires well-studied host systems, and a variety of proteins and pathways from heterologous sources. Although there are many candidates for host strains for metabolic engineering, the genus Geobacillus is especially attractive due to its unique temperature range (39-75 degrees Celsius.) This unique range makes Geobacillus species capable of expressing both mesophilic and thermophilic proteins. Among Geobacillus species, *Geobacillus stearothermophilus* NUB3621, GsNUB3621, is especially attractive because it is more transformable than other Geobacillus species.

In this work, we seek to improve upon the utility of GsNUB3621 as a host strain for metabolic engineering. To do this, we have sequenced its genome, which should provide insight into GsNUB3621's metabolic network. We have also developed two expression constructs, one inducible and one constitutive, that can be used to express foreign proteins.

These tools help improve the utility of GsNUB3621 as a strain for metabolic engineering. Other Geobacillus strains have already shown use as a vector for ethanol production, and these tools may allow GsNUB3621 to fulfill the same purpose. Because of GsNUB3621's higher transformation efficiency, it may be able to be used for other metabolic engineering purposes less feasible in other Geobacillus strains.

Geobacillus stearothermophilus NUB3621 as a vector for metabolic engineering

By

Kristen Hall Blanchard

B.S., Emory University, 2010

Advisor: Ichiro Matsumura, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Genetics and Molecular Biology
2016

Acknowledgements

First and foremost, I would like to thank my advisor, Ichiro Matsumura, for always pushing me to be my best and for teaching me almost everything I know about science. Thank you for genuinely caring about my development as a scientist. I would like to thank every member of my committee: Paul Doetsch, Judith Fridovich-Keil, Justin Gallivan, and Charles Moran. Thank you for your amazing guidance both scientifically and in terms of navigating the graduate school process.

I want to thank my parents, Susan and Tom, for being the two most supportive people in the universe. I’d also like to thank my sister for generally being awesome and for being someone I can look up to. (This is the only time I will ever admit that.) To my Grandad and Janice, thank you for always being proud of me and believing in me even when I didn’t believe in myself. To my Grandma, I truly wouldn’t have made it without your continual support and prayers. You are the best.

Finally, I want to thank all members of the Matsumura lab past and present. Jessica Eliason, whether it was joking about video games or commiserating about Unix, you have been an amazing labmate and friend. Thank you to Brindar Sandhu for learning so many techniques with me and always being encouraging. And to Srebrenica Robic, thank you for being a mentor, friend, and an incredible labmate.

Table of contents

Figures and tables

Chapter 1: Introduction

1. Metabolic engineering can generate enantiopure quantities of desired molecules

Both pharmaceutical and chemical industries depend on the ability to precisely and efficiently synthesize specific molecules. There are many strategies to chemically synthesize compounds, each with its own advantages and disadvantages. However, one limitation of chemical synthesis is that it results in racemic mixtures of the desired compound. Often the enantiomer of a desired molecule can have different properties than its pair. [1] The drug Ethambutol is one example of the importance of enantiomeric selectivity. The D-form of Ethambutol can be used to treat tuberculosis, but the L-form of this molecule causes blindness. [2] Similarly, the morning sickness drug Thalidomide and the arthritis medication Naproxen each have enantiomeric forms. While the R-form of these molecules is therapeutic, the S-form can cause birth defects when used by pregnant women. [3] [4]. Thus, the ability to synthesize enantiopure molecules can be especially important in pharmaceutical production.

Enzymatic synthesis can efficiently solve the problem of generating enantiopure quantities of a desired product. For example, 1,2-Propanediol (1,2-PDO) is a popular industrial molecule with both an R form and an S form. Although the racemic mixture of 1,2-PDO has been used in a variety of applications such as antifreeze, production of unsaturated polyester resins, and as a pharmaceutical solvent, enantiopure quantities of this molecule could potentially be used to synthesize specialty chemicals and chiral pharmaceuticals. [5] Fortunately, enzymatic synthesis of 1,2-PDO can be utilized to generate enantiopure quantities. [6] Zhu et al were able to engineer *E. coli* cells to produce 13.7 mM 1,2-PDO at >99% enantiopurity. [7] Microbial cell factories have been used to make building blocks for synthetic polymers, replacing the use of petroleum and other non renewable resources. [8] Alginate, a biopolymer used in medical applications as well as laboratory applications and even commercial products, can be produced by *Pseudomonas*, resulting in quantities with more defined chemical structures than alginate extracted from seaweed. [9] [10] Microbial cell factories have also been used to engineer squalene, a precursor molecule used to generate sterols, hormones, and vitamins; diols, compounds used to generate chemicals and fuels; and aromatics, building blocks for chemicals, plastics, and solvents. [11] [12] [13]

Production of a desired product within a microbial cell often requires metabolic engineering of the host in order to introduce heterologous pathways, improve flux of the desired pathway, or alter properties of existing enzymes in the pathway. For example, to generate *E. coli* producing the S form of 1,2-PDO, Zhu and colleagues had to delete the D-lactate dehydrogenase gene and replace it with an L-lactate dehydrogenase gene from *Bacillus coagulans*. They also had to inactivate genes from other branch metabolic pathways from glucose, and introduce an NADH and acetyl-CoA regeneration system. Although much work has been done to improve the utility of *E. coli* as a metabolic engineering chassis, the field of metabolic engineering greatly benefits from novel host strains that may be able to adapt to different metabolic stresses than can *E. coli*, and from novel sources of proteins that could potentially catalyze desired reactions. [14]

Thermophilic, and especially facultatively thermophilic strains are of special interest in metabolic engineering applications as both host strains and as sources of potential pathways and enzymes for product synthesis. Microbial cell factories that can survive at high temperatures are highly desirable; according to the Arrhenius equation, reaction rates should increase exponentially with temperature. Additionally, since many microbes cannot survive at elevated temperatures, high temperatures reduce the risk of microbial contamination. Many groups have used thermophilic organisms like *Caldicellulosiruptor bescii* as a host strain for ethanol production – the fact that these organisms can grow near the boiling point of ethanol is extremely advantageous as it reduces potential ethanol toxicity and improves the ease of its extraction. [15] Proteins from these thermophilic organisms have also been introduced into mesophilic hosts. Facultative thermophiles present an additional advantage. Because they can grow at both mesophilic and thermophilic temperatures, they can potentially be used as vehicles for *in vivo* evolution of thermostable proteins. Thus, facultative thermophilic species are attractive tools for metabolic engineering.

2. The genus *Geobacillus*

The genus *Geobacillus* is a member of the *Bacillus* superfamily. One of the first members of this genera was identified in 1920 and initially name *Bacillus stearothermophilus* [16]. At this time, any rod-shaped bacteria was included in the *Bacillus* superfamily. The first restructuring of this superfamily began in 1991 when Ash et al analyzed newly available 16s rRNA sequences for 51 *Bacillus* species. [17] They proposed that these species should be separated into five distinct genera; the “group 5” species represented what would one day be the *Geobacillus* genus. In 2001, Nazina et al further categorized and analyzed this subset of “*Bacillus*” species. [18] Nazina et al had previously identified two species of aerobic, thermophilic rod-shaped bacteria that possessed sequence similarity to the “group 5” 16s rRNA sequences as described by Ash et al. Thus, they proposed the use of a new genus, *Geobacillus*, to encompass *Bacillus stearothermophilus*, *Bacillus thermoleovorans*, *Bacillus thermocatenulatus*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius*, *Bacillus thermodenitrificans* and their two isolated strains. Members of this genus are rod-shaped cells with Gram positive cell walls, though they are capable of sporulation under harsh environmental conditions. They can grow under a wide range of pH, from 6.0 to 8.5. However, what makes this genus truly unique is the optimal temperature range, with many species growing at a range of 37-75 degrees Celsius. [18]

2.1 Geobacillus is an excellent vector for metabolic engineering

The unique temperature range favored by *Geobacillus* species makes these organisms an excellent tool in the field of metabolic engineering. Utilizing high temperatures for synthesis of desired products increases the speed of reactions and reduces the risk of microbial contamination. Thus, thermophilic proteins are highly desired for microbial biosynthesis designs. *Geobacillus* species have become attractive targets as sources of innately thermostable proteins. [19] [20] [21]

In addition to their utility as sources of thermostable proteins, *Geobacillus* species can also serve as vectors to evolve thermostable proteins. Because many *Geobacillus* species can grow at mesophilic temperatures (~37 degrees Celsius), it should be possible to transform *Geobacillus* cells with proteins from mesophilic sources and screen or select for protein activity. By mutating the protein’s gene sequence and selecting or screening for continued activity at higher temperatures, one can improve the thermostability of a protein with no rational design or any knowledge of a protein’s structure. The feasibility of this approach was demonstrated by Shamoo and Counago in 2006. [22] They replaced the adenylate kinase gene in *Geobacillus stearothermophilus* NUB3621 with the adenylate kinase from the mesophilic *Bacillus subtilis*. [23] Because the adenylate kinase gene from *B. subtilis* possessed low thermostability, Shamoo and Counago's recombinant strain could no longer grow at high temperatures (over 56 degrees). However, because of the strain's wide range of temperature tolerance, Shamoo and Counago were able to grow their recombinant strain at increasingly high temperatures and select for improved thermostability of the encoded adenylate kinase enzyme.

Geobacillus species can also serve as host vectors for biosynthesis of desired compounds. Because they can grow at elevated temperatures, reactions can proceed faster, and potential mesophilic microorganisms cannot survive and contaminate cultures. *Geobacillus* species have already been utilized for the production of ethanol and isobutanol [24] [25] [26] however, it is also potentially useful for xylose metabolism. [27] When transformed with cellulase and α-amylase, *Geobacillus* species can digest cellulose and insoluble starch, suggesting that they could potentially be exploited for the utilization of plant biomass. In this experiment, researchers found that *Geobacillus kaustophilus* HTA426 was even able to express gene products as soluble proteins that were insoluble in *E. coli,* suggesting yet another advantage to this system. However, transformation of *Geobacillus* has frequently proven difficult, thus any improvements to the ability to transform *Geobacillus* would greatly benefit the field.

2.2 GsNUB3621 is a uniquely useful *Geobacillus* strain

Geobacillus stearothermophilus NUB3621 was identified and studied before the *Geobacillus* genus had even been created. Welker and colleagues isolated a rod-shaped facultative thermophile originally named *Bacillus stearothermophilus* NUB36. From this parent strain, the Welker lab was able to do much of the pioneering work in understanding what would eventually be the *Geobacillus* genus. They generated a collection of mutants from this parent strain and using these mutants were able to generate a preliminary genetic map long before whole genome sequencing was feasible. One of these mutants, NUB3621 (GsNUB3621), was selected for its lack of a restriction modification system. Using this strain, Welker and colleagues were able to work out a transformation protocol. Though other transformation protocols have been developed, the Welker group's achieved transformation efficiency still hasn't been matched. This high efficiency is likely due to GsNUB3621's lack of a restriction modification system. Thus, GsNUB3621 is a uniquely attractive strain for protein engineering. Protein engineering often requires transforming a strain with hundreds of thousands of variants of a gene, so the process can be greatly expedited by using a host strain with a high transformation efficiency.

3. Project scope

The goal of this project was to improve the accessibility of *Geobacillus stearothermophilus* NUB3621 for metabolic engineering. Though the Welker lab made huge strides in improving the utility of this organism, many important tools were lacking. While other strains may seem like attractive alternatives due to existing genetic sequences, certainty of phylogentic position, etc. GsNUB3621 fills a unique niche among *Geobacillus* strains. Because it lacks a restriction modification system, it is likely to have the highest transformation efficiency among *Geobacillus* strains, regardless of the method of transformation used. High transformation efficiencies can be crucial for experiments involving large libraries, making GsNUB3621 the ideal host strain for metabolic engineering. The next chapter of this thesis will describe steps taken to improve the utility of GsNUB3621 as a laboratory organism. This work included sequencing and annotating the genome as well as creating two expression constructs for GsNUB3621 – a fluorescent reporter gene under the control of a constitutive promoter and an alpha galactosidase gene under the control of an inducible promoter. The appendices describe two projects intended to provide further insight into metabolic engineering that could potentially benefit engineering designs in GsNUB3632. The goal of the first experiment was to determine the evolutionary potential of the *E. coli* proteome, which should identify new targets that could potentially be evolved in metabolic engineering designs. The experiment detailed in the second appendix questions the flexibility of enzymatic reactions. By identifying potential flexibility, it may be possible to find new ways to exploit catalytic intermediates and design non-native metabolic schemes for product synthesis. The insights yielded by these experiments should further help to improve the utility of GsNUB3621 as a host strain for metabolic engineering.

References

1. Chhabra, N., M.L. Aseri, and D. Padmanabhan, *A review of drug isomerism and its significance.* Int J Appl Basic Med Res, 2013. **3**(1): p. 16-8.
2. Lim, S.A., *Ethambutol-associated optic neuropathy.* Ann Acad Med Singapore, 2006. **35**(4): p. 274-8.
3. Agranat, I., H. Caner, and J. Caldwell, *Putting chirality to work: the strategy of chiral switches.* Nat Rev Drug Discov, 2002. **1**(10): p. 753-68.
4. Chambers HF, D.D., *Basic and Clinical Pharmacology*. 2009, Noida, UP, India: Tata Macgraw- Hill.
5. Cameron, D.C., et al., *Metabolic engineering of propanediol pathways.* Biotechnol Prog, 1998. **14**(1): p. 116-25.
6. Liao, D.I., et al., *Crystal structure of substrate free form of glycerol dehydratase.* J Inorg Biochem, 2003. **93**(1-2): p. 84-91.
7. Zhu, L., et al., *Fermentative production of enantiomerically pure S-1,2-propanediol from glucose by engineered E. coli strain.* Appl Microbiol Biotechnol, 2016. **100**(3): p. 1241-51.
8. Tsuge, Y., et al., *Engineering cell factories for producing building block chemicals for bio-polymer synthesis.* Microb Cell Fact, 2016. **15**(1): p. 19.
9. Lee, K.Y. and D.J. Mooney, *Alginate: properties and biomedical applications.* Prog Polym Sci, 2012. **37**(1): p. 106-126.
10. Maleki, S., et al., *Alginate Biosynthesis Factories in Pseudomonas fluorescens: Localization and Correlation with Alginate Production Level.* Appl Environ Microbiol, 2015. **82**(4): p. 1227-36.

11. Ghimire, G.P., et al., *Advances in Biochemistry and Microbial Production of Squalene and Its Derivatives.* J Microbiol Biotechnol, 2016. **26**(3): p. 441-51.

12. Sabra, W., C. Groeger, and A.P. Zeng, *Microbial Cell Factories for Diol Production.* Adv Biochem Eng Biotechnol, 2016. **155**: p. 165-97.

13. Thompson, B., M. Machas, and D.R. Nielsen, *Creating pathways towards aromatic building blocks and fine chemicals.* Curr Opin Biotechnol, 2015. **36**: p. 1-7.

14. Liu, M., et al., *Metabolic engineering of Escherichia coli to improve recombinant protein production.* Appl Microbiol Biotechnol, 2015. **99**(24): p. 10367-77.

15. Chung, D., et al., *Cellulosic ethanol production via consolidated bioprocessing at 75 degrees C by engineered Caldicellulosiruptor bescii.* Biotechnol Biofuels, 2015. **8**: p. 163.

16. Donk, P.J., *A Highly Resistant Thermophilic Organism.* J Bacteriol, 1920. **5**(4): p. 373-4.

17. Ash, C., Farrow, J.A.E., Wallbanks, S. and Collins, M.D., *Phylogenetic heterogeneity of the genus Bacillus revealed by comparative analysis of small-subunit-ribosomal RNA sequences. .* Letters in Applied Microbiology, 1991(13): p. 202–206.

18. Nazina, T.N., et al., *Taxonomic study of aerobic thermophilic bacilli: descriptions of Geobacillus subterraneus gen. nov., sp. nov. and Geobacillus uzenensis sp. nov. from petroleum reservoirs and transfer of Bacillus stearothermophilus, Bacillus thermocatenulatus, Bacillus thermoleovorans, Bacillus kaustophilus,*

Bacillus thermodenitrificans to Geobacillus as the new combinations G. stearothermophilus, G. th. Int J Syst Evol Microbiol, 2001. **51**(Pt 2): p. 433-46.

19. Parashar, D. and T. Satyanarayana, *A chimeric alpha-amylase engineered from Bacillus acidicola and Geobacillus thermoleovorans with improved thermostability and catalytic efficiency.* J Ind Microbiol Biotechnol, 2016. **43**(4): p. 473-84.

20. Chen, Y., et al., *Identification of novel thermostable taurine-pyruvate transaminase from Geobacillus thermodenitrificans for chiral amine synthesis.* Appl Microbiol Biotechnol, 2016. **100**(7): p. 3101-11.

21. Bhalla, A., K.M. Bischoff, and R.K. Sani, *Highly Thermostable Xylanase Production from A Thermophilic Geobacillus sp. Strain WSUCF1 Utilizing Lignocellulosic Biomass.* Front Bioeng Biotechnol, 2015. **3**: p. 84.

22. Counago, R., S. Chen, and Y. Shamoo, *In vivo molecular evolution reveals biophysical origins of organismal fitness.* Mol Cell, 2006. **22**(4): p. 441-9.

23. Counago, R. and Y. Shamoo, *Gene replacement of adenylate kinase in the gram-positive thermophile Geobacillus stearothermophilus disrupts adenine nucleotide homeostasis and reduces cell viability.* Extremophiles, 2005. **9**(2): p. 135-44.

24. Kananaviciute, R. and D. Citavicius, *Genetic engineering of Geobacillus spp.* J Microbiol Methods, 2015. **111**: p. 31-9.

25. Van Zyl, L.J., et al., *Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius.* Appl Microbiol Biotechnol, 2014. **98**(3): p. 1247-59.

26. Cripps, R.E., et al., *Metabolic engineering of Geobacillus thermoglucosidasius for high yield ethanol production.* Metab Eng, 2009. **11**(6): p. 398-408.

27. Cordova, L.T., et al., *Complete genome sequence, metabolic model construction and phenotypic characterization of Geobacillus LC300, an extremely thermophilic, fast growing, xylose-utilizing bacterium.* Metab Eng, 2015. **32**: p. 74-81.

Chapter 2

Transformable facultative thermophile *Geobacillus stearothermophilus* NUB3621 as a host strain for metabolic engineering

Kristen Blanchard[1], Srebrenka Robic[2], Ichiro Matsumura[1]*

[1]Emory University School of Medicine, Department of Biochemistry, 1510 Clifton Road NE, room 4119, Atlanta, Georgia 30322, USA.

[2]Agnes Scott College, Department of Biology, 141 E. College Ave., Decatur, GA 30030, USA.

Published in *Applied Microbiology and Biotechnology* Volume 98, Issue 15, August 2014, Pages 6715-6723

Author contributions: KB, SR, and IM designed project scope. KB conducted genome assembly and analysis. KB and SR designed and tested expression constructs. KB and IM wrote the manuscript.

Abstract

Metabolic engineers offer inexpensive enantioselective syntheses of high value compounds, but their designs are sometimes thwarted by the misfolding of heterologously expressed proteins. *Geobacillus stearothermophilus* NUB3621 is a readily transformable facultative thermophile. It could be used to express, and properly fold, proteins derived from its many mesophilic or thermophilic *Bacillaceae* relatives, or to direct the evolution of thermophilic variants of mesophilic proteins. Moreover, its capacity for high temperature growth should accelerate chemical transformation rates in accordance with the Arrhenius equation, and reduce the risks of microbial contamination. Its tendency to sporulate in response to nutrient depletion lowers the costs of storage and transportation. Here we present a draft genome sequence of *G. stearothermophilus* NUB3621, and describe inducible and constitutive expression plasmids that function in this organism. These tools will help us, and others, to exploit the natural advantages of this system for metabolic engineering applications.

Introduction

Metabolic engineers manipulate microbes in order to convert inexpensive input compounds into valuable outputs. Their work is unfortunately impeded by design failure because foreign proteins, which are necessary for the construction of non-native metabolic pathways, often fail to fold properly in host cells. Protein misfolding is all too common because wild-type proteins are marginally stable (average $\Delta G_{folding}$ = -14 kcal/mol [1]); apparently modest changes in temperature, amino acid sequence or chemical environment can cause misfolding or unfolding. Heterologous expression causes foreign proteins, and their folding intermediates, to interact with high concentrations (300 - 400 mg/mL) of host protein, sometimes causing aggregation [2]. Most synthetic biologists acknowledge these limitations [3], but nevertheless rely heavily upon *E. coli* and other obligate mesophiles.

Geobacillus stearothermophilus NUB3621 (hereafter called GsNUB3621) offers a general solution to the protein misfolding problem. It belongs to the family *Bacillaceae*, which diversified to adapt to a variety of very different environments. The functional diversity of this superfamily suggests that natural evolution has already created a diverse set of compatible "parts" that could be artificially mixed and matched without aggregation. GsNUB3621 is a facultative thermophile that achieves balanced growth between 39 - 75° C [4]. It is therefore likely to fold proteins derived from its mesophilic and thermophilic relatives correctly, when propagated at the appropriate temperature. Moreover, others have already demonstrated that GsNUB3621 can serve as a host to direct the evolution of thermostable variants of mesophilic proteins (*vide infra*).

GsNUB3621 is not the only facultative thermophile, but it is particularly amenable to experimental manipulation. The parent strain, *G. stearothermophilus* NUB36, was extensively studied by Neil Welker and his colleagues. They developed protocols for the growth, transformation [5], and genetic analysis [6] of this strain, and created a genetic map [7]. They isolated the GsNUB3621 mutant, which lacked a functioning restriction-modification system. Protoplasts of this strain can be transformed with an efficiency of 10^7-10^8 transformants per microgram of DNA [5]. *Geobacillus thermoglucosidasius* [8] and *Geobacillus kaustophilus* HTA-426 [9] have also been transformed, but thus far much less efficiently.

Couñago et al. used GsNUB3621 to direct the evolution of thermostable variants of the adenylate kinase gene from *Bacillus subtilis* [10]. In this groundbreaking experiment, a mesophilic enzyme was evolved to fold AND function at high temperature. Additionally, Peña et al have utilized this organism as a vehicle for directed evolution to study the role of protein folding within evolution [11]. Though it is technically easier to express and fold proteins in *Escherichia coli* at 37° C, and to assay them *ex vivo* at higher temperatures [12], such an approach is less likely to favor mutations that promote proper folding at higher temperatures.

Couñago's feat was even more impressive when one considers the dearth of available genetic tools (plasmids, promoters and reporter genes) for this non-model organism. Such tools must be developed in tandem, as no “positive control” was available to troubleshoot potential constructs. The utility of GsNUB3621 was also limited by the absence of a genome sequence. Although a genetic map exists [7], it provides no specific sequences or the metabolic pathways. In our experience, degenerate primers based on other *Geobacillus* genome sequences rarely result in high PCR yields. Here we present a high-quality draft genome sequence of GsNUB36321, and describe two promoters, one inducible and one constitutive, and two reporter proteins, that facilitate the employment of this system for metabolic engineering.

Materials and Methods

Materials

All chemicals utilized in this study were from Sigma-Aldrich (St. Louis, MO.) DNA oligonucleotides (Table 1) were custom synthesized by Integrated DNA Technologies (Coraville, IA). All enzymes used for cloning and PCR amplification were purchased from New England Biolabs (Ipswitch, MA.) *E. coli* strain InvaF' (Invitrogen, Grand Island, NY) and custom BioBrick accepting vector, pIMBB [13] were used to clone genes. *Geobacillus stearothermophilus* NUB3621 and shuttle vector pNW33N were provided by the Bacillus Genetic Stock center (Columbus, OH.) The Qiaprep Spin Miniprep Kit (Qiagen, Valencia, CA) was used to purify all plasmids (Table 2).

Genome assembly and annotation

Genomic DNA was harvested using Qiagen's DNeasy DNA purification protocol for Gram-Positive Bacteria. The DNA was then sequenced by whole genome shotgun sequencing using Illumina technology, generating 37,651,593 100 bp reads. These reads were assembled using ABySS version 1.3.2 and setting the k-mer length to 88 [14]. This initial assembly yielded 336 contigs. These contigs were merged to form larger contigs by using iterative BLAST [15] searches to identify regions of overlap between contigs. A database of all of the contigs was generated using the standalone version of BLAST. Each of the 50 largest contigs was BLASTed against the database to identify contigs that overlapped at the ends. In many instances, multiple contigs were found to overlap with a search input. When possible, these multiple contigs were compared to sequences of other *Geobacillus* species. If one potential contig placement matched related *Geobacillus* species sequences better than the other possible placements, that contig was used in the final assembly.

In some instances contigs that overlapped in the same regions differed by only a few internal nucleotides, most frequently in genomic regions that are present in multiple, non-identical copies, such as 16s RNA. The selection of contigs in our final assembly was thus sometimes arbitrary. The consensus sequence for NUB3621 16s RNA determined here differs from the previously published NUB3621 16s RNA sequence [16] by less than 1% (5 nucleotides out of 1560), which is within the range of variance for two 16s RNA copies from the same organism for other *Geobacillus* species. This consensus sequence was used for all seven 16s RNA copies in the NUB3621 genome though it is unlikely that all are truly identical. This process of identifying contig overlap and extending the largest contigs resulted in ten final contigs containing 3621385 bp. The order and orientation of these ten contigs was determined by comparing to the genome sequence of *Geobacillus sp.* WCH70. *Geobacillus sp*. WCH70 was chosen because it possessed the highest degree of sequence similarity to the ten longest of the initial 336 contigs. *Geobacillus thermoglucosidasius* C56-YS93 appeared to be the next most closely related *Geobacillus* species; its genome sequence supported our current arrangement of contigs. Estimates of gap lengths vary based on whether *Geobacillus sp.* WCH70 or *Geobacillus thermoglucosidasius* C56-YS93 is used as the reference strain, so they are somewhat arbitrary.

The RAST server's online interface was used to annotate the open reading frames; the RAST gene caller was chosen as a basis for annotation. Gaps were set to an arbitrary size of 10 nucleotides, which is reflected in the feature start and end positions. For comparison of genes with *Geobacillus kaustophilus* HTA 426, this genome was also annotated using the RAST server and the same parameters. Features that were identical between the two genomes were deleted, duplications within a genome were deleted, and any genes designated as "hypothetical protein" with no additional information were deleted before comparing the genomes.

Phylogenetic analysis

Maximal unique matches were determined using MUMmer version 3.23 [17]. The minimum match length was set to 19 and both forward and reverse complement matches were set to be reported. MUM index was calculated using the perl script made available by Deloger et al. [18]. The tree was generated using Splitstree4 and choosing the neighbor-joining algorithm [19]. All genome sequences were taken from NCBI's database.

Sequence accession numbers

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AOTZ00000000. The version described in this paper is the first version, AOTZ01000000. The deposited version has been annotated by NCBI's Prokaryotic Genome Automatic Annotation Pipeline (PGAAP.)

Bacterial growth conditions

E. coli was grown in either liquid LB media or LB media supplemented with 1.5% (w/v) agar at a temperature of 37°C. GsNUB3621 was grown in modified LB (mLB) media [6], which contains 1.05 mM Nitrilotriacetic acid, 0.59 mM $MgSO_4 \cdot 7H_2O$, 0.91 mM $CaCl_2 \cdot 2H_2O$, 0.04 mM $FeSO_4 \cdot 7H_2O$ and 1.5% (w/v) agar as needed. GsNUB3621 was grown at 60°C unless otherwise noted. When grown in liquid culture, GsNUB3621 was grown with minimal volumes to maximize aeration (20 mL in a 250 mL flask.) The LB media was supplemented with 100 micrograms/mL ampicillin for selection of pIMBB-based plasmids in *E. coli,* or 34 micrograms/mL of chloramphenicol for selection of pNW33N-based plasmids. The mLB medium was augmented with 7 micrograms/mL of chloramphenicol for selection of GsNUB3621 transformed with pNW33N plasmids.

GsNUB3621 transformation

GsNUB3621 protoplasts were transformed as described previously [5] but with several modifications. Cultures were grown to saturation overnight, diluted 1/40 and propagated for 4.5 hours with shaking at 60° C until the cell density reached an OD 600 of approximately 1.8. For each transformation, 2 mL of log-phase culture was harvested in a microcentrifuge for one minute at 5000 RPM. The cells were resuspended in 500 microliters of protoplasting medium (mLB with 10% w/v lactose and 10 mM $MgCl_2 \cdot 7H_2O$); 10 microliters of 1 mg/mL lysozyme were added to each, and incubated for 20 minutes at 37° C in a water bath. The conversion of whole cells to protoplasts was verified by microscopy.

The protoplasts were then diluted with 500 microliters of protoplasting medium, harvested by centrifugation (5 minutes at 800 g), and resuspended in 100 microliters of protoplasting medium. One microgram of the appropriate plasmid DNA (100 ng/microliters x 10 microliters) was added to the washed protoplasts; 900 microliters of freshly prepared PEG solution (40% w/v PEG 6000 in protoplasting media) was subsequently added. The protoplast/DNA/PEG mixtures were then incubated for 2 minutes at 50° C with shaking (130 RPM), harvested by centrifugation for 5 minutes at 800 g, resuspended in 100 microliters of protoplasting media and incubated for an hour at 50° C with shaking at 130 RPM. The mixtures were then spread on regeneration plates (protoplasting media with 20 mM $CaCl_2 \cdot 2H_2O$, 0.8% agar, and 7 micrograms/mL chloramphenicol), incubated for 12 hours at 50° C, then at 60° C until the appearance of colonies. At the first sign of growth, the colonies were replica-plated onto fresh mLB plates supplemented with 7 microgram/mL of chloramphenicol and incubated at 60° C overnight.

Cloning of *surT*-P_{surP}-*agaN*-pNW33N and P_{RHIII}-sfGFP-pNW33N

The *agaN* reporter gene was PCR amplified from NUB3621 genomic DNA and blunt-end cloned into an *Eco*RV-cut pIMBB (Table 3) accepting vector [13]. The regulatory region upstream of the *surT* promoter to the *surP* start codon was PCR amplified from GsNUB3621 genomic DNA. An EcoRI site within this region was eliminated by introducing a synonymous single base pair change through two separate

PCRs that were then joined through overlap extension PCR. The *surT*-P$_{surP}$ regulatory region was then joined to *agaN* in the pIMBB cloning plasmid by cutting the *surT*-P$_{surP}$ PCR product with *Nsp*I and *Nco*I, cutting GFP-pIMBB with *Sph*I and *Spe*I, and cutting *agaN*-pIMBB with *Nco*I and *Spe*I and ligating the three fragments together. The *surT*-P$_{surP}$-*agaN* fragment was then cloned from the pIMBB vector into pNW33N by cutting both plasmids with *Eco*RI and *Pst*I. The plasmid was deposited into the Addgene repository (ID 44009).

The sfGFP gene was synthesized by GeneArt to reflect the codon bias of *Geobacillus stearothermophilus* [20]; internal *Eco*RI (nucleotide 669) and *Nde*I (nucleotide 231) sites were changed to synonymous codons. The BioBrick prefix was added to the 5' end and a BioBrick suffix was added to the 3' end. An *Nco*I site was added at the start codon and a glycine sequence was added in the second amino acid position to correct a frameshift that would otherwise have been caused by the addition of the *Nco*I site. The ribonuclease H III promoter was synthesized by IDT as a mini gene construct. Restriction endonucleases *Nco*I and *Pst*I were used to subclone sfGFP gene downstream of P$_{RHIII}$. *Eco*RI and *Pst*I were then used to subclone the P$_{RHIII}$-sfGFP cassette into pNW33N to create P$_{RHIII.}$-sfGFP-pNW33N (Addgene 52217).

Alpha galactosidase assays

Geobacillus stearothermophilus NUB3621 assays were carried out as described previously [21, 22] with slight modifications. Reactions containing 850 microliters of supernatant, 50 microliters of 1 M Tris pH 7.6, and 100 microliters of 40 mM 4-methylumbelliferyl-alpha-D-galactopyranoside (4-MU) were incubated at 55° C; 10

microliter aliquots were removed at 5 minute intervals and added to 100 microliters of 0.2 N NaOH in 96 well, white, clear bottom microtiter plates to quench the reaction. Fluorescence was measured in a SpectraMax M5 plate reader using an excitation wavelength of 355 nm and an emission wavelength of 460 nm. The photomultiplier automatic setting was used and the slit widths were set to the defaults (9 nm for excitation and 15 nm for emission.)

sfGFP assays

GsNUB3621 carrying P_{RHIII}-sfGFP-pNW33N was propagated overnight to saturation. The cells were harvested by centrifugation at 3500 RPM for 15 minutes; the cells were resuspended in 2 mL of 50 mM Tris pH 8.0, 10 mM EDTA, 2 mg/mL lysozyme and incubated at 37° for 30 minutes. The cell debris was removed by centrifugation for 2.5 minutes at 7000 RPM; 100 microliter aliquots of the supernatant were transferred to 96 well microtiter plates and assayed. Fluorescence was measured in a SpectraMax M5 plate reader using an excitation wavelength of 470 nm and a cutoff value of 495 nm. The photomultiplier automatic setting was used and the slit widths were set to the defaults (9 nm for excitation and 15 nm for emission.)

Results

Genome sequence of GsNUBC3621

Genome sequences have become *de rigueur* for metabolic engineering. Couñago et al., for example, worked without a genome sequence, which meant that they had to sequence their knock-out target gene in the GsNUB3621 chromosome and show that it was essential [23]. We used the Illumina method to sequence the GsNUB3621 genome.

Assembly of the reads resulted in ten ordered contigs (Table 1), an "improved high-quality draft genome" [24]. Scaffolding was completed by aligning the contigs against a related genome sequence, that of *Geobacillus sp.* WCH70. We emphasize that the gap estimates could differ substantially from the true gap sizes. In some instances, the first or last several kilobases of a contig had no sequence identity to *Geobacillus sp.* WCH70. Thus, the estimated “start” or “end” positions for some contigs correlate to positions a few kb within the contig, rather than the first or last nucleotide of the contig. We tried to resolve these gaps via targeted sequencing, but were unable to PCR amplify them. The gaps might be longer than we estimate, or contain repetitive sequences.

To verify the current order and orientation of the contigs, we also compared the contigs to the genome of *Geobacillus thermoglucosidasius* C56-YS93 (not shown). The BLAST hits against the *G. thermoglucosidasius* C56-YS93 genome confirm the general order and orientation of the contigs, though the gap estimates vary, and synteny within contigs is not always conserved between *Geobacillus sp.*WCH70 and *G. thermoglucosidasius* C56-YS93. We also attempted to validate this scaffold by comparison with the existing GsNUB36 genetic map [7]. Though we cannot be certain of the genetic loci that correspond to the phenotypes used to generate the map, we identified plausible candidate loci for many of the phenotypes, and our assignments were generally in concordance with the map (not shown).

The genome sequence was annotated using the RAST server [25]. The RAST gene caller identified 3929 features (annotations, listed in Supplemental Dataset 1) with an estimated eight features possibly missing from the sequence. The algorithm classified 47% of these features into RAST gene subsystems (Figure 1); it was unable to predict the functions of the majority of genes. The percentage of subsystem coverage, as well as the subsystem feature counts, are similar to those of *Geobacillus kaustophilus* HTA426 and *Geobacillus thermodenitrificans* NG80-2, which were already present in the RAST database's SEED viewer [26]. The features identified were then compared to those of the more extensively studied *Geobacillus kaustophilus* HTA426 genome [27, 28]. This comparison revealed 842 different features (Supplemental Dataset 2) between the two organisms (440 unique to GsNUB3621 and 402 unique to HTA426). Among the unique genes, *Geobacillus kaustophilus* HTA426 includes multiple restriction endonucleases, while GsNUB3621 has none. We hypothesize that this difference explains why the latter is more amenable to transformation.

The phylogenetic position of GsNUB3621 was somewhat uncertain [29, 30], so we compared our genome sequence to that of other *Geobacillus* genomes [31]. The MUM index (MUMi) is a computationally tractable method of estimating species relatedness utilizing whole genome data [18]. Briefly, MUMi utilizes the maximal unique matches (MUM) between two species as a measure of relatedness. MUMi values between GsNUB3621 and the eleven available *Geobacillus* whole genome sequences (*Geobacillus sp*. WCH70, accession number NC_012793.1; *Geobacillus thermoglucosidasius* C56-YS93, accession number NC_015660.1; *Geobacillus kaustophilus* HTA426, accession number NC_006510.1 [28]; *Geobacillus thermodenitrificans* NG80-2, accession number NC_009328.1 [32]; *Geobacillus thermoleovorans* CCB_US3_UF5, accession number NC_016593.1; *Geobacillus sp.* Y4.1MC1, accession number NC_014650.1; *Geobacillus sp.* Y412MC52, accession number NC_014915.1; *Geobacillus sp.* Y412MC61, accession number NC_013411.1;

Geobacillus sp. C56-T3, accession number NC_014206.1; *Geobacillus sp.* GHH01, accession number NC_020210.1; and *Geobacillus thermoglucosidans* TNO-09.020, accession number NZ_CM001483.1 [33]) were calculated and used to construct a phylogenetic tree (Figure 2). Our GsNUB3621 sequence is not identical to any of the other sequenced genomes. It remains unclear whether GsNUB3621 is closely related to other *G. stearothermophilus* strains, including the type strain, as none have yet been fully sequenced.

Inducible and constitutive expression of reporter genes in GsNUB3621.

Couñago et al. integrated a foreign gene into an existing operon within the GsNUB3621 chromosome [23]. A plasmid would enable more efficient transformations, which is essential for *in vitro* mutagenesis and recombination techniques. An inducible expression system would facilitate the heterologous expression of toxic genes. We constructed our inducible expression vector, *surT*-$\mathrm{P}_{\mathrm{SurP}}$-*agaN*-pNW33N, by combining parts that were developed or characterized by others. The parent vector for this plasmid, pNW33N, contains a thermostable chloramphenicol resistance gene (chloramphenicol acetyltransferase) that confers chloramphenicol resistance in both *E. coli* and GsNUB3621 [34]. This plasmid is fortuitously compatible with the BioBrick standard, a system of standardized restriction sites that facilitate the combinatorial assembly of multi-component biological devices [35].

The sucrose utilization operon regulatory region [36] consists of a sucrose phosphotransferase gene *surP* that is regulated by the *surT* gene product. The SurT antiterminator is believed to bind to palindromic *surR* region, allowing sucrose-induced transcription from the *surP* promoter (Figure 3). Our genome sequence data indicated that GsNUB3621 possesses two genes that encode alpha galactosidases. One matched the previously published GsNUB3621 *agaN* sequence [37], while the other was more distantly related. The former homologue was PCR amplified from GsNUB3621 and cloned downstream of *surT*-P_{surP}, thereby completing our *surT*-P_{SurP}-*agaN*-pNW33N expression vector. GsNUB3621 was transformed via protoplast transformation (Methods) and assayed for activity both in the presence and absence of sucrose. The alpha galactosidase activity in the supernatant increased five-fold when cultures were grown in sucrose (Figure 4a). Control cells carrying only the empty pNW33N plasmid exhibited little alpha-galactosidase activity, which suggests that the_chromosomal copy is not expressed under these conditions, and that our inducible promoter is somewhat leaky.

We looked through our sequence data to identify constitutive promoters. Among those considered, we focused on the promoter for ribonuclease H III (P_{RHIII}) gene because its -10 and -35 regions seemed most conserved across all promoter options. We coupled this promoter to the superfolding Green Fluorescent Protein (sfGFP), a thermostable variant of GFP [38]. Fluorescence was detected in cell extracts derived from GsNUB3621 transformed with P_{RHIII}-sfGFP-pNW33N, but not in control extracts in the empty vector (pNW33N) control (Figure 4b).

In addition to the *surT*-P_{SurP}-*agaN*-pNW33N and P_{RHIII}-sfGFP-pNW33N plasmids, we cloned and tested several others that did not work as well. We coupled either *agaN* or sfGFP, to the following promoters: T5 [39], T7 [40], tac [41], Spac [42] and the promoter from the chloramphenicol resistance gene present in pNW33N. We were unable to construct P_{Spac}-*agaN*-pNW33N and P_{RHIII}-*agaN*-pNW33N, which suggested to us that high levels of AgaN expression are toxic to *E. coli*. We constructed P_{Spac}-*sfGFP*-pNW33N, *lacI*-P_{T5}-*agaN*-pNW33N, and *lacI*-P_{tac}-*agaN*-pNW33N in *E. coli,* but were unable to obtain viable GsNUB3621 transformants. We transformed GsNUB3621 with P_{cat}-*agaN*-pNW33N and P_{cat}-*sfGFP*-pNW33N, but we did not observe any reporter protein expression. The *lacI*-P_{T7}-*agaN*-pNW33N plasmid apparently caused GsNUB3621 to express modest amounts of AgaN, less than 2 fold change over control cells carrying the empty pNW33N plasmid (data not shown).

Discussion

GsNUB3621 as source of robust proteins

Proteins that are robust to mutation tend to be more evolvable. Mutational robustness and conformational stability are correlated [43]; thermophilic bacteria are full of thermostable enzymes, but these might not be good starting points for directed evolution because few are active at mesophilic temperatures. In contrast, GsNUB3621 can grow at temperatures between 39 - 75° C so its proteins would probably be good starting points for the directed evolution of variants with novel catalytic functions (promiscuous or substrate ambiguous activities). The GsNUB3621 genome sequence could be used to identify candidate genes for PCR amplification, cloning in an *E. coli* expression vector and subsequent directed evolution in that mesophilic host. Alternatively, it should be relatively easy to assay GsNUB3621 cell extracts for desired catalytic activities, purify those thermostable activities via classical biochemical fractionation and to identify the gene through the mass or amino acid sequence of its product.

GsNUB3621 as a host strain for metabolic engineering applications

We developed an expression plasmid for GsNUB3621, and sequenced its genome, so that we and others can use this organism as a vehicle for the directed evolution of robust variants of mesophilic proteins. GsNUB3621 should also be a good host for metabolic engineering applications. Others have already shown that a related species, *G. themoglucosidasius*, is useful for biofuels production [8]. Enzyme-catalyzed reactions follow the Arrhenius equation, which predicts that reaction rates should increase exponentially with temperature (50 - 100% faster for each 10° C). Most enzymes denature above their optimum temperatures, but robust pathways within robust organisms are likely to be much faster and more efficient. Furthermore, bioreactors are much less likely to become contaminated at high temperatures. GsNUB3621 sporulates under starvation conditions [4]. *Geobacillus stearothermophilus* endospores are commonly used to test autoclaves and other disinfection protocols, so we tentatively expect that GsNUB3621 endospores will similarly be resilient. The capacity of endospores to persist indefinitely without refrigeration will save energy and simplify the transport of engineered strains to remote locations. The tools described here allow us and others to exploit the natural advantages GsNUB3621 for metabolic engineering applications.

Acknowledgements

We thank the Bacillus Genetic Stock Center for GsNUB3621, the pNW33N plasmid, and their protocols. We also thank the Emory Integrated Genomics Core for their service, and Paul Doetsch for the use of his spectrofluorimeter. This work was supported by the National Institute of General Medicine at the National Institutes for Health. KB and IM were supported by R01 GM086824 and KB was also supported by 5T32GM008490-19.

Tables

Table 1 Primers used in this study

Primer name	**Sequence**
5’surT for	5’cgccgtcggattgcgttccgaagcg3’
surT *Eco*RI del for	5’agtatgggaaagaatttgcctgcgcgcagaagatggc3’
surT *Eco*RI del rev	5’ccgccatcttctgcgcgcaggcaaattctttcccatactttg3’
5’surP rev	5’gcgacgcgttcgtaatccatgggcgaacccctctc3’
5’agaN for	5’gacggaggacaagccatggcaattgtatttgatcc3’
3’agaN rev	5’actagtgcctagccgcatgctagacacc3’

Table 2 Plasmids used in this study

Plasmid name	Description	Source
pNW33N	*Geobacillus* vector	BGSC, ECE136
pIMBB	BioBrick accepting vector for PCR cloning and assembly	[13]
pIM1638 (*GFP*-pIMBB)	Accepting vector for *surT* and *agaN* fragments	This study
pIM241 (*agaN*-pIMBB)	Source of *agaN* insert for pIM472	This study
pIM472 (*surT*-P$_{\mathrm{surP}}$-*agaN*-pIMBB)	*surT*-P$_{\mathrm{surP}}$-*agaN* in pIMBB plasmid	This study
pIM1708 (*surT*-P$_{\mathrm{surP}}$-*agaN*-pNW33N)	Inducible alpha galactosidase construct for GsNUB3621	This study
P$_{\mathrm{ldh}}$-*sfGFP*-pMK-RQ	GeneArt synthesized plasmid containing sfGFP	This study
P$_{\mathrm{RHIII}}$-pIDTSmart	IDT minigene construct containing RHIII promoter	This study
P$_{\mathrm{RHIII}}$-*sfGFP*-pIDTSmart	*sfGFP* cloned downstream of PRHIII promoter	This study
pIM1773 (P$_{\mathrm{RHIII}}$-*sfGFP*-pNW33N)	Constitutive sfGFP construct for GsNUB3621	This study

Table 3 Contigs in GsNUB3621 draft genome

Contig number	Length	Start[a]	Stop	Estimated gap[b]
1	159532 bp	105368	267177	17644 bp
2	615890 bp	284821	996044	3849 bp
3	411492 bp	999893	1424910	34581 bp
4	537992 bp	1458591	1812705	8684 bp
5	361973 bp	1821389	2055928	16933 bp
6	831535 bp	2072861	2838732	5278 bp
7	14434 bp	2844010	2847735	25660 bp
8	103904 bp	2873395	2969136	1853 bp
9	441948 bp	2970989	3415682	166 bp
10	142685 bp	3415848	91412	13956 bp

[a] The start and stop values represent the approximate mapping locations in *Geobacillus sp*. WCH70. A start value of 105368 means that the most 5’ BLAST hit matched the region starting at base pair 105368 in *Geobacillus sp.* WCH70.

[b] Gap estimate is for the gap following each contig.

Figures

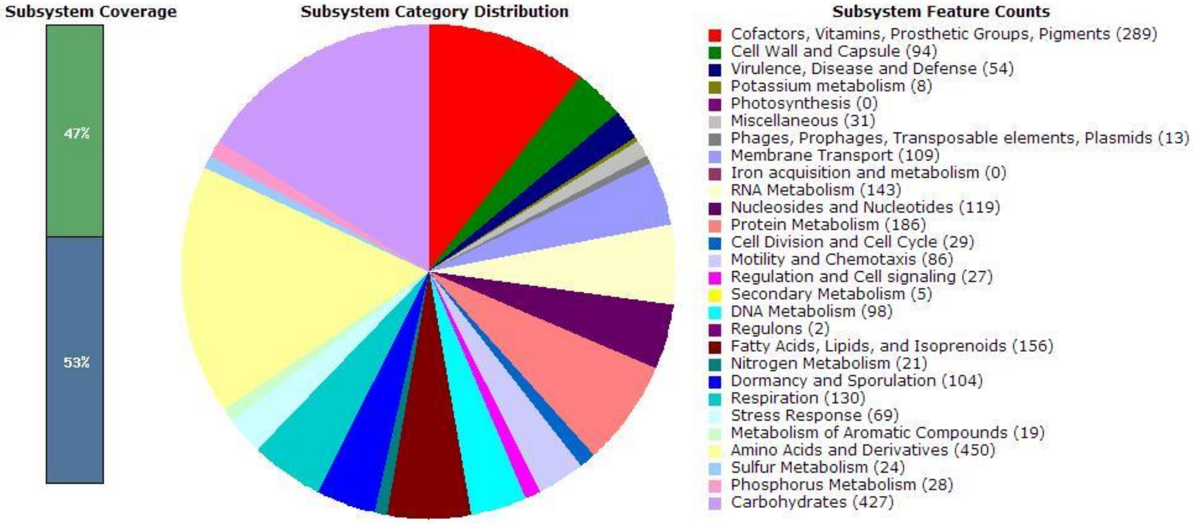

Fig. 1 The RAST server [25] was used to annotate the open reading frames of GsNUB3621. Of the 3832 features that were identified, 47% fell into known gene categories. The distribution of their functions is shown.

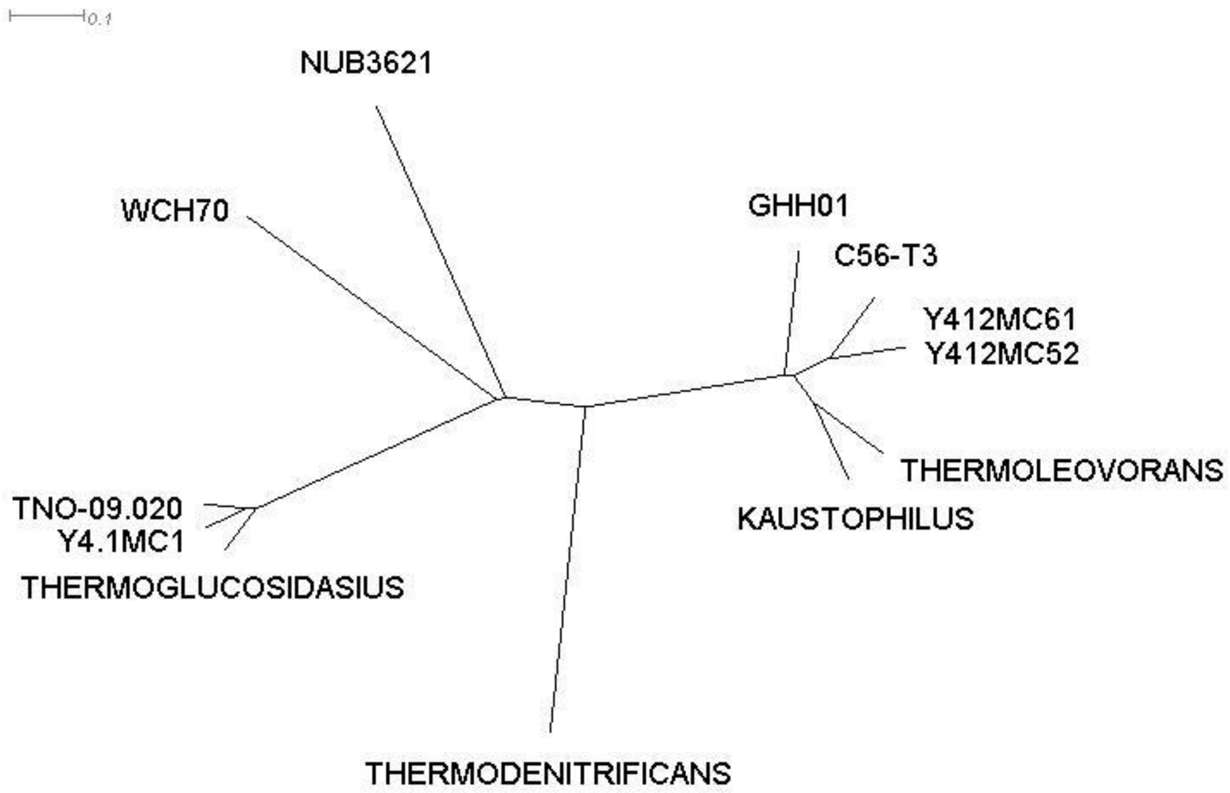

Fig. 2 The Maximal Unique Matches index [31] was used to compare the GsNUB3621 genome sequence with those of other *Geobacillus* species (namely, *Geobacillus sp.* WCH70, *Geobacillus thermoglucosidasius* C56-YS93, *Geobacillus kaustophilus* HTA426, *Geobacillus thermodenitrificans* NG80-2, *Geobacillus thermoleovorans* CCB_US3_UF5, *Geobacillus sp.* Y4.1MC1, *Geobacillus sp.* Y412MC52, *Geobacillus sp.* Y412MC61, *Geobacillus sp.* C56-T3, *Geobacillus sp.* GHH01, and *Geobacillus thermoglucosidans* TNO-09.020). The phylogenetic separation between *Geobacillus sp.* Y412MC52 and *Geobacillus sp.* Y412MC61 can only be detected at higher resolutions.

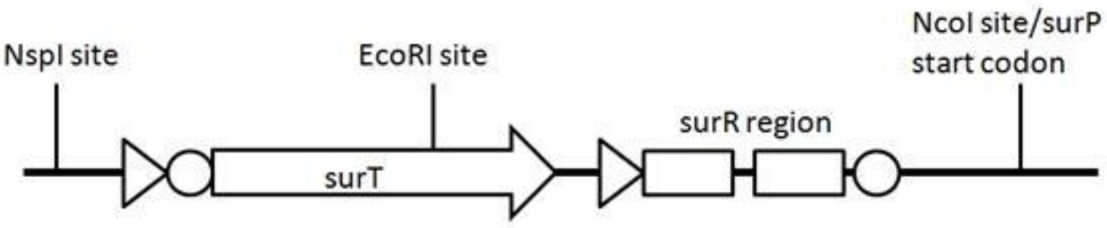

Fig. 3 The regulatory region of the sucrose utilization operon is schematized. Triangles denote promoters while circles represent Shine-Dalgarno sequences. The *surT* gene encodes an antiterminator that is thought to bind the *surR* region, allowing transcription from the *surP* promoter. The region was amplified via PCR; the *Eco*RI site within surT was incompatible with the BioBrick cloning standard, so it was eliminated by site-directed mutagenesis of a single base to effect a synonymous substitution. An NcoI site was created at the *surP* start codon; the region was cloned into the *E. coli*/GsNUB3621 shuttle plasmid pNW33N with restriction enzymes *Nco*I, *Nsp*I and *Sph*I.

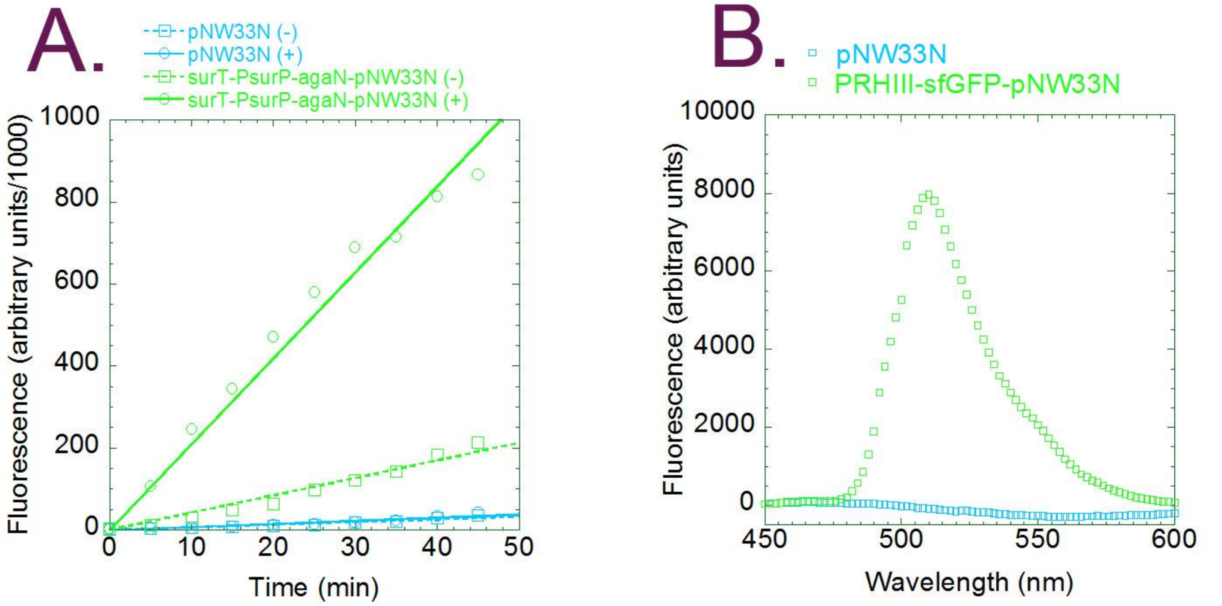

Fig. 4 The inducible *surP* promoter (A) and constitutive ribonuclease HIII promoter (B) were used to express reporter proteins alpha-galactosidase (AgaN) and superfolding Green Fluorescent Protein (sfGFP), respectively. A.) GsNUB3621 were transformed with the empty *E. coli*/GsNUB3621 shuttle vector pNW33N (blue) or *surT*-P$_{\mathrm{surP}}$-*agaN*-pNW33N (green). The transformants were propagated overnight in modified LB medium, in the presence (solid lines) or absence (dotted lines) of sucrose. The supernatants were reacted with 4-methylumbelliferyl-alpha-D-galactopyranoside. Aliquots were quenched in sodium hydroxide; the alpha-galactosidase activity (increase in fluorescence over time) was measured in a spectrofluorimeter. B.) GsNUB3621 was transformed with the empty pNW33N vector (blue) or expression vector P$_{\mathrm{RHIII}}$-*sfGFP*-pNW33N (green). The transformants were propagated overnight in mLB medium. The cells were harvested by centrifugation, resuspended in buffer and lysed by lysozyme-catalyzed hydrolysis of their cell walls. The fluorescence spectra were measured; the values were adjusted by subtracting the fluorescence of a blank (fresh mLB).

References

1. Jaenicke, R., *Stability and stabilization of globular proteins in solution.* J Biotechnol, 2000. **79**(3): p. 193-203.

2. Baneyx, F. and M. Mujacic, *Recombinant protein folding and misfolding in Escherichia coli.* Nat Biotechnol, 2004. **22**(11): p. 1399-408.

3. Woolston, B.M., S. Edgar, and G. Stephanopoulos, *Metabolic Engineering: Past and Future.* Annu Rev Chem Biomol Eng, 2013.

4. Wu, L. and N.E. Welker, *Temperature-induced protein synthesis in Bacillus stearothermophilus NUB36.* J Bacteriol, 1991. **173**(15): p. 4889-92.

5. Wu, L.J. and N.E. Welker, *Protoplast transformation of Bacillus stearothermophilus NUB36 by plasmid DNA.* J Gen Microbiol, 1989. **135**(5): p. 1315-24.

6. Chen, Z.F., S.F. Wojcik, and N.E. Welker, *Genetic analysis of Bacillus stearothermophilus by protoplast fusion.* J Bacteriol, 1986. **165**(3): p. 994-1001.

7. Vallier, H. and N.E. Welker, *Genetic map of the Bacillus stearothermophilus NUB36 chromosome.* J Bacteriol, 1990. **172**(2): p. 793-801.

8. Taylor, M.P., C.D. Esteban, and D.J. Leak, *Development of a versatile shuttle vector for gene expression in Geobacillus spp.* Plasmid, 2008. **60**(1): p. 45-52.

9. Suzuki, H. and K. Yoshida, *Genetic transformation of Geobacillus kaustophilus HTA426 by conjugative transfer of host-mimicking plasmids.* J Microbiol Biotechnol, 2012. **22**(9): p. 1279-87.

10. Counago, R., S. Chen, and Y. Shamoo, *In vivo molecular evolution reveals biophysical origins of organismal fitness.* Mol Cell, 2006. **22**(4): p. 441-9.

11. Pena, M.I., et al., *Evolutionary fates within a microbial population highlight an essential role for protein folding during natural selection.* Mol Syst Biol, 2010. **6**: p. 387.

12. Giver, L., et al., *Directed evolution of a thermostable esterase.* Proc Natl Acad Sci U S A, 1998. **95**(22): p. 12809-13.

13. Bryksin, A.V. and I. Matsumura, *Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids.* Biotechniques, 2010. **48**(6): p. 463-5.

14. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* Genome Res, 2009. **19**(6): p. 1117-23.

15. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

16. Goto, K., et al., *Application of the partial 16S rDNA sequence as an index for rapid identification of species in the genus Bacillus.* J Gen Appl Microbiol, 2000. **46**(1): p. 1-8.

17. Kurtz, S., et al., *Versatile and open software for comparing large genomes.* Genome Biol, 2004. **5**(2): p. R12.

18. Deloger, M., M. El Karoui, and M.A. Petit, *A genomic distance based on MUM indicates discontinuity between most bacterial species and genera.* J Bacteriol, 2009. **191**(1): p. 91-9.

19. Huson, D.H. and D. Bryant, *Application of phylogenetic networks in evolutionary studies.* Mol Biol Evol, 2006. **23**(2): p. 254-67.

20. Nakamura, Y., T. Gojobori, and T. Ikemura, *Codon usage tabulated from international DNA sequence databases: status for the year 2000.* Nucleic Acids Res, 2000. **28**(1): p. 292.

21. Beutler, E. and W. Kuhl, *Purification and properties of human alpha-galactosidases.* J Biol Chem, 1972. **247**(22): p. 7195-200.

22. Talbot, G. and J. Sygusch, *Purification and characterization of thermostable beta-mannanase and alpha-galactosidase from Bacillus stearothermophilus.* Appl Environ Microbiol, 1990. **56**(11): p. 3505-10.

23. Counago, R. and Y. Shamoo, *Gene replacement of adenylate kinase in the gram-positive thermophile Geobacillus stearothermophilus disrupts adenine nucleotide homeostasis and reduces cell viability.* Extremophiles, 2005. **9**(2): p. 135-44.

24. Chain, P.S., et al., *Genomics. Genome project standards in a new era of sequencing.* Science, 2009. **326**(5950): p. 236-7.

25. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC Genomics, 2008. **9**: p. 75.

26. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Res, 2005. **33**(17): p. 5691-702.

27. Takami, H., et al., *Genomic characterization of thermophilic Geobacillus species isolated from the deepest sea mud of the Mariana Trench.* Extremophiles, 2004. **8**(5): p. 351-6.

28. Takami, H., et al., *Thermoadaptation trait revealed by the genome sequence of thermophilic Geobacillus kaustophilus.* Nucleic Acids Res, 2004. **32**(21): p. 6292-303.

29. Studholme, D.J., R.A. Jackson, and D.J. Leak, *Phylogenetic analysis of transformable strains of thermophilic Bacillus species.* FEMS Microbiol Lett, 1999. **172**(1): p. 85-90.

30. Zeigler, D.R., *Application of a recN sequence similarity analysis to the identification of species within the bacterial genus Geobacillus.* Int J Syst Evol Microbiol, 2005. **55**(Pt 3): p. 1171-9.

31. Coorevits, A., et al., *Taxonomic revision of the genus Geobacillus: emendation of Geobacillus, G. stearothermophilus, G. jurassicus, G. toebii, G. thermodenitrificans and G. thermoglucosidans (nom. corrig., formerly 'thermoglucosidasius'); transfer of Bacillus thermantarcticus to the genus as G. thermantarcticus comb. nov.; proposal of Caldibacillus debilis gen. nov., comb. nov.; transfer of G. tepidamans to Anoxybacillus as A. tepidamans comb. nov.; and proposal of Anoxybacillus caldiproteolyticus sp. nov.* Int J Syst Evol Microbiol, 2011. **62**(Pt 7): p. 1470-85.

32. Feng, L., et al., *Genome and proteome of long-chain alkane degrading Geobacillus thermodenitrificans NG80-2 isolated from a deep-subsurface oil reservoir.* Proc Natl Acad Sci U S A, 2007. **104**(13): p. 5602-7.

33. Zhao, Y., et al., *Complete genome sequence of Geobacillus thermoglucosidans TNO-09.020, a thermophilic sporeformer associated with a dairy-processing environment.* J Bacteriol, 2012. **194**(15): p. 4118.

34. De Rossi, E., et al., *New shuttle vector for cloning in Bacillus stearothermophilus.* Res Microbiol, 1994. **145**(8): p. 579-83.

35. Shetty, R.P., D. Endy, and T.F. Knight, Jr., *Engineering BioBrick vectors from BioBrick parts.* J Biol Eng, 2008. **2**: p. 5.

36. Li, Y. and T. Ferenci, *Gene organisation and regulatory sequences in the sucrose utilisation cluster of Bacillus stearothermophilus NUB36.* Gene, 1997. **195**(2): p. 195-200.

37. Fridjonsson, O., et al., *Thermostable alpha-galactosidase from Bacillus stearothermophilus NUB3621: cloning, sequencing and characterization.* FEMS Microbiol Lett, 1999. **176**(1): p. 147-53.

38. Pedelacq, J.D., et al., *Engineering and characterization of a superfolder green fluorescent protein.* Nat Biotechnol, 2006. **24**(1): p. 79-88.

39. Bujard, H., et al., *A T5 promoter-based transcription-translation system for the analysis of proteins in vitro and in vivo.* Methods Enzymol, 1987. **155**: p. 416-33.

40. Studier, F.W. and B.A. Moffatt, *Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes.* J Mol Biol, 1986. **189**(1): p. 113-30.

41. de Boer, H.A., L.J. Comstock, and M. Vasser, *The tac promoter: a functional hybrid derived from the trp and lac promoters.* Proc Natl Acad Sci U S A, 1983. **80**(1): p. 21-5.

42. Yansura, D.G. and D.J. Henner, *Use of the Escherichia coli lac repressor and operator to control gene expression in Bacillus subtilis.* Proc Natl Acad Sci U S A, 1984. **81**(2): p. 439-43.

43. Bloom, J.D., et al., *Protein stability promotes evolvability*. Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5869-74.

Chapter 3: Discussion

Summary of contribution to the field

Metabolic engineering greatly benefits from the availability of diverse laboratory organisms. The properties of a host strain will often determine whether or not the desired product can be synthesized, so increasing the pool of well-characterized and genetically pliable host strains increases the probability of finding a strain amenable to synthesis of a desired product. Additionally, if a host strain lacks a necessary pathway or enzyme for the desired synthesis, heterologous expression of proteins or pathways from another strain can be utilized. Thus, identifying organisms with unique properties increases the pool of enzymes and pathways that can be coopted. Finally, if no known enzymes are suitable for a desired synthesis scheme, often existing proteins can be evolved to fit desired parameters. For this reason, strains that are highly evolvable or are excellent vectors for enzyme evolution are highly desirable.

The genus *Geobacillus* is attractive for metabolic engineering for all of these reasons. The *Geobacillus* species' wide range of growth temperatures makes them capable of expressing both mesophilic and thermophilic proteins. Because *Geobacillus* is a member of the diverse *Bacillus* superfamily, it is possible that enzymes from other Bacillus species will be more likely to properly fold in *Geobacillus* species than in more distantly related organisms like *E. coli*. Additionally, recent research suggests that *Geobacillus* species may be able to express some proteins that are insoluble in *E. coli*. Thus, *Geobacillus* species are excellent candidates for host strains for metabolic engineering.

Geobacillus species are also useful host strains for enzyme evolution, which can be used to modify proteins to improve synthesis designs. Thermostability is a desired trait for metabolic engineering schemes because growing host organisms at elevated temperatures can reduce the risk of microbial contamination by mesophilic species. Additionally, for synthesis of products such as ethanol, by growing the host organism near the vaporization point of the desired product, the product is quickly removed from the environment, easing its extraction and reducing its toxicity for the host organism. Although *in vitro* methods for evolving thermostable enzymes exist, these techniques often involve producing the enzyme in a mesophilic host, then increasing the temperature and screening for enzymes that retain their function. Thus, using these methods, it is only possible to screen for enzymes that *function* at high temperatures, but not whether the enzymes can properly *fold* at elevated temperatures. *In vitro* selection of thermostable enzymes allows for the selection of proteins that can both properly fold and function at elevated temperatures, but requires a host strain that can thrive at both mesophilic and thermophilic temperatures. Thus, *Geobacillus* species fit a unique niche for enzyme evolution.

Geobacillus stearothermophilus NUB3621 is a particularly attractive strain of *Geobacillus*, but its utility has been hampered by a lack of tools; the work described here remedies this situation and improves the utility of this unique strain. NUB3621 fortuitously lacks a restriction modification system, meaning it is likely to have a higher transformation efficiency than other *Geobacillus* strains regardless of the method of transformation. Enzyme evolution generally requires the screening and selection of large libraries of the desired enzyme; high transformation efficiency is crucial to constructing and screening these libraries. The tools described here greatly improve the utility of GsNUB3621. The annotated genome sequence is crucial to GsNUB3621's viability as a host strain for product synthesis. Engineering a cell to produce a desired compound generally requires at least a basic understanding of the host metabolome, thus the genome sequence and annotation provides a first step at understanding the GsNUB3621 metabolome. The expression constructs developed here allow heterologous genes to be manipulated and expressed in GsNUB3621, which is necessary for enzyme evolution or for utilizing GsNUB3621 as a cell factory.

Further steps to improve GsNUB3621's utility

Although the work here represents a stride forward for the use of GsNUB3621, there are many steps that can be taken to further improve the utility of GsNUB3621 and *Geobacillus* species in general. The work described here utilized protoplast transformation to introduce new genetic material into GsNUB3621. Protoplast transformation is more labor intensive than other methods such as electroporation or heat shock, and has likely served as a barrier for groups that may have otherwise considered utilizing GsNUB3621. Although other labs have reported successful electroporation of *Geobacillus* strains, the transformation efficiency is still lower than that reported by the Welker lab for GsNUB3621, and electroporation could not be replicated in our hands for GsNUB3621. Fortunately, other labs have recently investigated the possibility of conjugation as a method of introducing genetic material into *Geobacillus* species. [1] It would be interesting to see if these techniques could be applied successfully to

GsNUB3621, which is still likely to have a higher efficiency, as foreign DNA will not be degraded by host restriction enzymes.

The utility of GsNUB3621 could also be improved by fully mapping its metabolic network. Although the genome annotation described here provides insight into which pathways and enzymes are present in GsNUB3621, we currently lack an understanding of how these enzymes and pathways interact within the metabolome, and can only conjecture by comparing with metabolic networks from other species. A complete metabolic map of GsNUB3621 would greatly benefit anyone hoping to coopt GsNUB3621 to produce a desired compound. [2]

Projects involving GsNUB36

GsNUB3621 as a source of evolvable proteins

Proteins must fold correctly in order to function. Since most proteins are only marginally stable, protein stability can be a limiting factor when altering a protein's activity. Since many mutations that alter a protein's function can also be destabilizing, often protein evolution and protein stability can come at a tradeoff. Intuitively then, proteins with a higher initial stability can tolerate more mutations, and thus have more evolutionary potential, than proteins that are only marginally stable. Bloom et al demonstrated this principle experimentally by mutating variants of cytochrome P450 BM3. [3] It was found that only the thermostable variants of P450, and not its marginally stable relatives, could tolerate mutations that conferred a change in function. Thus, his results suggest that by utilizing thermostable variants as a starting point for evolution, proteins will have more evolutionary potential than marginally stable counterparts.

GsNUB3621 is likely to be both a source of highly stable proteins, and a host to evolve stable proteins. Because GsNUB3621 can survive at both mesophilic temperatures and thermophilic temperatures, all of its essential proteins must be able to properly fold and function over a wide range of temperatures. Though this premise is suggestive of the fact that GsNUB3621's proteins are more than marginally stable, it would be interesting to test this hypothesis experimentally. The Bloom experiment used thermostability as an indicator of mutational robustness; the hypothesis could be strengthened by testing the mutational robustness, or number of mutations tolerated before the protein unfolds, of proteins from a thermotolerant strain like GsNUB3621. Specifically, it would be beneficial to randomly mutate proteins from both *E. coli* and GsNUB3621 and measure the resulting protein stability (ΔG of unfolding) as the number of mutations increases. If GsNUB3621 is shown to be a strong source of mutationally robust proteins, this result would suggest that its proteins are excellent starting sources for evolution, and that the unique temperature range of GsNUB3621 may make it an excellent host to select for mutationally robust variants of heterologous proteins.

Evolution of a thermostable genome

How do whole genomes evolve? This question is important for understanding how organisms have evolved in the past, how ecological species might adapt to environmental challenges today, or even how pathogens can rapidly evolve in nature, such as developing resistance to antibiotics. Unfortunately, evolution on a whole genome level is difficult to monitor in the lab as changes are unlikely to be seen in a reasonable time frame. Even relatively "simple" genomes like that of *E. coli* are millions of base pairs long and contain thousands of genes that encode thousands of proteins. A simpler model, such as the genome of a phage, represents a more feasible starting strain to monitor changes that occur in whole genome evolution.

GsNUB3621 could potentially be used to evolve a mesophilic phage to become thermophilic. By infecting GsNUB3621 with a mesophilic phage and slowly increasing the growth temperature of GsNUB3621, it should be possible to select for changes in the phage that allow it to withstand high temperatures. Although there are many other Geobacillus species that grow at the same temperature as GsNUB3621, this organism is the strongest candidate for a host strain for phage evolution due to its lack of a restriction modification system. In strains with a restriction modification system, infecting phage DNA can be digested by microbial restriction enzymes, leading to lower rates of successful phage infection, thereby limiting the phage's evolution.

Final remarks

GsNUB3621 is an attractive vehicle for metabolic engineering and evolutionary studies. Although there are numerous thermostable microorganisms that are utilized in the lab, few bacteria can survive at both mesophilic and thermophilic temperatures. GsNUB3621 is especially unique in this regard as it also lacks a restriction modification system, meaning it is likely to be innately more transformable than other Geobacillus species. In this dissertation I have improved the utility of GsNUB3621 by sequencing and annotating its genome. The genome sequence provides insight into the proteome of GsNUB3621 which is crucial for engineering the strain to produce desired compounds. I have developed two expression constructs, one inducible and one constitutive that can be used to express heterologous genes in GsNUB3621. This work thus helps to validate the utility of GsNUB3621 as a beneficial strain for metabolic engineering.

References

1. Tominaga, Y., T. Ohshiro, and H. Suzuki, *Conjugative plasmid transfer from Escherichia coli is a versatile approach for genetic transformation of thermophilic Bacillus and Geobacillus species.* Extremophiles, 2016. **20**(3): p. 375-81.

2. Merlet, B., et al., *A Computational Solution to Automatically Map Metabolite Libraries in the Context of Genome Scale Metabolic Networks.* Front Mol Biosci, 2016. **3**: p. 2.

3. Bloom, J.D., et al., *Protein stability promotes evolvability.* Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5869-74.

Appendix 1: *E. coli* chromosomal evolution

Introduction

Metabolic engineering often requires introducing multiple enzymes and enzymatic reactions into a nonnative host. Predicting how these enzymes and enzymatic activities will affect and be influenced by the host metabolome is extremely difficult. Studies have revealed that the *E. coli* metabolome contains a wealth of previously underestimated flexibility; many native enzymes have catalytic promiscuity, substrate ambiguity, or other ways in which they can interact with substrates and perform enzymatic reactions other than their canonical function. [1] Appreciating the complexity of the host metabolome is crucial for metabolic engineering, thus it is important to study the potential flexibility of common host metabolomes like *E. coli*.

Our lab previously investigated the flexibility of the *E. coli* proteome. [1] Wayne Patrick et al discovered that many proteins within the *E. coli* proteome have secondary effects (often catalytic promiscuity or catalytic ambiguity) and can rescue auxotrophic knockouts of unrelated genes. Specifically, he was able to rescue 21 out of 104 auxotrophs. One fair criticism of his scheme is that it involved the use of high copy number plasmids expressing genes under very high induction conditions. This criticism raises the important question of whether or not these secondary effects can still be observed with proteins under their native expression condition, as would be seen in evolution in nature. Thus, the goal of this experiment was to determine how much flexibility is present within the *E. coli* proteome under conditions of native expression.

Methods

Mutagenesis

Auxotrophic strains from the KEIO collection were mutagenized using 2-amino purine (2AP). [2] Cultures were grown overnight in LB medium at 37°C. 50 uL of a 10^{-6} dilution of the overnight culture was used to inoculate a fresh culture of LB with 700 ug/mL of 2AP. This LB 2AP culture was then grown for approximately 24 hours then washed with M9 medium to remove any remaining LB or 2AP. 100 uL was plated on M9-glucose plates to screen for rescued auxotrophs, and diluted samples were plated on LB plates to check for rate of survival of the 2AP treatment. Plates were grown at 37°C and left until colonies formed.

Transduction

Transduction was first used to generate a clean genetic background for each of the surviving strains. To generate this clean genetic background, the kanamycin marker from the corresponding KEIO strain was transduced into fresh K12 cells. For example, to create a knockout of pdxJ in a K12 background, a lysate was generated from ∆pdxJ in the KEIO collection. This lysate was then used to transduce fresh K12 cells, selecting for kanamycin resistance. All transductions were done in high throughput. To generate the lysate, overnight LB-kanamycin cultures were grown to saturation. 200 uL of lysate medium (LB with 0.2% glucose and 5 mM $CaCl_2$) were added to the wells of a 96-well plate. This culture was innoculated with 10 uL of the overnight saturated culture and grown for 37°C with shaking. 20 uL of P1 dam rev6 phage was then added to each culture and the plate was incubated at 37°C with shaking. Lysis was monitored by measuring optical density (OD) over time. Phage was harvested when lysis appeared complete (approximately two to three hours of incubation) when the OD stopped declining. $CHCl_3$ was added to kill the phage, and lysates were stored at 4°C until use.

To transduce, an overnight culture of wildtype K12 was grown in LB at 37°C with shaking. In a 96 well plate, 250 uL of transduction medium (LB with 5 mM $CaCl_2$ and 10 mM $MgSO_4$) was added to each well utilized. 50 uL of the saturated K12 culture and 25 uL of appropriate lysate were added to the medium and incubated for 20 minutes at 37°C with shaking. Cells were centrifuged for 5 minutes at 5000 RPM and resuspended in LB with 5 mM sodium citrate. Cultures were incubated for one hour at 37°C in a water bath to allow for expression of the kanamycin resistance gene, plated on LB kanamycin, and grown overnight at 37°C.

To introduce the causative mutation into these K12 knockouts, the same procedure was followed. Lysates were generated from the rescued KEIO strains. These lysates were used to transduce the corresponding K12 knockouts (for example, the lysate from the rescued ΔpdxJ KEIO was added to ΔpdxJ K12.) For the transduction, the same procedure was followed, except that after incubating for 20 minutes, cultures were washed with 300 uL transduction media and resuspended in 200 uL of transduction media to remove any LB, and the preincubation step was skipped. Cultures were plated on minimal media to select for rescues, and the winners were restreaked on LB kanamycin.

Sequence analysis

Genomic DNA was prepared using Qiagen's DNeasy kit. The samples were sequenced by Illumina next generation sequencing. The reads were aligned to the Escherichia coli K-12 MG1655 (NCBI 2001-10-15) reference sequence using the BWA

Aligner through Illumina's Basespace. Reads that did not pass Illumina's chastity filter were removed before alignment. Variation from the reference sequence was called using samtools. [3] To verify that the kanamycin gene was in the expected region for each strain, a *de novo* assembly was generated for each sample. Assemblies were created using ABYSS with a kmer length of 88. [4] The location of the kanamycin resistance gene was then determined using BLAST to query the contigs for a match to the kanamycin resistance gene. [5] The location of the kanamycin gene was verified further by viewing the alignments in IGV to check for an area where no reads map, which should correspond to the kanamycin resistance gene. [6]

Results

Of the 104 auxotrophs mutagenized, 54 were rescued and restreaked on M9 medium. Of the 21 auxotrophs rescued in the Patrick *et al* study, 19 were also rescued in this experiment (Table 2.) Though the number of days it took the rescued strains to form colonies differed in this experiment compared with Patrick *et al*, there was no obvious overall pattern of whether the rescued strains grew faster with mutagenesis or with multicopy suppression.

To identify the causal mutations, we used transduction to try and introduce the causative mutation into a clean genetic background. The kanamycin resistance marker for each rescued strain was first introduced into wildtype K12 *E. coli* to create a new knockout. For example, a K12 knockout of serB was created by transducing wildtype K12 *E. coli* with P1 phage lysate from ΔserB from the KEIO collection. In this way, any unidentified mutations within the KEIO parent strain would be eliminated in further analysis. These new K12 knockouts were then transduced using lysate from each of the respective rescued strains. For example, the rescued ΔserB from the ASKA collection was used to generate a P1 lysate that was then used to infect the ΔserB K12 strain. Transductants could then be selected for growth on M9 medium. Of the 54 K12 knockout strains, only 18 could be rescued on M9 medium through transduction from the original rescued KEIO strains.

Of these 18 strains rescued through transduction, eight strains (pdxJ, purA, purL, purM, pyrB, proA, aroE, metA) were chosen for next generation sequencing. These strains were chosen because they were unable to be rescued in previous studies. Unfortunately, six of the eight samples appeared to be the incorrect strain. Five samples possessed the kanamycin gene in place of the slt gene, and one sample possessed the kanamycin gene in place of the fdoG gene. For the remaining two samples, the kanamycin gene mapped to the metA location and the aroE location respectively. In the metA sequence there was only one mutation identified (a change from AC to ACGC at base pair 4294403, a noncoding region between gltP and yjcO), however, this mutation was present in all eight strains, suggesting it was not causative for the rescue. For the aroE sequence, there was one unique mutation, a frameshift mutation in the yagA gene.

Discussion

The results of this study suggest that the flexibility in the *E. coli* chromosome may be greater than previously thought. The fact that 54 of the 104 auxotrophs could be rescued supports the work of Patrick *et al* and suggests that their results were not merely a consequence of high copy number plasmids and high expression conditions. [1]

Additionally, because more rescued strains were obtained via mutagenesis than by multicopy suppression, these results suggest that other mutations, instead of merely increases in protein expression, may help the proteome overcome auxotrophies.

The contamination of samples with ΔfdoG and Δslt cells represents an important limitation of this work. Because neither ΔfdoG nor Δslt are auxotrophs, both should grow normally on M9 medium. Thus, it is likely that the contamination occurred late in the experiment, possibly around the transduction step. If the contamination had occurred early or throughout the initial mutagenesis, it is likely that cell growth would have appeared much faster on the M9 plates. If strains were contaminated prior to transduction, then the selection on M9 should have yielded lawns, which was not the case for any strain. Additionally, the identity of rescued strains could be verified with PCR by using primers designed to anneal to the kanamycin gene and the region upstream. If the kanamycin gene were not in its expected location, the primers would not be close enough to allow for amplification. This technique was used to verify that the ΔproB strain rescued was truly ΔproB, but primers were not designed for all 54 strains.

There are currently no published studies on the role of yagA in *E. coli*, though it is predicted to be a DNA-binding transcriptional regulator. [7] Thus, it would be interesting to further elucidate the relationship between yagA and aroE. The first and simplest step would be to determine the phenotype of a yagA and aroE double knockout to verify if the identified frameshift mutation was able to rescue aroE due to the loss of function of yagA. Because yagA is predicted to be a regulatory protein, it would also be interesting to see if it affects other genes, especially those associated with other auxotrophies. Introducing a yagA deletion into other auxotrophic strains and checking for growth on

M9 could help to reveal whether yagA's regulatory effects are related to any other pathways besides that of chorismate synthesis.

Studying the flexibility of the *E. coli* proteome can lead to important insights for a variety of fields. The goal of the Patrick *et al* study was to help elucidate mechanisms by which new metabolic pathways evolve. As the results of this study suggest, adaptation and metabolic evolution can occur even without high expression of the evolving protein. These evolutionary questions are also important for the field of metabolic engineering, where a desired reaction must coexist with the cell's native metabolome. Understanding the inherent flexibility and evolvability of the host proteome can be helpful to ensuring that the desired reaction can persist and flourish for many generations.

Strain	Deficiency	Affected pathway	Days	Colony #
Δ*metC*	Cystathionine β-lyase	Methionine Biosynthesis	3	3
Δ*glnA*	Glutamine synthetase	Glutamine Biosynthesis	4	1
Δ*ilvE*	Branched-chain amino acid transferase	ILV Biosynthesis	3	1
Δ*carA*	carbamoyl phosphate synthetase	Arginine and pyrymidine Biosynthesis	2	>100
Δ*purF*	Amidophosphoribosyl transferase	Purine Biosynthesis	2	5
Δ*ppc*	phosphoenolpyruvate carboxylase	Fermentation	4	>100
Δ*ilvA*	Threonine deaminase	ILV Biosynthesis	3	>100
Δ*ptsI*	phosphoenolpyruvate-protein phosphotransferase	Carbohydrate transport	3	>100
Δ*serB*	Phosphoserine phosphatase	Serine biosynthesis	3	>100
Δ*hisH*	Imidazole glycerol phosphate synthase	Histidine biosynthesis	3	50
Δ*glyA*	Serine hydroxymethyltransferase	Glycine biosynthesis	3	>100
Δ*pabB*	para-aminobenzoate synthase component I	Folic acid biosynthesis	12	1
Δ*pabA*	para-aminobenzoate synthase component II	Folic acid biosynthesis	2	>100
Δ*metR*	Methionine DNA-binding transcriptional activator	Methionine biosynthesis	7	13
Δ*yhhK*	Panthothenate	Cofactor biosynthesis	3	
Δ*pdxB*	Erythronate 4-phosphate dehydrogenase	Pyridoxine biosynthesis	3	>100
Δ*fes*	ferric enterochelin esterase	Incorporation of metal ions	2	>100
Δ*serA*	D-3-phosphoglycerate dehydrogenase	Serine biosynthesis	6	2
Δ*serC*	Phosphoserine aminotransferase	Serine biosynthesis	12	2
Δ*metL*	Aspartate kinase II / homoserine dehydrogenase II	Homoserine/lysine biosynthesis	2	>100
Δ*thrA*	Aspartate kinase I / homoserine dehydrogenase I	Homoserine/lysine biosynthesis	2	>100
Δ*gltA*	citrate synthase	Anaerboic respiration	3	2
Δ*leuL*	Leu operon attenuation peptide	Leucine biosynthesis	4	>100
Δ*proB*	Glutamate-5-kinase	Proline biosynthesis	4	>100
Δ*tyrA*	chorismate mutase / prephenate dehydrogenase	Tyrosine/phenylalanine biosynthesis	5	>100
Δ*bioB*	Biotin synthase	Biotin	2	1
Δ*panC*	Pantothenate synthetase	Coenzyme A biosynthesis	2	>100
Δ*pdxA*	4-hydroxy-L-threonine phosphate dehydrogenase	Pyridoxine biosynthesis	2	3
Δ*bioD*	Dethiobiotin synthetase	Biotin biosynthesis	2	1
Δ*nadB*	L-aspartate oxidase	NAD biosynthesis	2	2
Δ*bioA*	7,8-diaminopelargonic acid synthase	Biotin biosynthesis	2	2
Δ*pyrF*	Orotidine-5'-phosphate-decarboxylase	Pyrimidine biosynthesis	2	14
Δ*pdxJ*	pyridoxine 5'-phosphate synthase	Pyridoxine biosynthesis	2	2
Δ*purA*	Adenylosuccinate synthetase	Purine biosynthesis	2	8
Δ*purD*	Phosphoribosylamine-glycine ligase	Purine biosynthesis	2	15
Δ*purL*	phosphoribosylformyl-glycineamide synthetase	Purine biosynthesis	2	3
Δ*purM*	phosphoribosylformylglycinamidine cyclo-ligase	Purine biosynthesis	2	2
Δ*pyrB*	aspartate carbamoyltransferase	Pyrimidine biosynthesis	2	2
Δ*pyrE*	Orotate phosphoribosyltransferase	Pyrimidine biosynthesis	2	4
Δ*bioH*	pimeloyl-ACP methyl ester carboxylesterase	Biotin biosynthesis	3	1
Δ*bioF*	8-Amino-7-oxononanoate synthase	Biotin synthesis	4	1
Δ*nadA*	Quinolinate synthase	NAD biosynthesis	4	1
Δ*proA*	Glutamate-semialdehyde dehydrogenase	Proline biosynthesis	4	>100
Δ*trpD*	Anthranilate phosphoribosyl transferase	Tryptophan biosynthesis	4	50
Δ*cysC*	Adenylylsulfate kinase	Sulfur metabolism	5	>100
Δ*cysD*	sulfate adenylyltransferase	Sulfur metabolism	5	>100
Δ*cysH*	3'-Phospho-adenylylsulfate reductase	Sulfur metabolism	5	>100
Δ*cysN*	sulfate adenylyltransferase	Sulfur metabolism	5	>100
Δ*pheA*	chorismate mutase / prephenate dehydratase	Tyrosine/phenylalanine	5	>100
Δ*guaB*	IMP dehydrogenase	Purine biosynthesis	6	2
Δ*metA*	Homoserine *O*-succinyltransferase	Methionine biosynthesis	7	1
Δ*metB*	cystathionine gamma-synthase	Methionine biosynthesis	7	1
Δ*aroE*	Shikimate dehydrogenase	Chorismate biosynthesis	9	2
Δ*purH*	AICAR transformylase/IMP cyclohydrolase	Purine biosynthesis	9	1

Table 1: Results of 2AP mutagenesis. The first column reveals the abbreviation for the gene knockout. The second column is the name of the gene's encoded protein and the third column is the pathway of the protein. The fourth column is the number of days it took for the first colonies to appear after mutagenesis and the last column represents the number of colonies appearing for each strain.

Table 2: Comparison of results with Patrick et al 2007

KEIO Strain	Days (2AP)	ASKA rescue	Days (ASKA)
ΔcarA	2	carB, cho, ygiT, yncK	3,6,6,6
ΔpurF	2	chbA	21
ΔpabA	2	menF, pabB	3,2
Δfes	2	setB, thiL	9,15
ΔmetC	3	alr, fimE, malY	6,10,2
ΔilvE	3	avtA	2
ΔilvA	3	emrD, tdcB	7,3
ΔptsI	3	fucP, galE, xylE	4,6,4
ΔserB	3	gph, hisB, ytjC	5,2,3
ΔhisH	3	hisF	5
ΔglyA	3	ltaE, rsd, tdh, yneH	4,11,8,10
ΔyhhK	3	murI, purF	4,4
ΔpdxB	3	purF, tdh	7,7
ΔglnA	4	asnB	20
Δppc	4	ecfM, yccT	28,27
ΔserA	6	yneH	18
ΔmetR	7	metE	2
ΔpabB	12	menF	4
ΔserC	12	yneH	21

Table 2: Comparison of results with the results of Patrick *et al* 2007. The first column lists strains that were rescued in both this study and the Patrick *et al* study. The second column lists the number of days before colonies appeared following 2AP mutagenesis in this study. The third column lists the genes that were responsible for the rescue in the Patrick *et al* study, and the last column lists the number of days for colonies to form following transformation with the ASKA plasmids in the Patrick *et al* study.

References

1. Patrick, W.M., et al., *Multicopy suppression underpins metabolic evolvability.* Mol Biol Evol, 2007. **24**(12): p. 2716-22.

2. Baba, T., et al., *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.* Molecular Systems Biology, 2006. **2**: p. 2006.0008-2006.0008.

3. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

4. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* Genome Res, 2009. **19**(6): p. 1117-23.

5. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

6. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

7. Keseler, I.M., et al., *EcoCyc: fusing model organism databases with systems biology.* Nucleic Acids Res, 2013. **41**(Database issue): p. D605-12.

Appendix 2: Catalytic modularity of glutamine synthetase

Introduction

Enzymes lower the activation energy of a reaction. In some instances, enzymes catalyze a reaction from substrate to product by first converting the substrate into a high energy intermediate. This high energy intermediate is still of lower free energy than the transition state of an uncatalyzed reaction, so the reaction proceeds faster in the presence of the enzyme. However, these high energy transition states present a different problem in cellular reactions. How can these high energy transitions states be shielded from aberrant catalysis? Understanding strategies for shielding reactive intermediates could allow for greater flexibility in designing metabolic pathways for microbial synthesis of desired products.

The most obvious solution is to shield the intermediate within the active site of the enzyme. This method would require one enzyme to be capable of catalyzing all steps of a reaction. Another possibility is that enzyme complexes have evolved as a solution to this dilemma. Complexes of sequential enzymes in a pathway allow intermediates to be shielded from aberrant catalysis by reducing the distance between two active sites. This strategy of shielding reactive intermediates could potentially be exploited for design of nonnative pathways in a metabolic engineering scheme. Catalytic modularity – or the idea that reactive intermediates generated by one enzyme may be able to be used by a different enzyme – could potentially increase the flexibility of engineering designs. Reactive intermediates in a desired reaction can then be shielded from aberrant catalysis within the host cell, and the intended product can be synthesized.

The relationship between glnA and proB may demonstrate this principle. glnA is the gene for glutamine synthetase, which catalyzes the transition of glutamate into glutamine. The reaction proceeds by forming a phosphorylated intermediate. proB is the gene for γ-glutamyl kinase. It forms a complex with proA, the gene for glutamate-5-semialdehyde dehydrogenase. When in complex with proA, proB phosphorylates glutamate, allowing proA to then form glutamate-5-semialdehyde. Both glnA and proB form the same intermediate. (Figure 1) Thus, it may be possible to use glnA to replace proB. If proA could utilize the reactive intermediate produced by glnA, it would demonstrate a strategy to shield these reactive intermediates in nonnative contexts. This relationship could then serve as a proof of principle for exploiting novel intermediate shielding strategies in microbial product synthesis.

GlnA forms a dodecamer. Each subunit contains an active site where glutamate is first phosphorylated, and then a free ammonia catalyzes the transition into glutamine. The Villafranca lab at Pennsylvania State University did a thorough investigation of each of the active site residues. [1] By mutating several key residues within the active site, they were able to form hypotheses about the roles of each of these residues in the glnA reaction. One residue in particular, residue 327, is of particular interest. They found that by introducing a glutamate to alanine mutation at this site, the activity of glnA against ammonia was impaired, while the activity against glutamate was unchanged. Their data thus suggested that the 327 residue plays an important role in the addition of ammonia to the phosphorylated intermediate, but the residue is not involved in phosphorylating glutamate. Their work was further refined by the Eisenberg lab at UCLA, studying the 3D structure of glnA. [2] Thus the 327 residue is believed to play two roles: it helps recruit the ammonium ion by forming a negatively charged pocket and it helps trap the glutamyl substrate in the active site until the reaction is complete. [3] The 327 residue is therefore an excellent candidate to alter in order to allow glnA to rescue ΔproB.

Methods

Creating E327A mutant

The E327A mutation was created using site directed mutagenesis of the glnA gene on the pCA24N plasmid from the ASKA collection. [4] The forward primer used to introduce the mutation was 5'-gcgtctggtcccgggctatgcggcaccggtaatgctggc and the reverse primer was 5'-gccagcattaccggtgccgcatagcccgggaccagacgc. The primer used to amplify the region upstream of the mutation was 5'-ctcaccatcaccatcaccatacgg and the primer used to amplify the downstream region was 5'-gagtagtgacaagtgttggccatggaacaggtag. An overlap extension PCR was done to combine the fragment upstream and including the E327A mutation and the fragment downstream and including the E327A mutation. The final PCR product was cloned into pCA24N using SfiI. To remove the GFP tag, the plasmid was digested with NotI. The proA gene from proA-pCA24N was cloned into pCDF using SfiI digestion. Both plasmids were transformed into ΔproB from the KEIO collection.

Growth assays

Cells were grown overnight in 5 mL LB cultures with chloramphenicol and spectinomycin. 1 mL of saturated culture was centrifuged and resuspended in equal volume M9 salts, and then centrifuged again and resuspended in M9 medium with glucose and 1 mM IPTG. Innoculated medium was then aliquoted across a 96 well plate in 200 uL volumes per well. The OD 600 was measured for each well every half hour for 48 hours.

Results

The E327A mutation rescued ΔproB, allowing for growth on minimal medium, while the wildtype glnA gene under the same induction levels could not. ΔproB cells were first transformed with ~1 ug of glnA-pCA24N DNA and grown on M9 minimal medium with 1 mM IPTG to induce glnA expression. Under these conditions, after two days only two small colonies appeared. When plasmids from these colonies were recovered and sequenced, they were found to have a stop codon in the 3' GFP tag normally found in glnA-pCA24N. To further improve growth of ΔproB, the GFP tag was more cleanly removed by digesting the plasmid with SfiI, and proA-PCDF was cotransformed to increase the level of proA within the cell. In liquid growth assays, the E327A mutation clearly improved growth over cells expressing wildtype glnA from the pCA24N plasmid (Figure 2.)

The ability of glnA to restore growth in a ΔproB background could be modulated by the amount of ammonia in the medium. The previous experiment was repeated, however instead of using M9 medium with ammonia, asparagine was used as the nitrogen source. In this low ammonia medium, there was no clear difference between the wildtype glnA and glnA with the E327A mutation. (Figure 3.) This result suggested that in the previous experiment, when wildtype glnA synthesizes glutamyl-5-phosphate, it is more susceptible to catalysis by free ammonia in the cell than when the intermediate is synthesized by glnA possessing the E327A mutation.

Discussion

This work serves as a proof of principle for the idea of catalytic modularity in microbial product synthesis. By altering the properties of glnA, or by altering the metabolic environment within the cell, we were able to shield reactive intermediates in novel ways. Thus, glnA can be utilized to synthesize glutamyl-5-phosphate that can then be utilized by proA. Demonstrating this principle is an important step toward developing new strategies to shield reactive intermediates in nonnative environments.

An important next step would be to study the relationship between glnA and proA *in vitro*. Measuring the rate of glutamate-5-semialdehyde formation by proA when glnA provides the substrate rather than proB would further support the hypothesis that the activity of glnA is modular and can be used to form more than just glutamine. Additionally, *in vitro* measurements of the effect of ammonia concentration on the formation of glutamate-5-semialdehyde could further support the hypothesis that reactive molecules within the cell can lead to aberrant catalysis of reactive intermediates.

Another important step would be to see whether the idea of catalytic modularity holds true for enzymes beyond glnA and proB. The first and simplest step would be to see if the glnA intermediate can be utilized by any other enzyme. There are many enzymes that utilize 5-glutamyl-phosphate such as γ-glutamate-cysteine ligase, gshA. [5] If this enzyme can utilize 5-glutamyl-phosphate when produced by glnA, the result would suggest that this pattern is not unique to glnA and proB.

The results of the free ammonia experiment described here suggest that modularity is not solely dependent on enzymatic activity. By modifying the environment within a cell, intermediates can be exposed to other reactive molecules. It may thus be possible to synthesize diverse products by using this modular strategy.

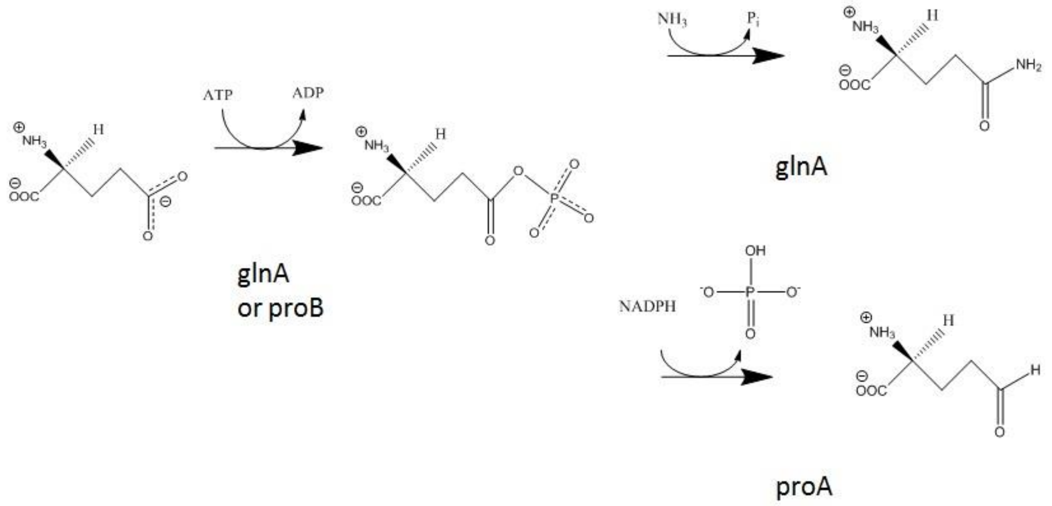

Figure 1: The reactions catalyzed by glnA and the proA-proB complex. The intermediate, 5-glutamyl-phosphate, is the same for both reactions.

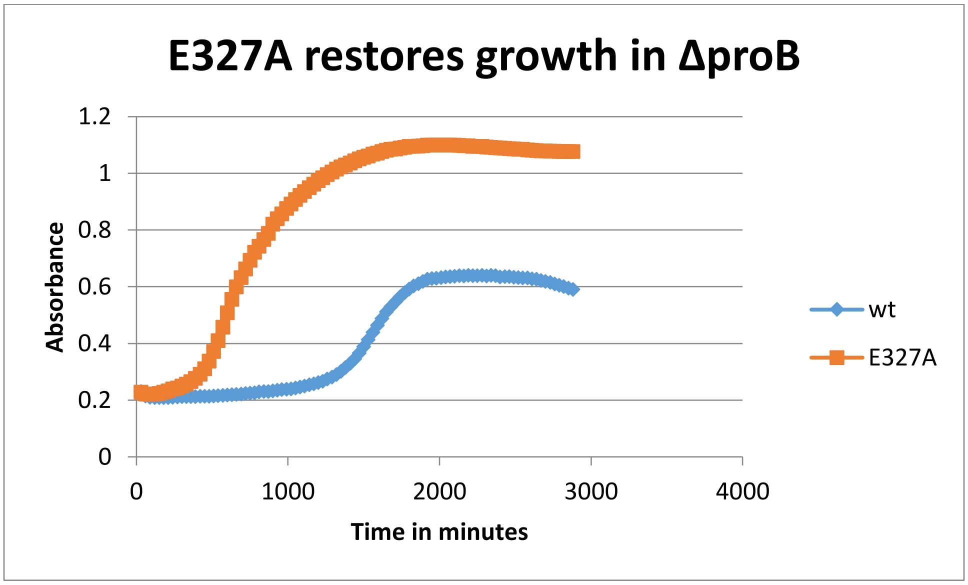

Figure 2: ΔproB cells expressing glnA with the E327A mutation display improved growth over cells expressing wildtype glnA.

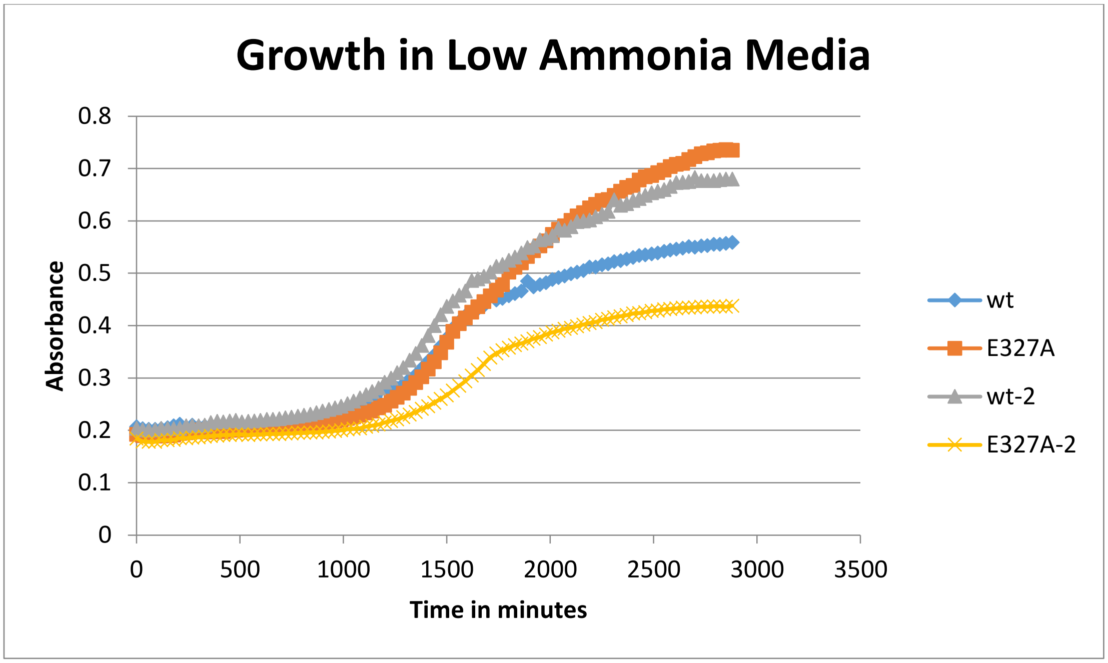

Figure 3: When ΔproB cells are grown in low ammonia medium, wildtype glnA is able to rescue growth at the same level as glnA with the E327A mutation.

References

1. Alibhai, M. and J.J. Villafranca, *Kinetic and mutagenic studies of the role of the active site residues Asp-50 and Glu-327 of Escherichia coli glutamine synthetase.* Biochemistry, 1994. **33**(3): p. 682-6.
2. Almassy, R.J., et al., *Novel subunit-subunit interactions in the structure of glutamine synthetase.* Nature, 1986. **323**(6086): p. 304-9.
3. Eisenberg, D., et al., *Structure-function relationships of glutamine synthetases.* Biochim Biophys Acta, 2000. **1477**(1-2): p. 122-45.
4. Kitagawa, M., et al., *Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research.* DNA Res, 2005. **12**(5): p. 291-9.
5. Musgrave, W.B., et al., *Probing the origins of glutathione biosynthesis through biochemical analysis of glutamate-cysteine ligase and glutathione synthetase from a model photosynthetic prokaryote.* Biochem J, 2013. **450**(1): p. 63-72.