UNIVERSITY OF CALIFORNIA
RIVERSIDE

Developing *Pseudomonas chlororaphis* as a Biotechnology Host for Phenazine Production Using a Population Genomics-Guided Metabolic Engineering Approach

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemical and Environmental Engineering

by

Sarah Ann Thorwall

December 2022

Dissertation Committee:
Dr. Ian Wheeldon, Chairperson
Dr. Yanran Li
Dr. Jason Stajich

The Dissertation of Sarah Ann Thorwall is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGMENTS

The text of Chapter 1 of this dissertation is a reprint of the material as it appears in the Perspective article “Stress-tolerant non-conventional microbes enable next-generation chemical biosynthesis” published in issue 16 of Nature Chemical Biology in 2020. Cory Schwartz assisted in writing the first draft of the manuscript, Justin W. Chartron wrote the section “Acute and long-term responses to environmental stress,” and Sarah Thorwall and Ian Wheeldon wrote and edited the remaining manuscript.

The remaining text is, in part, based off work by Sarah Thorwall, Varun Trivedi, and Ian Wheeldon which has been submitted for publication in 2022. Varun Trivedi performed the microbial GWAS analysis, and Ian Wheeldon supervised the research which forms the basis for this dissertation. Varun Trivedi wrote the methods section for the mGWAS analysis.

Bacterial strains used in this study were obtained from the American Type Culture Collection, German Collection of Microorganisms and Cell Cultures, and the Westerdijk Fungal Biodiversity Institute’s Netherlands Culture Collection of Bacteria.

This material is based upon work supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program under Award Number DE-SC0019093, Air Force Office of Scientific Research award FA9550-17-1-0270, Army Research Office MURI award W911NF1410263, and National Science Foundation award NSF-CBET 1706545 for funding. The UCR Plants3D NSF NRT Grant 1922642 also provided funding and mentoring support.

ABSTRACT OF THE DISSERTATION

Developing *Pseudomonas chlororaphis* as a Biotechnology Host for Phenazine Production Using a Population Genomics-Guided Metabolic Engineering Approach

by

Sarah Ann Thorwall

Doctor of Philosophy, Graduate Program in Chemical and Environmental Engineering
University of California, Riverside, December 2022
Dr. Ian Wheeldon, Chairperson

Metabolic engineering improves industrial biochemical production through beneficial edits in the microbial host's genome. Traditional metabolic engineering approaches modify the relevant biosynthesis pathway or induce genomic mutations in response to environmental stress, often within a few well-characterized host strains. With the advent of next generation sequencing and new synthetic biology tools (e.g., CRISPR genome editing), more microbes could be sequenced and engineered with a genome-wide approach as novel bioproduction hosts. Further sequencing technology improvements and cost decreases have recently made it economically feasible to sequence collections of many microbial genomes. These large genomic datasets could be used to determine the genetic underpinnings of industrially-relevant phenotypes and identify non-intuitive genome-wide metabolic engineering targets.

This study focuses on developing *Pseudomonas chlororaphis* as a biotechnology host for the production of phenazines, its colorful, redox-active secondary metabolites which have current and potential uses as agricultural fungicides and within bioectrochemical devices. First, I experimentally characterized the phenazine production, biofilm formation, and temperature tolerance of 34 strains and used that phenotype data to identify the best phenazine production strain. Next, I sequenced all strains with both Illumina and Oxford Nanopore technologies and assembled their genomes with multiple state-of-the-art assembly algorithms (i.e., SPAdes, Unicycler, Flye) to ensure I had optimal genomic input for downstream analysis. The final genome assemblies and phenazine phenotype data were passed to the microbial GWAS algorithm DBGWAS, which provided a list of 330 hits statistically associated with phenazine production. Finally, I overexpressed the 7 coding sequences (CDSs) associated with the top phenotype scores within strain DSM 21509, the strain which naturally has the highest phenazine-1-carboxamide titers. I confirmed that 3 of these CDSs, a putative transcriptional repressor, carboxypeptidase, and histidine transporter, significantly affected phenazine-1-carboxamide titers. This work demonstrates the potential of a population genomics-guided approach to metabolic engineering, which identifies engineering targets from a collection of genomes rather than a single genomic reference.

Table of Contents

List of Figures

List of Tables

Chapter 1: Stress-tolerant non-conventional microbes enable next-generation chemical biosynthesis

Abstract

Microbial chemical production is a rapidly growing industry, with much of the growth fueled by advances in synthetic biology. New approaches have enabled rapid strain engineering for the production of various compounds; however, translation to industry is often problematic because native phenotypes of model hosts prevent the design of new low-cost bioprocesses. Here, we argue for a new approach that leverages the native stress-tolerant phenotypes of non-conventional microbes that directly address design challenges from the outset. Growth at high temperature, high salt and solvent concentrations, and low pH can enable cost savings by reducing the energy required for product separation, bioreactor cooling, and maintaining sterile conditions. These phenotypes have the added benefit of allowing for the use of low-cost sugar and water resources. Non-conventional hosts are needed because these phenotypes are polygenic and thus far have proven difficult to recapitulate in the common hosts *Escherichia coli* and *Saccharomyces cerevisiae*.

Introduction

Industrial biotechnology generates over $324 billion in US revenues annually, representing more than 2% of the US gross domestic product (GDP)[1]. Industrial biochemicals (for example, biofuels, food additives, biopolymers, and other commodity chemicals) represent more than $125 billion of these revenues and are the fastest growing subsector with year-over-year growth consistently exceeding 10% over the last decade. As a whole, industrial biotechnology is a substantial driver of the US economy, but a new generation of biochemical processes is needed for the industry to continue to expand. The remarkable advancement of synthetic biology and metabolic engineering has provided the technical capabilities to rapidly design and create cells with optimized pathways, but our ability to design low-cost bioprocesses has not kept pace with the new technical innovations driving this sector (for example, genome editing by CRISPR–Cas9 and other similar systems[2], low cost and rapid DNA synthesis[3], and automated combinatorial gene assembly[4]).

Current challenges limiting the competitiveness of bioprocessing include expensive feedstocks, high energy and water use, loss of productivity due to contamination, and high downstream separation costs. These technical challenges have proven difficult to overcome, in part, because of inherent limitations in model host organisms. The genetics and metabolism of *E. coli* and *S. cerevisiae*, common lab-scale hosts, are well understood, and genome-editing techniques are advanced; however, both are mesophilic, do not tolerate environmental stresses, and have limited native or engineered capacity to metabolize carbon sources other than glucose[5, 6]. Efforts have

been made to address these limitations, but success is often limited because these phenotypes are frequently polygenic, requiring the manipulation of multiple genes to bring about substantial changes. Moreover, the detailed molecular mechanisms underpinning phenotype are not always fully understood.

Many non-conventional microbes (i.e., microbes not widely used in lab-scale studies and/or not common in industrial bioprocessing) have phenotypes that address the technical challenges in bioprocessing, but these are often not selected as new production hosts because their genetics are not sufficiently tractable and/or the synthetic biology tools necessary to optimize product titer, rate, and yield are not available. The expansion of synthetic biology into non-model hosts promises to enable a new generation of bioprocessing, as microbes that natively possess industrially favorable phenotypes can be engineered as production strains with desired biosynthetic pathways. This is not to say that one host could address every process challenge, but by increasing the number of industrially viable hosts, process designers can best match microbial phenotypes with process needs.

In this Perspective, we discuss how non-conventional microbes can enable new, low-cost bioprocessing for chemical production. The discussion begins with an analysis of the process benefits that can be realized by exploiting complex traits including tolerance to high temperatures, extreme pH, non-aqueous solvents and organic acids, and high osmotic stress. Following this, we discuss the molecular underpinnings of stress-tolerance traits and engineering efforts to enhance these traits in both non-model and model hosts. We also describe the technical challenges in engineering non-conventional

microbes. Lastly, we comment on the future directions of the field. This Perspective is not intended to be a comprehensive review of all potentially relevant microbes and their phenotypes, but instead uses selected examples (see Table 1.1) to illustrate our view that non-conventional microbial hosts and their unique traits will play a central role in the future of biochemical processing.

Table 1.1. Microbial hosts discussed in this chapter along with their industrially relevant phenotypes and metabolic engineering products.

Microbial hosts and their limits	Products
Thermotolerance	
Thermoanaerobacterium saccharolyticum (45–65 °C)	Ethanol, n-butanol[29]
Clostridium thermocellum (50–60 °C)	Ethanol, isobutanol[30]
Geobacillus thermoglucosidasius (40–70 °C)	Ethanol, lactate, formate, acetate, pyruvate[31]
Halomonas campaniensis (20–55 °C)	Polyhydroxybutyrate[32]
Kluyveromyces marxianus (52 °C)	Ethanol, ethyl acetate, inulase, galactosidase[11], esters[34], fatty acids[38], hexanoic acid
Issatchenkia orientalis (30–45 °C)	Ethanol[24], succinate[35], lactate[75]
Hansenula polymorpha (37–48 °C)	Ethanol[33]
Metallosphaera sedula (70–75 °C)	Metallic copper[79]
Model hosts*: *E. coli* (37 °C); *S. cerevisiae* (30 °C)	
Osmotolerance	
Halomonas campaniensis (1–20% (w/v) NaCl)	Polyhydroxybutyrate[32]
Halomonas elongata (>20% (w/v) NaCl)	Ectoine[48]
Model hosts*: *E. coli* (<3% (w/v) NaCl); *S. cerevisiae* (<4% (w/v) NaCl)	
Solvent tolerance	
Pseudomonas putida (1–3% (v/v) n-butanol)	n-Butanol[55], polyhydroxyalkanoates, rhamnolipids, phenol, p-coumarate, p-hydroxybenzoate[65]
Clostridium acetobutylicum (1–2% (v/v) n-butanol)	Acetone, n-butanol, ethanol[66]
Model hosts*: *E. coli* (1% (v/v) n-butanol); *S. cerevisiae* (1–2% (v/v) n-butanol)	
Acid tolerance	
Lactobacillus casei (pH 3.3-6.5)	Lactate, acetate, fermented foods[68]
Issatchenkia orientalis (pH 1.5-8)	Ethanol[24], succinate[35], lactate[75]
Kluyveromyces marxianus (pH 3.8)	Ethanol, ethyl acetate, inulase, galactosidase[11], esters[34], fatty acids[38], hexanoic acid[39]
Acidithiobacillus ferrooxidans (pH 1–2)	Isobutyric acid, heptadecane, metallic copper[77]
Metallosphaera sedula (pH 1-4.5)	Metallic copper[79]
Model hosts*: *E. coli* (pH 4.5); *S. cerevisiae* (pH 4)	

*In some cases, high-tolerance phenotypes have been engineered, but typically are accompanied by other altered phenotypes that can be detrimental to bioprocessing (see text).

Matching bioprocess challenges with microbial phenotypes

Typically, microbial fermentations are conducted at moderate temperatures (for example, 30–37 °C); however, substantial benefits can be realized with microbes that sustain high growth rates at elevated temperatures. Bioreactor cooling is necessary because microbial growth is exothermic, releasing up to half of the energy stored in its substrates as heat. High cooling costs are particularly acute in tropical climates, in which sugarcane and other carbon sources are abundant at low cost. Operating at high temperatures can alleviate some of these costs, because lower cost cooling water is typically suitable for heat exchange, whereas lower bioreactor temperatures (<~35 °C) require expensive refrigerated cooling water to maintain a sufficient temperature gradient for temperature control. Moreover, elevated process temperatures also reduce the heating required for the separation of volatile bioproducts such as ethanol and butanol from spent media.

The bioprocessing costs and environmental impact associated with handling water, both as input and waste streams, can be substantial. For example, ethanol plants use an average of five gallons of treated water per gallon of ethanol produced[7]. Untreated water sources such as ocean water are abundant, but microbial growth is inhibited owing to high salt content. An alternative is to use growth media with high concentrations of carbon sources and nutrients to drive high cell density cultures, but this requires hosts that are tolerant to the osmotic pressure created by the media. The decrease in fermentation volume associated with high cell density cultures is also beneficial

because smaller volumes require less purification equipment and smaller reaction vessels, which are simpler to sterilize and maintain.

Historical bioprocessing data reveals an inverse relationship between product titer and separation costs; products at low concentrations are difficult and costly to separate from large volumes of aqueous fermentation broth that not only contain the desired product but also waste compounds and side products[8]. Increasing product titer increases production and decreases separation costs, but high concentrations of some products can be toxic. This effect is readily apparent with *S. cerevisiae* ethanol production, wherein industrial titers are limited by ethanol toxicity[9]. This represents another area in which non-conventional microbial hosts with native tolerance to valuable bioproducts (for example, alcohols and organic acids) can benefit bioprocess design.

Microbial contamination reduces process yields by parasitically consuming the carbon source and by producing inhibitors that increase fermentation time. This is particularly problematic in bioprocesses with narrow profit margins in which even minimal losses in yield can substantially affect process economics. For example, bacterial contamination in *S. cerevisiae* cultures has been shown to reduce ethanol titers by more than 10 g/L, likely owing to inhibition of *S. cerevisiae* growth by acetic acid produced by the bacteria[10]. Non-aseptic culturing enabled by hosts that can withstand process conditions that are inhospitable to other microbes can provide major cost savings, as input streams do not need to be sterilized and simple reactor designs made from low-cost materials can be used. Of course, the process conditions for non-aseptic systems and/or conditions specific to a single stress must also be balanced with the needs of the desired

product (for example, high-temperature fermentation would not be suitable for the biosynthesis of thermally labile compounds and culture pH must be matched with the desired product protonation state).

It is also important to consider how the traits of a host affect the design–build–test cycle. Critical to this cycle is the ability to make genomic edits within a reasonable timeframe, which is challenging when growth is slow. With current workflows, multiple edits can be made in *E. coli* and *S. cerevisiae* within a week or two. The growth rates of many of the non-conventional hosts discussed here are comparable to their conventional counterparts, thus facilitating rapid engineering. For example, *Kluyveromyces marxianus* is one of the fastest-growing eukaryotes with growth rates nearly double that of *S. cerevisiae*[11]. In the case of slower-growing non-conventional hosts (for example, *Acidithiobacillus ferrooxidans*), strain engineering may be slowed, but beginning with a set of phenotypes that match the desired process can limit the number of edits required to create the final production strain.

Acute and long-term responses to environmental stress

When exposed to environmental stress, microorganisms will try to counteract its effect, remove it, or repair its damage. How cells respond after acute stress differs from that during both gradual acclimation and evolutionary adaptation[12]. Adaptations used by stress-tolerant microbes are often reflected in numerous changes to cell physiology. Responses are transient mechanisms to prevent and repair damage, including the upregulation of protein-folding chaperones, transporters and other membrane-modulating factors, nucleic acid repair enzymes, and proteolytic enzymes that degrade terminally

misfolded proteins[13]. Many of these responses are universally conserved and are triggered by shifts relative to the ideal conditions for a given species. For instance, thermophiles and mesophiles express homologous molecular chaperones when rapidly heated 5–10 °C above their optimal growth temperature[14]. When slowly acclimated to a stress, these organisms can also survive in conditions in which an acute shock would be lethal[15]. Acclimating to prolonged stress, while maintaining some of the same patterns as the acute response, ultimately leads to distinct transcriptional states and a stress-tolerant phenotype[16].

Many stresses converge on the proteome, as the ΔG between the folded and unfolded states is typically only a few kcal mol^{-1}. Protein state is strongly influenced by temperature, pH, and osmolarity, and the response to proteotoxic challenges is well documented[17]. For example, a rapid increase in temperature leads to expression of aptly named heat shock proteins (HSPs). However, this response is better described as a general mechanism to protect the cell against protein unfolding[13, 18]. In *S. cerevisiae*, HSPs are also induced by exposure to peroxides, diamides, osmotic shock, nitrogen or amino acid starvation, and the diauxic shift[19], thus pointing to the generality of the stress response. Similarly, in *E. coli*, a variety of acute proteotoxic stresses, including extremes of temperature, pH, and accumulation of ethanol lead to a shared response[20].

Microbial stress-tolerance phenotypes

Thermotolerance

The polygenic underpinnings of thermotolerance include adaptations such as altered membrane composition, efficient DNA repair, structural changes to genome

packing, and increased HSP expression[21]. The various adaptations are illustrated by the diverse membrane compositions of stress-tolerant microbes (Figure 1.1). Archaeal phospholipids have thermostable ether linkages and branched acyl chains[22]. The branched-chain phospholipids reduce membrane permeability, helping to maintain ion gradients under stress conditions. Higher production of unsaturated fatty acids in ethanol-producing strains of the thermotolerant yeasts *K. marxianus*[23] and *Issatchenkia orientalis*[24] is distinctly different from the thermotolerance solution adapted by many bacteria that produce saturated acyl chains under temperature stress[21].

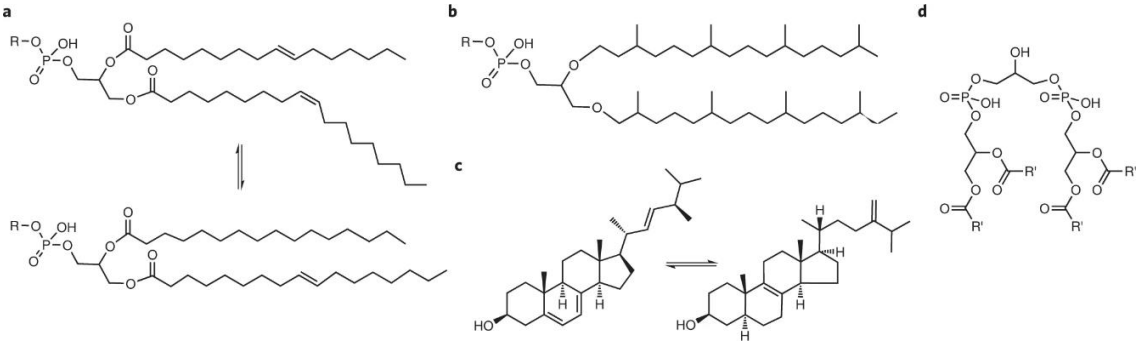

Figure 1.1. Membrane chemistries of stress-tolerant microbes. a) In native and engineered thermo- and solvent-tolerant microbes, cis acyl chains are converted to the trans configuration, and the ratio of saturated to unsaturated chains is increased. b) Thermophilic and acidophilic archaea have branched acyl chains and stable ether linkages that reduce membrane permeability and increase stability. R is any phospholipid head group modification. c) In engineered yeast, membrane sterol composition is altered to favor the production of 'bent' fecosterol (right) over 'flat' ergosterol (left). d) Changing the phospholipid head group allows tolerant microbes to adapt to higher temperatures, solvent concentrations, and acidities. R' is any acyl chain found in membrane phospholipids.

The various mechanisms that bring about thermotolerance are also evident from engineering efforts in *E. coli* and *S. cerevisiae*. Two long-term adaptive laboratory evolution studies were successful in isolating new strains of *E. coli* with maximum growth temperatures of ~48 °C[25, 26]. In addition, rational design[23], laboratory

evolution[27], and genome shuffling[28] have been successful in increasing tolerance of *S. cerevisiae* to ~45 °C. Although these studies increased thermotolerance, negative phenotypic effects were observed in both organisms; specifically, the loss of growth on glycerol and reduced growth rates at standard temperature. These results again highlight the complex genetic interactions of tolerance phenotypes; in this case, the laboratory evolution of higher tolerance detracted from other critical traits.

When considering past successes in metabolic engineering, genetic tractability, and the availability of synthetic biology tools, the following bacterial hosts should be considered for thermotolerant applications: *Thermoanaerobacterium saccharolyticum*, *Clostridium thermocellum*, *Geobacillus thermoglucosidasius*, and *Halomonas campaniensis* (Table 1.1). Each species has been shown to have a high capacity for growth at temperatures upward of 60 °C, and standard metabolic engineering tools are available (i.e., transformation protocols, expression plasmids, promoters, and gene deletion/integration methods). *T. saccharolyticum*, *C. thermocellum*, and *G. thermoglucosidasius* have primarily been the focus of bioenergy and consolidated bioprocessing research and have been engineered to produce n-butanol (1.05 g/L from xylose[29]), isobutanol (5.4 g/L from cellulose[30]), and ethanol (14.8 g/L on cellobiose[31]). *H. campaniensis* has been proposed as a multistress-tolerant host and has been demonstrated as a platform for low-cost polyhydroxyalkanoate production[32].

K. marxianus, along with the other thermotolerant yeasts *I. orientalis* and *Hansenula polymorpha*, also have the native capacity to metabolize a range of carbon sources including glycerol, C5 and C6 sugar monomers, and C12 disaccharides[33-35].

This flexible metabolism benefits process design, as it allows the use of the lowest cost carbon source available. Pathway engineering success has also been demonstrated. In our own work, we enhanced the production of ethyl acetate in *K. marxianus* via a multiplexed CRISPR interference (CRISPRi) engineering strategy[36]. *K. marxianus* has also been engineered to overproduce esters, fatty acids, and ethanol[37-39]. Similarly, the native stress tolerance of *I. orientalis* has been leveraged to create new overproduction strains for succinic acid biosynthesis[35], and *H. polymorpha* has been explored as a high-temperature ethanol production host[40].

Osmotolerance

Despite the general toxicity of high extracellular osmolyte concentrations, a number of different microbes have evolved robust tolerance. Some produce internal osmolytes including ectoine, trehalose, and glycine betaine, which balance the osmotic gradient across the cell membrane (Figure 1.2). Other halotolerant microbes balance the osmotic pressure by allowing salts such as KCl to accumulate intracellularly. These microbes evolved proteomes with high acidic amino acid content, which is thought to maintain protein stability and activity under high salt concentrations[41]. Though this strategy is less energetically costly than producing internal osmolytes, the proteomes can denature at low salt concentration, thus limiting the range of conditions in which the microbe can survive[42]. Another strategy uses active ion pumps, most commonly Na^+/H^+ antiporters, to maintain Na^+ and pH homeostasis[43]. This pumping action also comes at a cost because the antiporters reduce the proton gradient across the outer membrane, which is necessary for ATP generation.

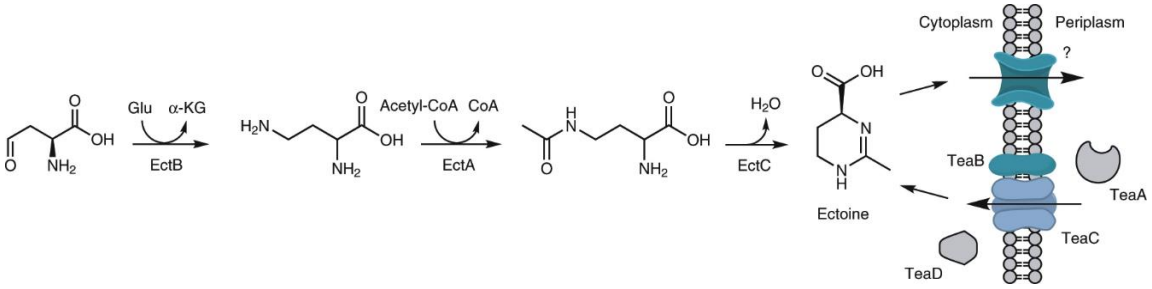

Figure 1.2. Metabolism and transport of the osmolyte ectoine. Much of the mechanistic understanding of ectoine as an osmolyte comes from the study of *Halomonas elongata*, which converts L-aspartate-β-semialdehyde into ectoine via the ectABC gene cluster[48]. *H. elongata* utilizes exogenous ectoine through the tripartite ATP-independent transporter TeaABC. TeaA is a substrate-binding protein, TeaC and TeaB span the cytoplasmic membrane, and TeaD is a negative regulator of TeaABC[47]. An unknown transporter (question mark) is thought to remove ectoine from the cytoplasm. Both the biosynthesis and transmembrane transport of ectoine are regulated by environmental salinity and by intra- and extracellular osmolyte levels. α–KG, α-ketoglutarate.

Engineering microbes to withstand high osmotic conditions presents some challenges: ion pumping places a metabolic burden on the cell, proteome-wide adaptations would be laborious to implement, and in cases in which high osmolyte concentrations are needed for protection, the yield of desired bioproduct could be reduced. The osmotolerance phenotype must also be well suited to the range of process conditions experienced over the course of a single production run. In fermentations rich in sugars, the media composition substantially changes over time, thus presenting a variable environmental stress: high osmotic stress at the outset and stress from high product titers toward the end of a production batch. In addition, some of these strategies may not be compatible with other tolerance phenotypes. For example, some Na^+/H^+ antiporters function at neutral pH, severely limiting the range of process conditions wherein this strategy could be implemented.

Many osmotolerance adaptions found in non-conventional microbes have inspired the engineering of model hosts. Synthetic trehalose production in *E. coli* increased NaCl

and glucose minimum inhibitory concentrations to over 550 and 800 mM, respectively; however, the enhanced tolerance was accompanied by increased growth inhibition by ethanol[44]. Another study demonstrated that heterologous expression of a sodium export pump in *E. coli* could improve growth at elevated NaCl concentrations, but final cell densities were significantly reduced in comparison to low salt conditions[45]. *Saccharomyces pastoranus*, a close relative of *S. cerevisiae*, has also been evolved for higher osmotolerance, which resulted in reducing fermentation time from 8 to 4.5 days when grown on a high-sugar medium. Interestingly, the evolved strain showed considerably less trehalose production, suggesting that an alternative compensation mechanism may be responsible for the improved fermentation[46].

One source of highly salt-tolerant non-conventional bacteria is the genus *Halomonas*. Microbes from this genus primarily fight osmotic stress by producing the osmolyte ectoine[47, 48]. Researchers have developed stable plasmids, high-expression promoters, and CRISPR–Cas9 genome-editing and gene-regulation systems[49, 50]. By growing engineered strains under conditions inhospitable to most contaminants, products such as polyhydroxyalkanoates (~42 g/L PHB) have been produced in open, non-sterile bioreactors using ocean-water-based growth media[32]. Other compounds such as ectoine (15.9 g/L) have also been produced[51]. A recent report of a novel strain of *S. cerevisiae* capable of growing in sterilized ocean water and successful ethanol fermentation with *K. marxianus* grown on substrates produced in 50% ocean water are valuable examples of yeasts with native osmotolerance[52, 53].

Toxic product tolerance

As various toxic metabolites cause different types of damage throughout cells, no single strategy has emerged to achieve broad product tolerance. Pseudomonads, which tolerate a wide range of solvents and have been extensively studied, are useful examples of a multipronged approach[54, 55]. Solvent tolerance in *P. putida* has been shown to arise from altered membrane composition (Figure 1.1d)[56] and the upregulation of RND (resistance-nodulation-division) efflux pumps (Figure 1.3)[57, 58] and protein-folding chaperones[59]. With some success, *E. coli* has been engineered to increase product tolerance by exploiting these mechanisms. Growth screens with a library of 43 different efflux pumps with homology to a *P. putida* protein[60] identified candidates that increased culture viability in the presence of various terpenoids (but not butanol or isopentanol). In another example, heterologous expression of an isomerase from *P. aeruginosa* increased octanoic acid titers by 29%[61]. Lastly, overexpressing the chaperone GroESL improved cell growth and viability in the presence of ethanol, n-butanol, and 2-butanol[62]. Although these engineering examples show some success, these methods vary in effectiveness, are likely not addressable by a single engineering strategy, and do not achieve the high-tolerance phenotypes natively found in pseudomonads.

In the example of ethanol tolerance in *S. cerevisiae*, the model species already possesses industrially relevant phenotypes; that is, high fermentative capacity to produce ethanol and tolerance to reasonable concentrations of the desired product. Despite a natural tolerance to ethanol, the economics of commercial production has driven research

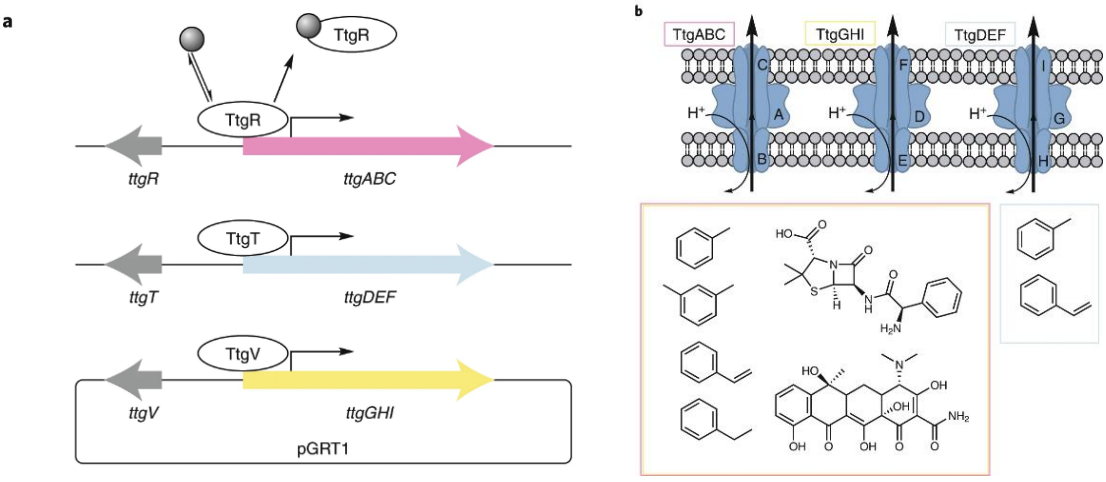

Figure 1.3. High solvent tolerance in *Pseudomonas putida* through effective efflux pumping.
a) Gene clusters of a set of conserved RND (resistance–nodulation–division) efflux pumps in *P. putida*. The ttg clusters are named for their ability to process toluene (ttg, toluene tolerance genes). Each pump is encoded by a single operon; ttgABC and ttgDEF are chromosomal, whereas the ttgGHI operon is expressed from a self-transmissible plasmid (pGRT1). Expression is regulated by a repressor specific to each operon. TtgR, TtgT, and TtgV regulate ttgABC, ttgDEF and ttgGHI, respectively, with some cross-talk between TtgT and TtgV[58]. b) TtgABC and TtgGHI remove toluene, m-xylene, styrene, ethylbenzene (left, top to bottom, respectively), and the antibiotics tetracycline (middle, bottom) and ampicillin (middle, top), among others[57]. TtgDEF is specific to toluene and styrene (right, top and bottom, respectively).

to improve tolerance to enable higher titers. One study increased ethanol titers up to 80%[9] by adding KCl and KOH to the media. The added salt restored K^+ and H^+ gradients across the cell membrane that were disrupted because of increased membrane permeability in high ethanol. The observed effect was successfully translated to a genetic strategy of hyperactivating a potassium importer and upregulating a proton exporter.

The high tolerance of *P. putida* to solvent stress represents a valuable opportunity to create new high-titer processes, as the tolerance of *E. coli* to solvents has proven limited. A number of synthetic biology tools, including promoters, expression vectors, and genome-editing systems (for example, CRISPR–Cas9 and λ-red recombineering) have also been developed[63]. These tools have been used to create new strains for both biodegradation[64] and chemical biosynthesis (for example, polyhydroxyalkanoates,

omega-3 fatty acids, and various organic acids, among others)[65]. The solventogenic *Clostridium acetobutylicum* is also a viable host. Notably, this host was used for the industrial fermentation of acetone to support war efforts in the early 20th century. More recently, advanced metabolic engineering tools have been used to control the ratio of acetone, butanol, and ethanol[66]; chemical synthesis routes have been developed to convert the fermented solvents into long chain gasoline-like molecules[67]; and *in situ* product-removal strategies have been used to improve product titers.

Acid tolerance

Natural adaptations to acidic environments are similar to other tolerance mechanisms, including altered membrane composition, active efflux pumps, and changes to the proteome. To preserve ion gradients under acid stress, tolerant microbes have membranes with reduced permeability, such as the highly methylated membranes of some archaea[22]. Active ion pumps can maintain a neutral cytoplasm by pumping out excess H^+ at the cost of ATP[68]. Proteome adaptations also allow proteins to function at low pH. For example, acidophiles may also have proteins with fewer acidic amino acids to prevent electrostatic repulsion[22]. In addition to tolerating a low-pH environment, organic acid-producing hosts need to withstand the stress caused by the toxic organic acid product. Maintaining a low extracellular pH increases the fraction of organic acids in the protonated state, which allows the product to be directly extracted from the culture media, but also leads to toxic effects as the protonated organic acids diffuse across the cell membrane and dissociate in the near-neutral cytoplasm.

The common hosts *E. coli* and *S. cerevisiae* have been engineered for tolerance and production of organic acids, including lactic[69, 70] succinic[71, 72], and acetic acids[73, 74]. In *S. cerevisiae*, adaptive laboratory evolution improved growth in the presence of ~5 g/L of acetic acid at pH 4[73]. This engineered tolerance was conditional, as strains were unable to grow after acute exposure to these same conditions. Adaptive evolution has also been applied to *E. coli*, in which a strain was evolved to withstand high succinate concentrations[72]. Other studies have also been conducted to separate the effects of low pH and high concentrations of a specific organic acid. One study used multiple genomic libraries and growth selections to identify the mechanisms behind acetate tolerance in *E. coli*[74]. The broad range of mechanisms found to be responsible for enhanced acetate tolerance again supports the idea that tolerance is polygenic and arises from not one, but a set of parallel mechanisms.

Native microbial producers of organic acids have also been engineered for enhanced tolerance. Laboratory evolution of *Lactobacillus casei*, a species commonly used to produce fermented foods, improved biomass formation by 60% at a pH of ~4[68]. Another example of a natively acid-tolerant production host is a strain of *I. orientalis* engineered by Cargill, Inc. for lactic acid fermentation[75]. This host was chosen for its acid tolerance, as it must withstand the low pH caused by high titers of organic acids. Another potential yeast host is the previously mentioned *K. marxianus*, which maintains fast growth kinetics at pH ~3[11].

Acidophilic extremophiles can also potentially serve as next-generation hosts. The bacteria *Acidithiobacillus ferrooxidans* is used for industrial copper bioleaching and

grows optimally at pH ~2. As a chemolithoautotroph, it can fix carbon from CO_2 and obtain energy from iron and sulfur oxidation. These interesting phenotypes as well as the increasing availability of genetic engineering tools (for example, a completely sequenced genome[76], functional plasmids[77], chromosomal integration methods[78]) are also advantageous. One proof-of-concept study successfully engineered *A. ferrooxidans* to produce isobutyric acid and heptadecane (570 μg/L and 0.6 μg/L, respectively), supporting the idea that electrochemical cells coupled to *A. ferrooxidans* could convert CO_2 into useful chemicals[77]. Another potential host is the archaea *Metallosphaera sedula*, which is both thermophilic (70–75 °C) and acidophilic (pH ~3). *M. sedula* also has pathways for carbon fixation, sulfur oxidation, and copper bioleaching as well as a fully sequenced genome and functional gene-knockout system[79-81]. Although *M. sedula* has yet to be engineered for biochemical production, its native combination of interesting phenotypes and the development of a system enabling transformation and selection suggests that future efforts toward more advanced metabolic engineering will be fruitful.

Challenges in engineering non-conventional microorganisms

The use of multitolerant microbes will help avoid technical challenges associated with biochemical process scale-up; however, engineering these microbes is almost always more difficult than engineering *E. coli* and *S. cerevisiae*. The extensive investment put toward gaining fundamental knowledge of the genetics and metabolism and the development of genetic-engineering tools that enable such studies provide a clear advantage for common hosts. The complete genomes of *E. coli* and *S. cerevisiae* have

been sequenced for over 20 years, metabolic models have been constructed for each[82, 83], and genome-scale engineering techniques have been developed[84, 85]. Moreover, transformation protocols, plasmids and genome-integration strategies, and collections of genetic parts are standardized and widely available[86, 87]. These synthetic biology tools greatly facilitate engineering through pathway expression level tuning, high-throughput screening of mutational libraries, and rapid construction of new strains with desired heterologous pathways.

In many of the cases discussed in the previous sections, the bioproduct titers achieved with non-conventional hosts do not reach the levels of highly engineered *E. coli* or *S. cerevisiae*. For example, high-temperature ethanol production in *K. marxianus* or *G. thermoglucosidasius* is much lower than the >100 g/L of ethanol produced in industrial processes. The development of non-conventional microbes as production hosts is at various stages. Many have traditional transformation and transfection methods, whereas others have a small set of synthetic biology tools, but the multiplexed editing and combinatorial mutant libraries have yet to be developed (Figure 1.4). As the technology gap decreases, engineered non-conventional hosts will not only provide a phenotypic advantage, but also through strain engineering will achieve titers that are comparable to or surpass those from *E. coli* and *S. cerevisiae*.

The lack of tools and biological understanding is not the only hurdle to overcome in engineering non-conventional hosts, as they often have characteristics that make them recalcitrant to engineering, including resistance to harboring heterologous genes and low rates of foreign DNA incorporation, as well as prevention of easy access to genetic

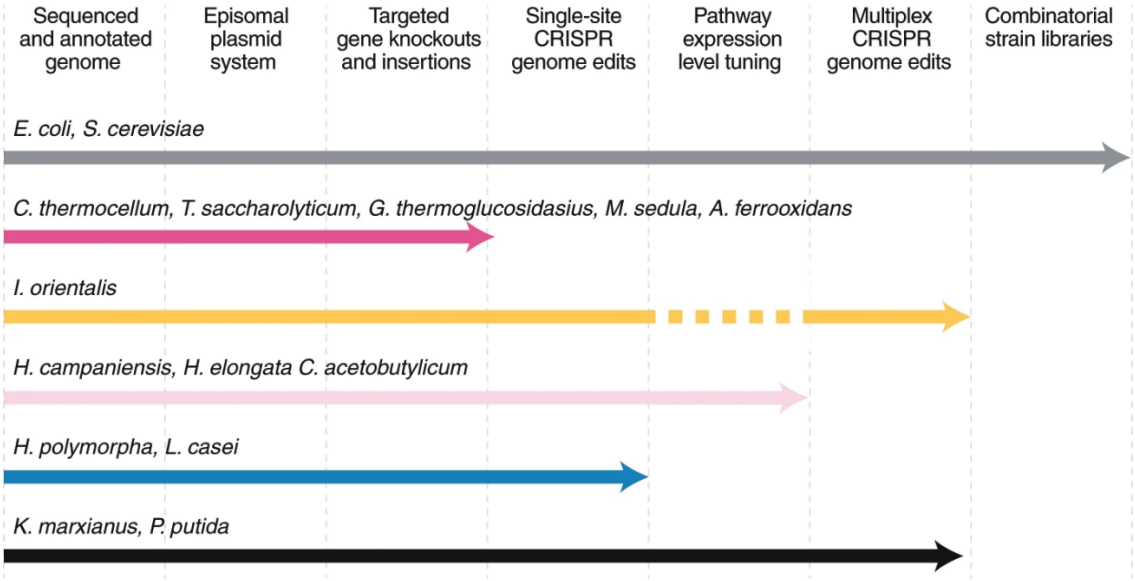

Figure 1.4. Progress of synthetic biology tool development. The current stages for selecting non-conventional microbial hosts compared to the conventional hosts *E. coli* and *S. cerevisiae*. Solid lines indicate developed synthetic biology tools; dashed lines indicate that the listed tools are under development.

material owing to the biochemistry of tolerance phenotypes. Much of the engineering in *S. cerevisiae* relies upon its high native capacity to perform homologous recombination, a capacity that is not present in most other yeasts and higher eukaryotes[88]. This limitation can be minimized by disrupting or repressing competing DNA repair through nonhomologous end-joining (NHEJ), but such strategies can present challenges of their own, including the need for transient repression systems and/or the permanent disruption of native DNA repair mechanisms that can lead to long-term strain instability[89].

Despite challenging genetic backgrounds, synthetic biology and metabolic engineering tools have been developed for many non-conventional organisms. With advances in DNA sequencing, synthesis, and synthetic biology, almost any microorganism that can be cultured and transformed with DNA can be adapted and used

as a production host with designed pathways. Engineering of non-conventional microbes has recently been enabled by the development of CRISPR–Cas9 and related systems, which can allow facile and targeted genome editing. In non-model eukaryotes, CRISPR–Cas9 has been used to knock genes out, integrate heterologous pathways, repress and activate transcription, and perform genome-wide studies[90-92]. In prokaryotes, CRISPR–Cas9 can be more challenging to use because of the absence of native DNA repair pathways that are present in eukaryotes (for example, NHEJ). Even so, CRISPR–Cas9 still enables targeted genome editing and has been applied in a range of species[2, 11, 50, 63, 90, 93]. Improvements in systems biology have also made it easier to identify genetic parts for testing (for example, promoters, terminators, and centromeric sequences for stable plasmid construction)[94], to construct metabolic models to inform engineering decisions[95], and to identify possible gene clusters for novel products[96]. There is also promise in creating tools that are functional across diverse species that will enable more rapid engineering of new isolates[97]. Current successes in tool development and engineering of non-conventional hosts demonstrate that many of the challenges in engineering these novel microbes can be overcome with sustained research efforts that build on past successes with the model microbes and by leveraging new innovations across diverse fields.

Conclusion and outlook

In working toward the next generation of bioprocesses, we envision near-term effort focused on matching our ability to engineer non-conventional microorganisms with the advanced synthetic biology tools that have been developed for *E. coli* and *S. cerevisiae*. Tools that can tightly control gene expression, rapidly create new pathways, and balance metabolic flux, as well as technologies that generate combinatorial strain libraries are needed to rapidly engineer new non-conventional production strains that maximize product titer, rate, and yield. These efforts are currently underway for a limited number of hosts, including the multitolerant hosts *K. marxianus*[11, 36], *H. campaniensis* LS21[32, 49], and *P. putida*[63]. These are all examples of how the recent innovation of CRISPR–Cas genome editing has exponentially increased the speed with which advanced synthetic biology can be applied in a selected host. Once a functional endonuclease-active CRISPR system is established for gene disruption, translation of that system to enable tuning of transcription via CRISPR interference and CRISPR activation can be rapidly achieved[89, 93, 98]. Although these tools have enabled the engineering of these non-conventional hosts, more advanced tools such as high-throughput screening of mutational libraries and automated systems to create combinatorial libraries are needed to match current capacities in model hosts. Additional investment is also needed to apply these strategies in newly identified microbes with advantageous phenotypes.

In the longer term, we envision considerable research effort to identify new candidate hosts that have robust growth under environmental stresses. Such phenotypes are not present in the common hosts *E. coli* and *S. cerevisiae*, and although engineering

efforts have helped reveal the mechanism of stress tolerance, these efforts have not produced enhanced phenotypes that rival those of many non-conventional hosts. New stress-tolerant organisms may be identified through bioprospecting in extreme environments, by searching existing physical collections, and through bioinformatic analysis of sequenced genomes. Bioprospecting is already routinely done and has identified novel isolates with industrially useful thermotolerant enzymes. Further bioprospecting promises to identify additional microbes with potent multitolerances. Several extensive repositories of microorganisms already exist, including the broad collections of microbes in the American Type Culture Collection, Agricultural Research Service Culture Collection (NRRL), and the German Collection of Microorganisms and Cell Cultures. Other, more curated repositories also exist. The Phaff Yeast Culture Collection at the University of California, Davis contains food and environmental yeast isolates, and the collection at the University of Western Ontario Department of Plant Sciences contains mostly ascomycetes isolated from plants and insects[99]. Although these repositories collectively have tens of thousands of different isolates, screening for desirable phenotypes within those collections is a slow and labor-intensive process and represents an area in which technical innovation is needed. Though advances in DNA sequencing have produced a large number of partial and complete genomes, computationally identifying organisms with tolerance phenotypes from only genome sequences is difficult and is an area of active research. Bioinformatic analysis across many species also provides insight into the genetic basis for tolerance phenotypes and promises to accelerate strain generation for industrial biotechnology.

In this Perspective, we have focused on how stress-tolerance phenotypes can address challenges in bioprocessing, but many non-conventional microbes can also metabolize a wide range of carbon sources. This flexibility is advantageous, as it allows the use of the lowest cost available carbon source. For example, *C. thermocellum* produces multi-enzyme complexes that breakdown the long glucan and xylan chains in lignocellulosic biomass, an ability that is also present in new isolates of anaerobic gut fungi[100]. Oleaginous species such as the yeast *Yarrowia lipolytica* (itself a promising non-conventional host with recently established synthetic biology tools[89, 91-93]) are also viable options for industrial streams rich in fatty acids, alkanes, or glycerol.

As new microbes with advantageous phenotypes are identified, there is an immediate need to create novel synthetic biology tools that function across a wide range of species to accelerate onboarding of novel organisms into engineering workflows. The core concept is selecting a host for its desired native traits, and synthetic biology tools can be used to optimize the desired native or heterologous biosynthetic pathways, achieving industrially relevant titers. This strategy promises to generate poly-tolerant strains with carefully tuned metabolism toward production of a chemical of interest, enabling production strains that can be grown non-aseptically while approaching theoretical maxima for titer, rate, and yield.

References

1. Carlson, R., *Estimating the biotech sector's contribution to the US economy.* Nat Biotechnol, 2016. **34**(3): p. 247-55.

2. Jinek, M., et al., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.* Science, 2012. **337**(6096): p. 816-21.

3. Caruthers, M.H., *A brief review of DNA and RNA chemical synthesis.* Biochem Soc Trans, 2011. **39**(2): p. 575-80.

4. Chao, R., et al., *Engineering biological systems using automated biofoundries.* Metab Eng, 2017. **42**: p. 98-108.

5. Hong, K.K. and J. Nielsen, *Metabolic engineering of Saccharomyces cerevisiae: a key cell factory platform for future biorefineries.* Cell Mol Life Sci, 2012. **69**(16): p. 2671-90.

6. Pontrelli, S., et al., *Escherichia coli as a host for metabolic engineering.* Metab Eng, 2018. **50**: p. 16-46.

7. Shapouri, H. and P. Gallagher, *USDA's 2002 Ethanol Cost-of-Production Survey*. Vol. 22. 2005.

8. Blanch, H.W. and D.S. Clark, *Biochemical engineering*. 1996, New York: M. Dekker. xii, 702 p.

9. Lam, F.H., et al., *Engineering alcohol tolerance in yeast.* Science, 2014. **346**(6205): p. 71-5.

10. Rich, J.O., et al., *Biofilm formation and ethanol inhibition by bacterial contaminants of biofuel fermentation.* Bioresour Technol, 2015. **196**: p. 347-54.

11. Lobs, A.K., et al., *CRISPR-Cas9-enabled genetic disruptions for understanding ethanol and ethyl acetate biosynthesis in Kluyveromyces marxianus.* Biotechnol Biofuels, 2017. **10**: p. 164.

12. Shui, W., et al., *Understanding the Mechanism of Thermotolerance Distinct From Heat Shock Response Through Proteomic Analysis of Industrial Strains of Saccharomyces cerevisiae.* Mol Cell Proteomics, 2015. **14**(7): p. 1885-97.

13. Richter, K., M. Haslbeck, and J. Buchner, *The heat shock response: life on the verge of death.* Mol Cell, 2010. **40**(2): p. 253-66.

14. Phipps, B.M., et al., *Structure of a Molecular Chaperone from a Thermophilic Archaebacterium.* Nature, 1993. **361**(6411): p. 475-477.

15. Parsell, D.A. and S. Lindquist, *The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins.* Annu Rev Genet, 1993. **27**: p. 437-96.

16. Strassburg, K., et al., *Dynamic transcriptional and metabolic responses in yeast adapting to temperature stress.* OMICS, 2010. **14**(3): p. 249-59.

17. Balchin, D., M. Hayer-Hartl, and F.U. Hartl, *In vivo aspects of protein folding and quality control.* Science, 2016. **353**(6294): p. aac4354.

18. Verghese, J., et al., *Biology of the heat shock response and protein chaperones: budding yeast (Saccharomyces cerevisiae) as a model system.* Microbiol Mol Biol Rev, 2012. **76**(2): p. 115-58.

19. Gasch, A.P., et al., *Genomic expression programs in the response of yeast cells to environmental changes.* Mol Biol Cell, 2000. **11**(12): p. 4241-57.

20. Rodrigues, J.L. and L.R. Rodrigues, *Potential Applications of the Escherichia coli Heat Shock Response in Synthetic Biology.* Trends Biotechnol, 2018. **36**(2): p. 186-198.

21. Lewin, A., A. Wentzel, and S. Valla, *Metagenomics of microbial life in extreme temperature environments.* Curr Opin Biotechnol, 2013. **24**(3): p. 516-25.

22. Jacquemet, A., et al., *Archaeal tetraether bipolar lipids: Structures, functions and applications.* Biochimie, 2009. **91**(6): p. 711-7.

23. Li, P., et al., *The transcription factors Hsf1 and Msn2 of thermotolerant Kluyveromyces marxianus promote cell growth and ethanol fermentation of Saccharomyces cerevisiae at high temperatures.* Biotechnol Biofuels, 2017. **10**: p. 289.

24. Seong, Y.J., et al., *Physiological and Metabolomic Analysis of Issatchenkia orientalis MTY1 With Multiple Tolerance for Cellulosic Bioethanol Production.* Biotechnol J, 2017. **12**(11): p. 1700110.

25. Blaby, I.K., et al., *Experimental evolution of a facultative thermophile from a mesophilic ancestor.* Appl Environ Microbiol, 2012. **78**(1): p. 144-55.

26. Rudolph, B., et al., *Evolution of Escherichia coli for growth at high temperatures.* J Biol Chem, 2010. **285**(25): p. 19029-34.

27. Caspeta, L., et al., *Biofuels. Altered sterol composition renders yeast thermotolerant.* Science, 2014. **346**(6205): p. 75-8.

28. Shi, D.J., C.L. Wang, and K.M. Wang, *Genome shuffling to improve thermotolerance, ethanol tolerance and ethanol productivity of Saccharomyces cerevisiae.* J Ind Microbiol Biotechnol, 2009. **36**(1): p. 139-47.

29. Bhandiwad, A., et al., *Metabolic engineering of Thermoanaerobacterium saccharolyticum for n-butanol production.* Metab Eng, 2014. **21**: p. 17-25.

30. Lin, P.P., et al., *Consolidated bioprocessing of cellulose to isobutanol using Clostridium thermocellum.* Metab Eng, 2015. **31**: p. 44-52.

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

32. Yue, H.T., et al., *A seawater-based open and continuous process for polyhydroxyalkanoates production by recombinant Halomonas campaniensis LS21 grown in mixed substrates.* Biotechnology for Biofuels, 2014. **7**(108): p. 108.

33. Voronovsky, A.Y., et al., *Development of strains of the thermotolerant yeast Hansenula polymorpha capable of alcoholic fermentation of starch and xylan.* Metab Eng, 2009. **11**(4-5): p. 234-42.

34. Lobs, A.K., et al., *High throughput, colorimetric screening of microbial ester biosynthesis reveals high ethyl acetate production from Kluyveromyces marxianus on C5, C6, and C12 carbon sources.* Biotechnology Journal, 2016. **11**(10): p. 1274-1281.

35. Xiao, H., et al., *Exploiting Issatchenkia orientalis SD108 for succinic acid production.* Microb Cell Fact, 2014. **13**: p. 121.

36. Lobs, A.K., et al., *Highly Multiplexed CRISPRi Repression of Respiratory Functions Enhances Mitochondrial Localized Ethyl Acetate Biosynthesis in Kluyveromyces marxianus.* ACS Synth Biol, 2018. **7**(11): p. 2647-2655.

37. Abdel-Banat, B.M., et al., *High-temperature fermentation: how can processes for ethanol production at high temperatures become superior to the traditional process using mesophilic yeast?* Appl Microbiol Biotechnol, 2010. **85**(4): p. 861-7.

38. Cernak, P., et al., *Engineering Kluyveromyces marxianus as a Robust Synthetic Biology Platform Host.* MBio, 2018. **9**(5): p. e01410-18.

39. Cheon, Y., et al., *A biosynthetic pathway for hexanoic acid production in Kluyveromyces marxianus.* J Biotechnol, 2014. **182-183**: p. 30-6.

40. Dmytruk, K., et al., *Development of the Thermotolerant Methylotrophic Yeast Hansenula polymorpha as Efficient Ethanol Producer*, in *Yeast Diversity in Human Welfare*, T. Satyanarayana and G. Kunze, Editors. 2017, Springer Singapore: Singapore. p. 257-282.

41. Lenton, S., et al., *Structural evidence for solvent-stabilisation by aspartic acid as a mechanism for halophilic protein stability in high salt concentrations.* Phys Chem Chem Phys, 2016. **18**(27): p. 18054-62.

42. Oren, A., *Microbial life at high salt concentrations: phylogenetic and metabolic diversity.* Saline Systems, 2008. **4**(1): p. 2.

43. Padan, E., et al., *Na(+)/H(+) antiporters.* Biochim Biophys Acta, 2001. **1505**(1): p. 144-57.

44. Purvis, J.E., L.P. Yomano, and L.O. Ingram, *Enhanced trehalose production improves growth of Escherichia coli under osmotic stress.* Appl Environ Microbiol, 2005. **71**(7): p. 3761-9.

45. Yang, L., et al., *A primary sodium pump gene of the moderate halophile Halobacillus dabanensis exhibits secondary antiporter properties.* Biochem Biophys Res Commun, 2006. **346**(2): p. 612-7.

46. Ekberg, J., et al., *Adaptive evolution of the lager brewing yeast Saccharomyces pastorianus for improved growth under hyperosmotic conditions and its influence on fermentation performance.* FEMS Yeast Res, 2013. **13**(3): p. 335-49.

47. Schweikhard, E.S., et al., *Structure and function of the universal stress protein TeaD and its role in regulating the ectoine transporter TeaABC of Halomonas elongata DSM 2581(T).* Biochemistry, 2010. **49**(10): p. 2194-204.

48. Kunte, H., G. Lentzen, and E. Galinski, *Industrial Production of the Cell Protectant Ectoine: Protection Mechanisms, Processes, and Products.* Current Biotechnology, 2014. **3**(1): p. 10-25.

49. Chen, X., et al., *Reprogramming Halomonas for industrial production of chemicals.* J Ind Microbiol Biotechnol, 2018. **45**(7): p. 545-554.

50. Qin, Q., et al., *CRISPR/Cas9 editing genome of extremophile Halomonas spp.* Metab Eng, 2018. **47**: p. 219-229.

51. Chen, R., et al., *Optimization of the extraction and purification of the compatible solute ectoine from Halomonas elongate in the laboratory experiment of a commercial production project.* World J Microbiol Biotechnol, 2017. **33**(6): p. 116.

52. Zaky, A.S., et al., *The establishment of a marine focused biorefinery for bioethanol production using seawater and a novel marine yeast strain.* Sci Rep, 2018. **8**(1): p. 12127.

53. Yuan, W.J., et al., *Ethanol fermentation with Kluyveromyces marxianus from Jerusalem artichoke grown in salina and irrigated with a mixture of seawater and freshwater.* J Appl Microbiol, 2008. **105**(6): p. 2076-83.

54. Ramos, J.L., et al., *Efflux pumps involved in toluene tolerance in Pseudomonas putida DOT-T1E.* J Bacteriol, 1998. **180**(13): p. 3323-9.

55. Ruhl, J., A. Schmid, and L.M. Blank, *Selected Pseudomonas putida strains able to grow in the presence of high butanol concentrations.* Appl Environ Microbiol, 2009. **75**(13): p. 4653-6.

56. Ramos, J.L., et al., *Mechanisms for solvent tolerance in bacteria.* J Biol Chem, 1997. **272**(7): p. 3887-90.

57. Rojas, A., et al., *Three efflux pumps are required to provide efficient tolerance to toluene in Pseudomonas putida DOT-T1E.* J Bacteriol, 2001. **183**(13): p. 3967-73.

58. Teran, W., et al., *Complexity in efflux pump control: cross-regulation by the paralogues TtgV and TtgT.* Mol Microbiol, 2007. **66**(6): p. 1416-28.

59. Segura, A., et al., *Proteomic analysis reveals the participation of energy- and stress-related proteins in the response of Pseudomonas putida DOT-T1E to toluene.* J Bacteriol, 2005. **187**(17): p. 5937-45.

60. Dunlop, M.J., et al., *Engineering microbial biofuel tolerance and export using efflux pumps.* Mol Syst Biol, 2011. **7**: p. 487.

61. Tan, Z., et al., *Membrane engineering via trans unsaturated fatty acids production improves Escherichia coli robustness and production of biorenewables.* Metab Eng, 2016. **35**: p. 105-113.

62. Zingaro, K.A. and E. Terry Papoutsakis, *GroESL overexpression imparts Escherichia coli tolerance to i-, n-, and 2-butanol, 1,2,4-butanetriol and ethanol with complex and unpredictable patterns.* Metab Eng, 2013. **15**: p. 196-205.

63. Cook, T.B., et al., *Genetic tools for reliable gene expression and recombineering in Pseudomonas putida.* J Ind Microbiol Biotechnol, 2018. **45**(7): p. 517-527.

64. Gong, T., et al., *Metabolic Engineering of Pseudomonas putida KT2440 for Complete Mineralization of Methyl Parathion and gamma-Hexachlorocyclohexane.* ACS Synth Biol, 2016. **5**(5): p. 434-42.

65. Nikel, P.I. and V. de Lorenzo, *Pseudomonas putida as a functional chassis for industrial biocatalysis: From native biochemistry to trans-metabolism.* Metab Eng, 2018. **50**: p. 142-155.

66. Bormann, S., et al., *Engineering Clostridium acetobutylicum for production of kerosene and diesel blendstock precursors.* Metab Eng, 2014. **25**: p. 124-30.

67. Anbarasan, P., et al., *Integration of chemical catalysis with extractive fermentation to produce fuels.* Nature, 2012. **491**(7423): p. 235-9.

68. Zhang, J., et al., *Enhanced acid tolerance in Lactobacillus casei by adaptive evolution and compared stress response during acid stress.* Biotechnology and Bioprocess Engineering, 2012. **17**(2): p. 283-289.

69. Zhou, L., et al., *Improvement of D-lactate productivity in recombinant Escherichia coli by coupling production with growth.* Biotechnol Lett, 2012. **34**(6): p. 1123-30.

70. Lee, J.Y., et al., *Engineering cellular redox balance in Saccharomyces cerevisiae for improved production of L-lactic acid.* Biotechnol Bioeng, 2015. **112**(4): p. 751-8.

71. Yan, D., et al., *Construction of reductive pathway in Saccharomyces cerevisiae for effective succinic acid fermentation at low pH value.* Bioresour Technol, 2014. **156**: p. 232-9.

72. Kwon, Y.D., et al., *Long-term continuous adaptation of Escherichia coli to high succinate stress and transcriptome analysis of the tolerant strain.* J Biosci Bioeng, 2011. **111**(1): p. 26-30.

73. Wright, J., et al., *Batch and continuous culture-based selection strategies for acetic acid tolerance in xylose-fermenting Saccharomyces cerevisiae.* FEMS Yeast Res, 2011. **11**(3): p. 299-306.

74. Sandoval, N.R., et al., *Elucidating acetate tolerance in E. coli using a genome-wide approach.* Metab Eng, 2011. **13**(2): p. 214-24.

75. Suominen, P., et al., *Genetically Modified Yeast Of The Species Issatchenkia Orientalis And Closely Relates Species, And Fermentation Processes Using Same*. 2012, CARGILL INC: US.

76. Valdes, J., et al., *Acidithiobacillus ferrooxidans metabolism: from genome sequence to industrial applications.* BMC Genomics, 2008. **9**: p. 597.

77. Kernan, T., et al., *Engineering the iron-oxidizing chemolithoautotroph Acidithiobacillus ferrooxidans for biochemical production.* Biotechnol Bioeng, 2016. **113**(1): p. 189-97.

78. Inaba, Y., et al., *Transposase-Mediated Chromosomal Integration of Exogenous Genes in Acidithiobacillus ferrooxidans.* Appl Environ Microbiol, 2018. **84**(21): p. e01381-18.

79. Maezato, Y., et al., *Metal resistance and lithoautotrophy in the extreme thermoacidophile Metallosphaera sedula.* J Bacteriol, 2012. **194**(24): p. 6856-63.

80. Auernik, K.S., et al., *The genome sequence of the metal-mobilizing, extremely thermoacidophilic archaeon Metallosphaera sedula provides insights into bioleaching-associated metabolism.* Appl Environ Microbiol, 2008. **74**(3): p. 682-92.

81. Zeldes, B.M., et al., *Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals.* Front Microbiol, 2015. **6**: p. 1209.

82. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**: p. 121.

83. Forster, J., et al., *Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.* Genome Res, 2003. **13**(2): p. 244-53.

84. Wang, H.H., et al., *Programming cells by multiplex genome engineering and accelerated evolution.* Nature, 2009. **460**(7257): p. 894-U133.

85. Warner, J.R., et al., *Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides.* Nat Biotechnol, 2010. **28**(8): p. 856-62.

86. Lee, M.E., et al., *A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly.* ACS Synth Biol, 2015. **4**(9): p. 975-86.

87. Xu, P., et al., *ePathBrick: a synthetic biology platform for engineering metabolic pathways in E. coli.* ACS Synth Biol, 2012. **1**(7): p. 256-66.

88. Lobs, A.K., C. Schwartz, and I. Wheeldon, *Genome and metabolic engineering in non-conventional yeasts: Current advances and applications.* Synth Syst Biotechnol, 2017. **2**(3): p. 198-207.

89. Schwartz, C., et al., *CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica.* Biotechnol Bioeng, 2017. **114**(12): p. 2896-2906.

90. Cao, M., et al., *CRISPR-Mediated Genome Editing and Gene Repression in Scheffersomyces stipitis.* Biotechnol J, 2018. **13**(9): p. e1700598.

91. Schwartz, C., et al., *Standardized Markerless Gene Integration for Pathway Engineering in Yarrowia lipolytica.* ACS Synth Biol, 2017. **6**(3): p. 402-409.

92. Schwartz, C., et al., *Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast Yarrowia lipolytica.* Metab Eng, 2019. **55**: p. 102-110.

93. Schwartz, C.M., et al., *Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica.* ACS Synth Biol, 2016. **5**(4): p. 356-9.

94. Cao, M., et al., *Centromeric DNA Facilitates Nonconventional Yeast Genetic Engineering.* ACS Synth Biol, 2017. **6**(8): p. 1545-1553.

95. Lewis, N.E., H. Nagarajan, and B.O. Palsson, *Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.* Nat Rev Microbiol, 2012. **10**(4): p. 291-305.

96. Schwalen, C.J., et al., *Bioinformatic Expansion and Discovery of Thiopeptide Antibiotics.* J Am Chem Soc, 2018. **140**(30): p. 9494-9501.

97. Brophy, J.A.N., et al., *Engineered integrative and conjugative elements for efficient and inducible DNA transfer to undomesticated bacteria.* Nat Microbiol, 2018. **3**(9): p. 1043-1053.

98. Schwartz, C., et al., *Multiplexed CRISPR Activation of Cryptic Sugar Metabolism Enables Yarrowia Lipolytica Growth on Cellobiose.* Biotechnol J, 2018. **13**(9): p. e1700584.

99. Boundy-Mills, K.L., et al., *Yeast culture collections in the twenty-first century: new opportunities and challenges.* Yeast, 2016. **33**(7): p. 243-60.

100. Solomon, K.V., et al., *Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes.* Science, 2016. **351**(6278): p. 1192-5.

Chapter 2: Selecting a strain of *Pseudomonas chlororaphis* as a microbial host for phenazine production

Abstract

Strain selection is an important part of bioprocess design; a host should be chosen which has industrially advantageous traits, such as thermotolerance, tolerance to toxic products, favorable growth characteristics, and the ability to produce high titers of the desired product. This work focuses on identifying a strain of *Pseudomonas chlororaphis* as a host to produce phenazine compounds, as both the bacteria itself and the phenazines it produces have biocontrol activity against many devastating fungal plant pathogens. While some *P. chlororaphis* strains naturally produce high titers of phenazines, they may also have traits which are less favorable for bioprocessing, such as biofilm formation and low growth temperatures. To identify an optimal strain of *P. chlororaphis* for phenazine production, I characterized phenazine production, growth temperature, and biofilm formation in 34 *Pseudomonas* isolates purchased from 3 international culture collections. Strain DSM 21509 was determined to be the best host strain for phenazine production as it had the highest phenazine titers (477 ± 163 mg/L phenazine-1-carboxamide) and low biofilm formation in both tested media conditions.

Introduction

Recent advances in synthetic biology, bioinformatics, and next-generation sequencing allow non-conventional microbes to be developed as biotechnology hosts. Synthetic biology toolsets and annotated genome assemblies can be generated quickly for many novel strains, allowing them to be engineered for industrial use. Rather than choosing from a handful of well-characterized strains, strains can now be selected based on their natural properties which make them suited for producing a desired product and for bioprocessing in general. Selecting strains with process-beneficial phenotypes (i.e., thermotolerance, toxic product tolerance, high product titers) can lead to associated bioprocess benefits such as enhanced separations, non-aseptic conditions, and reduced cooling costs. Screening a collection of potential host strains for these phenotypes can allow researchers to select the best host strain for their metabolic engineering studies.

This work seeks to identify a microbial host for phenazines, a class of heterocyclic redox-active organic compounds with current applications as agricultural fungicides and potential applications within bioelectrochemical devices. Due to their antibiotic activities, a host must be chosen which can tolerate these toxic products. Some pseudomonads, Gram-negative rod-shaped bacteria belonging to the genus *Pseudomonas*, naturally produce phenazines as secondary metabolites. Pseudomonads make promising biotechnology hosts due to their innate tolerance to many aromatic organic solvents and toxic products[1, 2] as well as their many available synthetic biology tools[3] and genomic resources[4].

Some pseudomonads have phenotypes which are unfavorable for bioprocessing as well. While some *Pseudomonas* strains can grow at temperatures greater than or equal to 43 °C, many strains can only grow at low temperatures around 30 °C [5]. Higher growth temperatures are preferable for bioprocessing because operating at higher fermentation temperatures could reduce costs associated with bioreactor cooling. Many researchers are interested in biofilm formation in pseudomonads, as biofilm formation in human and plant pathogens such as *P. aeruginosa* can lead to persistent, antibiotic-resistant infections[6, 7]. Biofilm formation is a phenotype of interest for this study due to its concerns for industrial biotechnology, as biofilms can be difficult to sanitize and remove from industrial bioreactors[8].

This work performs a high-throughput phenotype screening of a collection of *Pseudomonas chlororaphis* strains for phenazine production, biofilm formation, and growth temperature with the goal of selecting an industrial biotechnology host strain for phenazine production. Not only does *P. chlororaphis* naturally produce multiple phenazine derivatives, but it is non-pathogenic to humans and plants and commercially available as a biocontrol product. It already demonstrates potential as a phenazine production host, as it has been successfully engineered to produce phenazines and other products within bioreactors[9, 10]. By characterizing and comparing these phenotypes, we hope to identify a strain of *Pseudomonas chlororaphis* to use in future work as a phenazine production host.

Materials and methods

Strain selection and culturing

All strains designated as *Pseudomonas chlororaphis* that were available as of April and October 2019 were ordered from the American Type Culture Collection (ATCC; Manassas, VA); all strains designated as *Pseudomonas chlororaphis* that were not already purchased and available as of March 2020 were ordered from the German Collection of Microorganisms and Cell Cultures (DSMZ GmbH; Braunschweig, Germany) and the Westerdijk Fungal Biodiversity Institute's Netherlands Culture Collection of Bacteria (NCCB; Utrecht, Netherlands). Strains which appeared to have more than one colony morphology were separated into distinct isolates (denoted by [1] and [2], which were arbitrarily assigned). All isolates were sequenced with 16s rRNA sequencing (GENEWIZ®; South Plainfield, NJ), and the 33 confirmed *P. chlororaphis* isolates were used in this study (Table 2.1). One PCA-producing isolate identified as *Pseudomonas synxantha* from the 16s sequencing was also included in this study as a phylogenetic outgroup, bringing the total to 34 isolates used in this study.

Strains were initially revived according to the guidance of each culture collection then subsequently cultured at 30 °C in liquid King's Media B (KMB) or on solid King's Media B agar which was prepared according to the methods of King et al.[11]. Briefly, 16 g Bactopeptone, 8 mL glycerol, and 8 mL 15% (w/v) potassium phosphate dibasic (K_2HPO_4) were dissolved in 780 mL milliQ water and the pH was adjusted to 7.2 using hydrochloric acid (HCl). To prevent precipitation, 8 mL of sterile 15% (w/v)

Table 2.1. List of strains used in this study. All strains were confirmed as *P. chlororaphis* with 16 rRNA sequencing with the exception of ATCC 17413 which was determined to be *P. synxantha*.

Strain name	Isolate used in this study	Source
Pseudomonas chlororaphis subsp. *chlororaphis* (ATCC® 9446™)	ATCC 9446	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 9447™)	ATCC 9447	American Type Culture Collection
Pseudomonas chlororaphis subsp. *aureofaciens* (ATCC® 13985™)	ATCC 13985	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 13986™)	ATCC 13986[1]	American Type Culture Collection
	ATCC 13986[2]	
Pseudomonas chlororaphis (ATCC® 15926™)	ATCC 15926	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17411™)	ATCC 17411	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17413™)	ATCC 17413	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17414™)	ATCC 17414	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17415™)	ATCC 17415[1]	American Type Culture Collection
	ATCC 17415[2]	
Pseudomonas chlororaphis (ATCC® 17417™)	ATCC 17417	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17418™)	ATCC 17418[1]	American Type Culture Collection
	ATCC 17418[2]	
Pseudomonas chlororaphis (ATCC® 17419™)	ATCC 17419	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17809™)	ATCC 17809	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17810™)	ATCC 17810	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17811™)	ATCC 17811	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17814™)	ATCC 17814	American Type Culture Collection
Pseudomonas chlororaphis subsp. *aurantiaca* (ATCC® 33663™)	ATCC 33663[1]	American Type Culture Collection
	ATCC 33663[2]	
Pseudomonas chlororaphis subsp. *aureofaciens* DSM-6508	DSM 6508	German Collection of Microorganisms and Cell Cultures
Pseudomonas chlororaphis subsp. *piscium* DSM-21509	DSM 21509	German Collection of Microorganisms and Cell Cultures
Pseudomonas chlororaphis subsp. *aureofaciens* DSM-29578	DSM 29578[1]	German Collection of Microorganisms and Cell Cultures
	DSM 29578[2]	
Pseudomonas chlororaphis subsp. *chlororaphis* NCCB 47033	NCCB 47033	Netherlands Culture Collection of Bacteria
Pseudomonas chlororaphis subsp. *aureofaciens* NCCB 60037	NCCB 60037	Netherlands Culture Collection of Bacteria
Pseudomonas chlororaphis subsp. *aureofaciens* NCCB 60038	NCCB 60038	Netherlands Culture Collection of Bacteria
Pseudomonas chlororaphis subsp. *aureofaciens* NCCB 82053	NCCB 82053[1]	Netherlands Culture Collection of Bacteria
	NCCB 82053[2]	
Pseudomonas chlororaphis subsp. *chlororaphis* NCCB 88062	NCCB 88062[1]	Netherlands Culture Collection of Bacteria
	NCCB 88062[2]	
Pseudomonas chlororaphis NCCB 100368	NCCB 100368[1]	Netherlands Culture Collection of Bacteria
	NCCB 100368[2]	

magnesium sulfate heptahydrate (MgSO4·7H2O) was added to the media after autoclaving the other components together. Solid media also contained 15 g/L agar. King's Media B + Fe (KMB+Fe) was created based off the work of van Rij et al., where supplementing KMB with ferric iron ions in the form of sodium ferric ethylenediaminetetraacetate (FeNaEDTA) increased phenazine production in *P. chlororaphis*[12]. KMB+Fe was prepared by adding 8 mL sterile 10 mM FeNaEDTA to KMB after autoclaving, accounting for the volume increase. For phenazine production experiments, KMB and KMB+Fe contained 1.5 g/L $MgSO_4$ instead of $MgSO_4 \cdot 7H_2O$.

Liquid culturing was done using sterile 2 mL 96-deep well plates within an INFORS HT Multitron Pro plate shaker incubator at 1000 rpm and ~88% humidity. Overnight cultures were started by inoculating 500 µL media of interest with the respective colony or glycerol stock. After the overnight culture incubated with shaking at 30°C for 22-24 hours, the plate was spun down in Beckman Coulter Allegra 25R centrifuge for 10 minutes at 5,000 rcf. To reduce phenazine transfer and to ensure biofilm-forming strains were well-mixed, old media was removed, and cultures were resuspended in fresh media. To start experimental cultures, 500 µL of desired media was inoculated with 10 µL of resuspended respective overnight culture.

Phenazine production

After 48 hours of liquid culture at 30 °C, phenazine compounds were extracted from each culture using ethyl acetate liquid-liquid extraction. Whole cultures were acidified with 10 µL of 3 M HCl, then 1.2 mL ethyl acetate was added to each culture and mixed well by pipetting. Each mixture was transferred to a microcentrifuge tube,

vortexed at maximum speed for 1 minute, and spun down to separate liquid phases. 900 µL of each ethyl acetate phase was transferred to a 96-well plate and evaporated overnight. The extract was resuspended in either 200 µL methanol to concentrate the sample or 900 µL methanol and filtered for quantification via HPLC.

All phenazines were quantified with a photodiode-array detector on a Shimadzu Nexera-i LC-2040C 3D liquid chromatograph with an Agilent Poroshell 120 EC-C18 2.7 µm 3.0 x 75 mm column and 3.0 mm x 5.0 mm guard column at 40 °C. To resolve such similar compounds, the following method with gradients of methanol and ammonium acetate buffer (pH 5.0) was used: 2 µL sample injection, 5 min of 20% methanol, 2 min gradient from 20% to 30% methanol, and 13 min gradient from 30 to 40% methanol with subsequent steps to wash and re-equilibrate the column. All steps had a 1 mL/min flow rate. PCA and PCN peaks were identified by comparing retention times to those of purchased PCA and PCN (ChemScene; Monmouth Junction, NJ). Because pure 2-HP and 2-HPCA were not commercially available, the identities of these HPLC peaks were confirmed with LC-MS following the same protocol. Phenazines were quantified by converting peak areas at a wavelength of 254 nm and bandwidth of 4 nm to concentrations using extinction constants calculated from the purchased PCA and PCN.

Biofilm and growth characterization

Biofilm formation phenotyping was characterized using a crystal violet staining assay following the protocol of George A. O'Toole[13]. A 2% inoculum of overnight culture in the respective media was used to start 100 µL stationary cultures which were incubated without agitation at 30 °C for 48 hours in sterile non-culture-treated

polystyrene 96-well plates with lid. Loose cells were rinsed away with milliQ water and blotted dry with paper towel 3 times, and the remaining biofilm was stained with 150 µL 0.4% crystal violet solution for 15 minutes. After rinsing off the excess dye with water and blotting dry with paper towel 3 times, the plates were dried upside down overnight. 150 µL 30% (v/v) acetic acid was added to each well and incubated for 15 minutes. The absorbance was measured at 550 nm on a BioTek Synergy Neo2 plate reader after transferring the acetic acid solution to a new microtiter plate. Growth temperature phenotyping was performed by streaking overnight cultures onto the respective solid media in patches and incubating the plates at either 30 ℃ or 37 ℃ for 48 hours. Plates were assessed visually for growth based on the density of the culture patch.

Results

I purchased all non-redundant strains of *Pseudomonas chlororaphis* which were available from the international culture collections that were accessible to our institution, resulting in 26 strains purchased from three culture collections (Table 2.1). Upon receipt, eight of these strains had two distinct colony morphologies present which appeared to vary in pigment production. An example of a strain with these distinct morphologies is ATCC 13986 (Figure 2.1). After a few days of culture, one morphology (ATCC 13986[1]) produced orange pigment which diffused into the media whereas the other (ATCC 13986[2]) produced bright yellow diffusible pigment. Because these variants likely differ in the phenotypes of interest for this study and may have genetic variation associated with those differing phenotypes, they were treated as separate isolates moving forward (arbitrarily denoted by [1] and [2]).

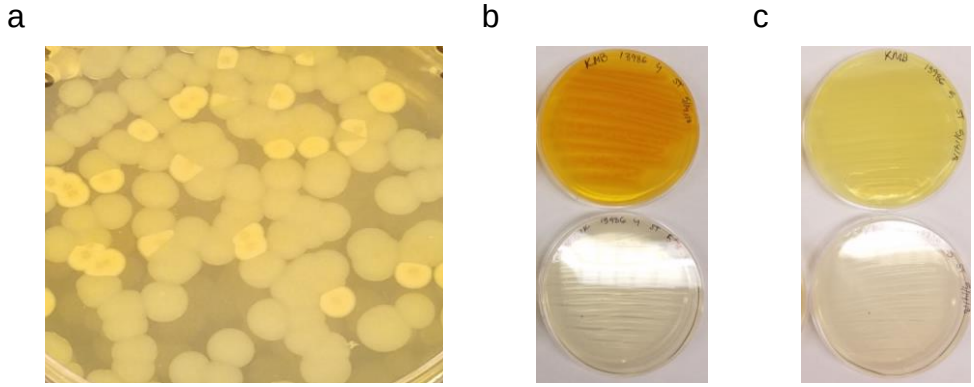

Figure 2.1. Photographs of ATCC 13986 isolates cultured on King's Media B. a) ATCC 13986 upon initial culturing had 2 distinct colony morphologies present, so this strain was separated into 2 isolates: b) ATCC 13986[1] and c) ATCC 13986[2]. In b) and c) the respective strain is streaked onto solid KMB (top) and KMB + 50 µg/mL kanamycin sulfate (bottom). After 6 days of growth (pictured), there was distinctly different pigment production on KMB and no growth on KMB + 50 µg/L kanamycin sulfate. Both ATCC 13986[1] and ATCC 13986[2] were confirmed as *P. chlororaphis* with 16 rRNA sequencing.

Even though all purchased strains were labeled as *P. chlororaphis*, all isolates were 16s rRNA sequenced to confirm their species identity. Many of these strains were initially isolated and deposited into their respective culture collections over 50 years ago[5, 14] when there were different methodologies and criteria for classifying various *Pseudomonas* species[11, 15, 16]. Additionally in the cases with multiple colony morphologies, 16s sequencing would confirm that the variants were in fact variants with different phenotypes rather than plate contaminants. 33 isolates were confirmed as *P. chlororaphis*, but isolate ATCC 17413 was identified as *Pseudomonas synxantha*. Due to its ability to produce phenazine-1-carboxylic acid, ATCC 17413 was still included in this study where it could serve as a phylogenetic outgroup for any future comparative genomics studies. All this culturing yielded a total of 34 *Pseudomonas* isolates for this study.

For the phenazine production phenotyping, I quantified the 4 phenazine compounds naturally produced by *Pseudomonas chlororaphis*: 2-hydroxyphenazine (2-HP), 2-hydroxyphenazine-1-carboxylic acid (2-HPCA), phenazine-1-carboxylic acid (PCA) and phenazine-1-carboxamide (PCN) (Figure 2.2). In pseudomonads, phenazines are produced by a highly conserved biosynthesis operon where *phzABDEFG* are responsible for converting chorismic acid generated by the shikimate pathway into PCA. *Pseudomonas chlororaphis* isolates have either *phzH* or *phzO* immediately downstream of *phzABCDEFG* which convert PCA into PCN or 2-HPCA, respectively. 2-HP is a product of the spontaneous decarboxylation of 2-HPCA.

I first characterized phenazine production in King's Media B (KMB), the standard culture media for fluorescent pseudomonads (Figure 2.3). Under these conditions, fewer than half the isolates produced more than 10 mg/L of total phenazines. These low titers suggest that phenotyping in KMB may underestimate the phenazine production capacity of this strain collection. Because accurately characterizing phenazine production is very important for selecting a phenazine-production strain and for the future mGWAS analysis, I sought to improve phenazine production by supplementing KMB with ferric iron in the form of FeNaEDTA which has been previously reported to enhance phenazine production in some strains of *P. chlororaphis*[12, 17]. KMB+Fe media improved total phenazine production in 24 isolates, and 10 isolates did not produce significant phenazines in either medium. Due to its positive effects for most of the strains and its neutral effects on the remaining strains, I did additional phenotyping in KMB+Fe as well as KMB.

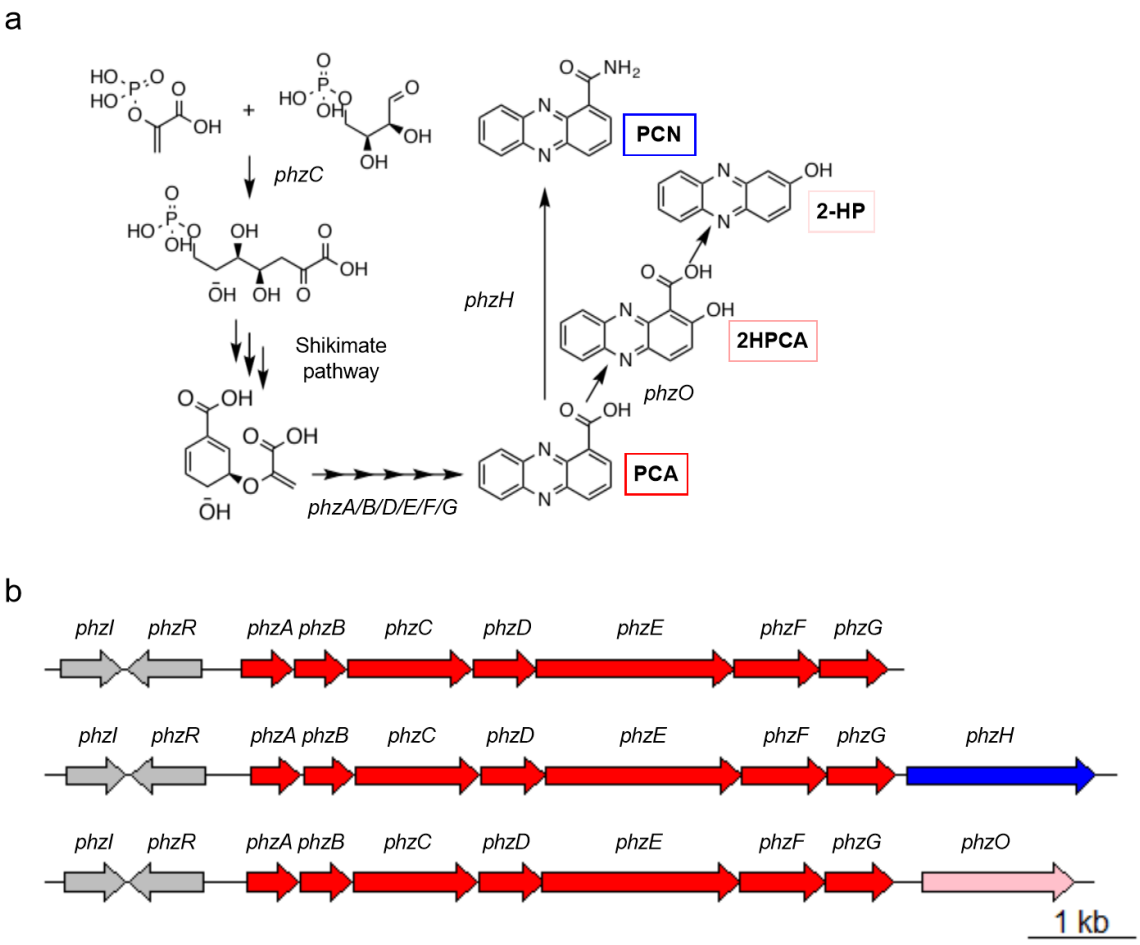

Figure 2.2. Phenazine biosynthesis pathway and operon in *P. synxantha* and *P. chlororaphis.*
a) Phenazine biosynthesis pathway in *P. chlororaphis*. *Pseudomonas chlororaphis* naturally produces 4 phenazines: 2-hydroxyphenazine (2-HP), 2-hydroxyphenazine-1-carboxylic acid (2-HPCA), phenazine-1-carboxylic acid (PCA) and phenazine-1-carboxamide (PCN). Chorismate from the shikimate pathway is converted into PCA in all strains. PCA can be converted into PCN or acid 2-HPCA by PhzH or PhzO, respectively. 2-HP is a byproduct of the spontaneous decarboxylation of 2-HPCA. b) In fluorescent pseudomonads, phenazines are produced by a highly conserved core phenazine biosynthesis operon. *phzI/phzR* encode two-component quorum-sensing system which regulates the expression of the following *phz* genes. *phzABCDEFG* are responsible for the production of PCA. Strains of *Pseudomonas chlororaphis* contain either *phzO,* which converts PCA into 2-HPCA (which spontaneously decomposes into 2-HP), or *phzH*, which converts PCA into PCN carboxamide (PCN). *phzH* or *phzO* occurs immediate downstream of *phzG*.

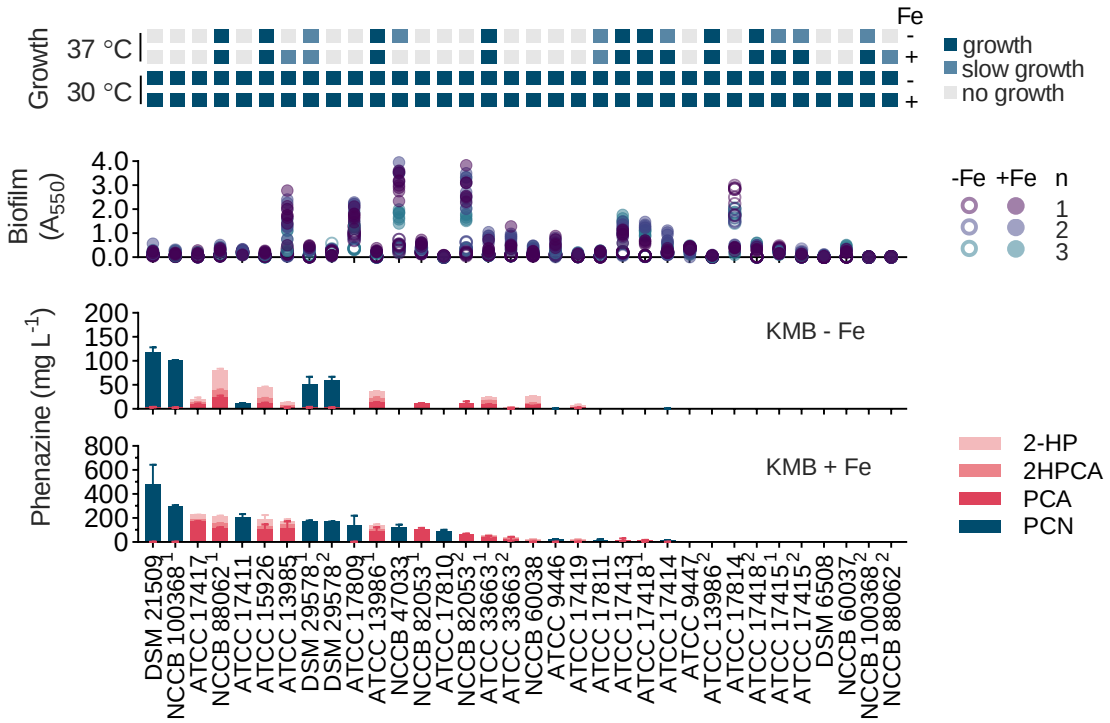

Figure 2.3. Phenazine production, biofilm formation, and growth temperature phenotype characterization for all isolates used in this study. All phenotyping data was collected after 48 hours of culture in either King's Media B (KMB-Fe, -Fe) or King's Media B + Fe (KMB+Fe, +Fe). 2-hydroxyphenazine (2-HP), 2-hydroxyphenazine-1-carboxylic acid (2HPCA), phenazine-1-carboxylic acid (PCA) and phenazine-1-carboxamide (PCN) were quantified using HPLC.

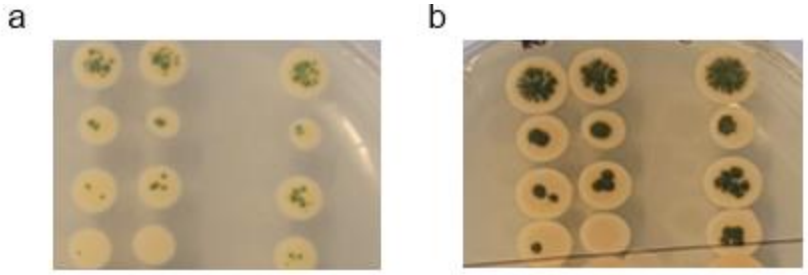

Figure 2.4. Photographs of DSM 21509 pigment production on solid King's Media B + Fe. These photographs are of the same plate after a) 2 days and b) 5 days of culture at 30 °C. Each column is a replicate, and each row is a dilution series with the top row as 2 µL of overnight culture and each spot moving down is a 1/10 dilution.

The PCN-producers primarily produced PCN, with only very small amounts of the PCA precursor detected (< 5 mg/L), if at all. Four isolates (DSM 21509, NCCB 100368, DSM 29578[1], DSM 29578[2]) produced at least 45 mg/L PCN in both media conditions. Isolate ATCC 17411 produced high levels of PCN in KMB+Fe (247 ± 66 mg/L) as well but not in KMB (12 ± 0.5 mg/L). Among all strains, the top producer of PCN and total phenazines in both media conditions was DSM 21509, which produced 477 ± 163 mg/L PCN in KMB+Fe and 114 ± 11 mg/L PCN in KMB. Figure 2.4 shows the large amounts of non-diffusible green pigment strain DSM 21509 produces on solid KMB+Fe, which visually confirms its capacity to produce large quantities of PCN.

Strain ATCC 17417 produced the most PCA (171 ± 4 mg/L) and combined PCA/2-HPCA/2-HP (~228 mg/L) in KMB+Fe. Strain ATCC 15926 produced the most 2-HP (59 ± 36 mg/L) in KMB+Fe with strain NCCB 88062[1] close behind (58 ± 6 mg/L 2-HP). Strain NCCB 88062[1] produced the most 2-HPCA (35 ± 2 mg/L) in KMB+Fe and the most PCA (25 ± 2 mg/L), 2-HPCA (14 ± 2 mg/L), and 2-HP (41 ± 4 mg/L) and combined PCA/2-HPCA/2-HP (~80 mg/L) in KMB. When considering the performance in both media conditions, NCCB 88062[1] was the top producer of PCA/2-HPCA/2-HP.

I characterized growth on solid media for all strains at 30 °C and 37 °C, which are fermentation temperatures for many common microbial hosts. These temperatures are appropriate for *P. chlororaphis* as it is typically cultured at 28 °C or 30 °C, and some strains have been reported to grow at up to 37 °C[18]. All isolates could grow at 30 °C on both media, so all other culturing and phenotyping was performed at this temperature (Figure 2.4). At 37 °C about half the strains were capable of growth but no colorful

pigment was observed on either media (Figure 2.5). Overall, biofilm formation was a lot lower in KMB than KMB+Fe. Only two strains (ATCC 17809 and ATCC 17814) have biofilm formation over $A_{550} = 1$ in KMB whereas 5 strains had biofilm formation above the same absorbance cutoff in KMB+Fe. Notably strain ATCC 17814 had high biofilm formation in KMB, but not KMB+Fe.

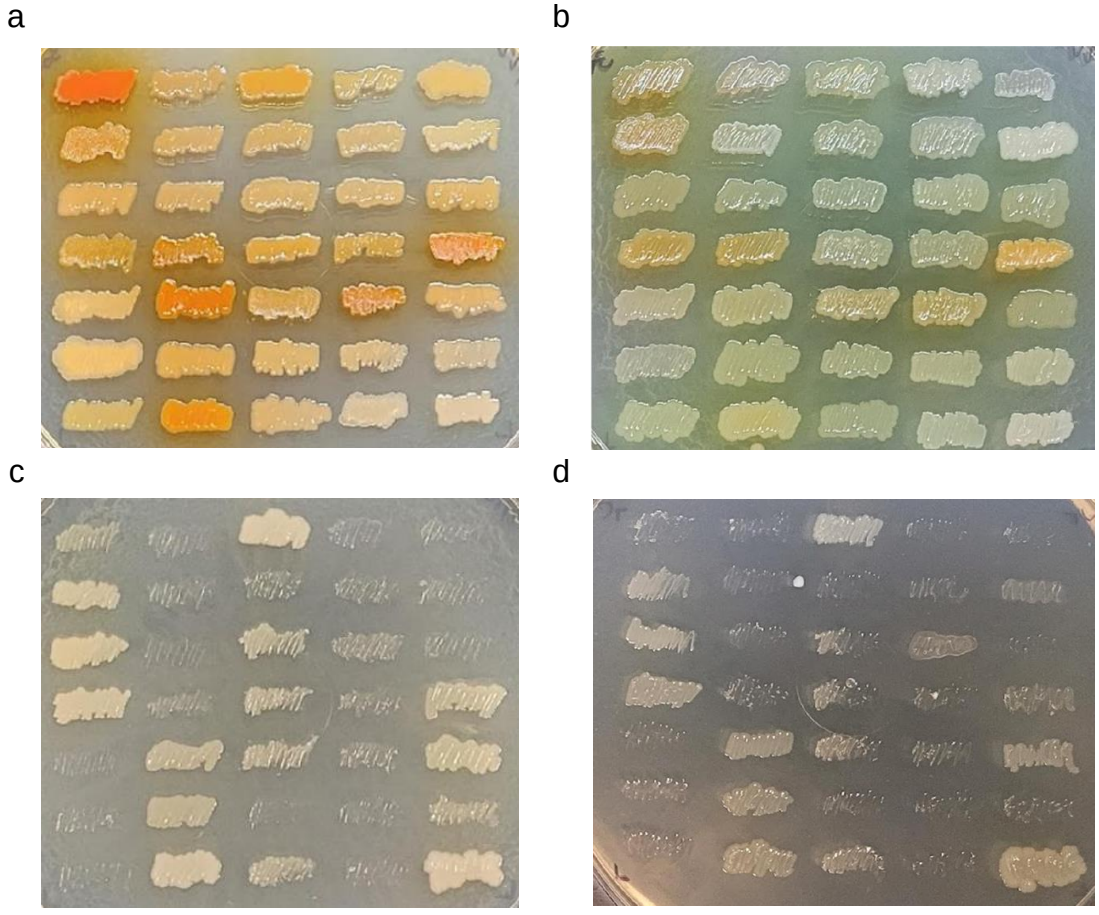

Figure 2.5. Photographs of all isolates streaked on King's Media B agar and King's Media B + Fe agar plates after 48 hours of culture at 30 °C and 37 °C. Growth on solid King's Media B + Fe (a, c) and King's Media B (b, d) at 30 °C (a, b) and 37 °C (c, d). In each photograph, strains are listed appear in the following order moving down each column starting in the lefthand column: ATCC 13985, ATCC 13986[1], ATCC 13986[2], ATCC 15926, ATCC 17411, ATCC 17809, ATCC 17810, ATCC 17814, ATCC 9447, ATCC 9446, ATCC 17419, ATCC 17418[1], ATCC 17418[2], ATCC 33663[1], ATCC 33663[2], ATCC 17414, ATCC 17415[1], ATCC 17415[2], ATCC 17417, ATCC 17811, DSM 6508, DSM 21509, DSM 29578[1], DSM 29578[2], NCCB 60037, NCCB 60038, NCCB 82053[1], NCCB 82053[2], NCCB 47033, NCCB 100368[1], NCCB 100368[2], NCCB 88062[1], NCCB 88062[2], ATCC 17413. The strain in the bottom righthand corner is *P. putida* KT2440 which served as a positive control.

Discussion

Out of all 34 strains we characterized, strain DSM 21509 is the best host strain overall for phenazine-1-carboxamide and total phenazine production due to its high titers of PCN and low biofilm formation in both media conditions. DSM 21509 is the type strain of *Pseudomonas chlororaphis* subsp. *piscium*, and it was isolated from the intestine of a European perch from Lake Neuchâtel, Switzerland, in 2005[19]. Strain NCCB 88062[1] was the best host strain for PCA/2-HPCA/2-HP when considering the production of all three compounds in both media conditions. This strain originated from the Netherlands and was deposited into NCCB in 1988 by E. van den Bosch, but there are no available publications [https://wi.knaw.nl/page/NCCB_strains_display/24262].

While these are the best *P. chlororaphis* phenazine production strains available from ATCC, DSMZ, or NCCB, there may be other strains available from other global culture collections or private laboratory stocks that produce higher titers of these phenazines. Wildtype *P. chlororaphis* HT66 (deposited into the China Center for Type Culture Collection, CCTCC) and LX24 have been reported to produce 424.87 mg/L PCN[20] and 158.6 mg/L 2-HP[21], respectively, in KMB at 28 °C which are the highest reported titers produced by wildtype strains to date.

Most phenotyping studies quantify phenazine production in a few strains, but a recent study screens a collection of 63 fluorescent *Pseudomonas* strains (including 38 *P. chlororaphis* strains and 6 *P. synxantha* strains) for phenazine production and biocontrol activity phenotypes[22]. That study also observed low phenazine production in KMB with large error bars even after 5 days of culture, with their top PCN, PCA, and 2-HP

producers making 80.5 μmol/L (~18 mg/L), 181 μmol/L (~43 mg/L), and <10 mg/L, respectively, and they were unable to detect 2-HPCA. Their phenotyping included a few of the same strains as this study, including DSM 21509 which produced only ~50 μmol/L (approximately 11 mg/L) PCN in KMB which is 10 times less that what I reported in KMB. These differences in phenazine production among studies once again highlight the importance of phenotyping a collection under uniform conditions when constructing a phenotype dataset and when drawing comparisons among strains, as phenazine formation is heavily dependent on environmental conditions.

Many environmental variables impact phenazine production profiles within fluorescent pseudomonads, including aeration[23], pH[24], electrochemical potential[25], and various mineral ions, such as iron, magnesium, zinc, and sulfate[23, 26]. Just by supplementing the media with one compound, FeNaEDTA, phenazine titers improved by hundreds of milligrams per liter for some strains in this study. Other studies have successfully improved phenazine production on the order of grams per liter within *P. chlororaphis* through media and fermentation optimization[27, 28]. Understanding the genetic link between environmental conditions and phenazine production could provide new metabolic engineering targets.

As can be seen in this study, these characterized phenotypes are not necessarily independent. Fermentation temperature could affect growth rate and biomass accumulation, which in turn affects phenazine production as expression of the *phz* operon is controlled by quorum sensing. Even though some *P. chlororaphis* strains can grow at 37 °C, they are unable to produce phenazines at that temperature. Most *P. chlororaphis*

strains stop producing phenazines around 34-35 °C, but *P. aeruginosa* can produce phenazines at 37 °C. Understanding the link between growth temperature and phenazine production could possibly allow the *P. chlororaphis* strains which can grow at higher temperatures to be engineered to produce phenazines those higher temperatures. This link could also be important for developing a biocontrol strain which is effective over a broader temperature range or for understanding virulence in *P. aeruginosa*.

Biofilm formation and phenazine production are connected as well. Both the PhzR/PhzI quorum sensing system, which directly regulates phenazine production, and phenazines themselves are required for biofilm formation in *P. chlororaphis* 30-84[29], so it makes sense that the strains made more biofilm in KMB+Fe where phenazine production was also improved. This relationship between a positive and negative phenotype should be kept in mind as a phenazine production host is developed.

While most of these strains have been phenotyped in some capacity, many do not have available quantitative phenazine production data or qualitative biofilm data or growth data at 37 °C[5]. This work provides phenotypic data about previously uncharacterized strains collected under uniform culture conditions which allows them to be compared to one another. Selecting a strain at the outset such as DSM 21509 which has naturally high phenazine production could reduce the number of metabolic engineering steps required for future strain development.

References

1. Rojas, A., et al., *Three efflux pumps are required to provide efficient tolerance to toluene in Pseudomonas putida DOT-T1E.* J Bacteriol, 2001. **183**(13): p. 3967-73.

2. Teran, W., et al., *Complexity in efflux pump control: cross-regulation by the paralogues TtgV and TtgT.* Mol Microbiol, 2007. **66**(6): p. 1416-28.

3. Cook, T.B., et al., *Genetic tools for reliable gene expression and recombineering in Pseudomonas putida.* J Ind Microbiol Biotechnol, 2018. **45**(7): p. 517-527.

4. Winsor, G.L., et al., *Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database.* Nucleic Acids Res, 2016. **44**(D1): p. D646-53.

5. Stanier, R.Y., N.J. Palleroni, and M. Doudoroff, *The aerobic pseudomonads: a taxonomic study.* J Gen Microbiol, 1966. **43**(2): p. 159-271.

6. Walker, T.S., et al., *Pseudomonas aeruginosa-plant root interactions. Pathogenicity, biofilm formation, and root exudation.* Plant Physiol, 2004. **134**(1): p. 320-31.

7. Whiteley, M., et al., *Gene expression in Pseudomonas aeruginosa biofilms.* Nature, 2001. **413**(6858): p. 860-4.

8. Samandoulgou, I., et al., *Murine Norovirus Interaction with Pseudomonas aeruginosa Biofilm in a Dynamic Bioreactor.* Food Environ Virol, 2021. **13**(4): p. 485-492.

9. Yue, S.J., et al., *Enhanced Production of 2-Hydroxyphenazine from Glycerol by a Two-Stage Fermentation Strategy in Pseudomonas chlororaphis GP72AN.* J Agric Food Chem, 2019.

10. Wang, S., et al., *Enhanced biosynthesis of arbutin by engineering shikimate pathway in Pseudomonas chlororaphis P3.* Microb Cell Fact, 2018. **17**(1): p. 174.

11. King, E.O., M.K. Ward, and D.E. Raney, *Two simple media for the demonstration of pyocyanin and fluorescin.* J Lab Clin Med, 1954. **44**(2): p. 301-7.

12. van Rij, E.T., et al., *Influence of environmental conditions on the production of phenazine-1-carboxamide by Pseudomonas chlororaphis PCL1391.* Mol Plant Microbe Interact, 2004. **17**(5): p. 557-66.

13. O'Toole, G.A., *Microtiter dish biofilm formation assay.* J Vis Exp, 2011(47).

14. Haynes, W.C., et al., *PSEUDOMONAS AUREOFACIENS KLUYVER AND PHENAZINE α-CARBOXYLIC ACID, ITS CHARACTERISTIC PIGMENT.* Journal of Bacteriology, 1956. **72**(3): p. 412-417.

15. Haynes, W.C. and L.J. Rhodes, *Comparative taxonomy of crystallogenic strains of Pseudomonas aeruginosa and Pseudomon as chlororaphis.* J Bacteriol, 1962. **84**: p. 1080-4.

16. Gaby, W.L. and E. Free, *Occurrence and identification of nonpigmented strains of Pseudomonas aeruginosa in the clinical laboratory.* J Bacteriol, 1953. **65**(6): p. 746.

17. Chin, A.W.T.F., et al., *Root colonization by phenazine-1-carboxamide-producing bacterium Pseudomonas chlororaphis PCL1391 is essential for biocontrol of tomato foot and root rot.* Mol Plant Microbe Interact, 2000. **13**(12): p. 1340-5.

18. Zhang, L., et al., *Genome analysis of plant growth-promoting rhizobacterium Pseudomonas chlororaphis subsp. aurantiaca JD37 and insights from comparasion of genomics with three Pseudomonas strains.* Microbiol Res, 2020. **237**: p. 126483.

19. Burr, S.E., et al., *Pseudomonas chlororaphis subsp. piscium subsp. nov., isolated from freshwater fish.* Int J Syst Evol Microbiol, 2010. **60**(Pt 12): p. 2753-2757.

20. Peng, H., et al., *Enhanced biosynthesis of phenazine-1-carboxamide by engineered Pseudomonas chlororaphis HT66.* Microb Cell Fact, 2018. **17**(1): p. 117.

21. Liu, W.H., et al., *Characterization and Engineering of Pseudomonas chlororaphis LX24 with High Production of 2-Hydroxyphenazine.* J Agric Food Chem, 2021. **69**(16): p. 4778-4784.

22. Biessy, A., et al., *Inhibition of Three Potato Pathogens by Phenazine-Producing Pseudomonas spp. Is Associated with Multiple Biocontrol-Related Traits.* mSphere, 2021. **6**(3): p. e0042721.

23. Korth, H., *Einfluß von Eisen und Sauerstoff auf die Pigmentbildung bei verschiedenen Pseudomonas-Spezies.* Archiv für Mikrobiologie, 1971. **77**(1): p. 59-64.

24. Slininger, P.J. and M.A. Shea-Wilbur, *Liquid-culture pH, temperature, and carbon (not nitrogen) source regulate phenazine productivity of the take-all biocontrol agentPseudomonas fluorescens 2-79.* Applied Microbiology and Biotechnology, 1995. **43**(5): p. 794-800.

25. Bosire, E.M. and M.A. Rosenbaum, *Electrochemical Potential Influences Phenazine Production, Electron Transfer and Consequently Electric Current Generation by Pseudomonas aeruginosa.* Front Microbiol, 2017. **8**: p. 892.

26. Slininger, P.J. and M.A. Jackson, *Nutritional factors regulating growth and accumulation of phenazine 1-carboxylic acid by Pseudomonas fluorescens 2-79.* Applied Microbiology and Biotechnology, 1992. **37**(3): p. 388-392.

27. Cui, J.J., et al., *Enhanced Phenazine-1-Carboxamide Production in Pseudomonas chlororaphis H5ofleQorelA through Fermentation Optimization.* Fermentation-Basel, 2022. **8**(4).

28. Li, L., et al., *Metabolic Engineering of Pseudomonas chlororaphis Qlu-1 for the Enhanced Production of Phenazine-1-carboxamide.* Journal of Agricultural and Food Chemistry, 2020. **68**(50): p. 14832-14840.

29. Maddula, V.S., et al., *Quorum sensing and phenazines are involved in biofilm formation by Pseudomonas chlororaphis (aureofaciens) strain 30-84.* Microb Ecol, 2006. **52**(2): p. 289-301.

Chapter 3: Constructing a pangenome for potential phenazine-producers: Draft genome assemblies of 34 *Pseudomonas* isolates

Abstract

With the development and constant improvement of next-generation sequencing technologies, it is now economically feasible to sequence the entire genomes of collections of many strains. In addition, there are many sequencing technologies and bioinformatics algorithms available to choose from to sequence and assemble these genomes, each with their own strengths and weaknesses. Here I sequenced 34 *Pseudomonas* strains using 2 next-generation sequencing platforms: Illumina and Oxford Nanopore. I assembled these reads using 3 established bioinformatics algorithms (i.e., SPAdes, Unicycler, and Flye) for *de novo* bacterial genome assembly. The best genome assemblies were the hybrid Unicycler assemblies which used both reads sets as input due to their high contiguity and completeness metrics (L_{50} = 1 and BUSCO score >98% for all strains). I assembled the pangenome from these final hybrid genome assemblies and found that the pangenome contained 12674 coding sequences with 3471 core genes. By assembling this collection of 34 isolates, we doubled the potential number of targets for our future metabolic engineering efforts.

Introduction

Over the past 20 years the cost of DNA sequencing has decreased from hundreds of millions of dollars per genome to only hundreds of dollars thanks to developments in next-generation sequencing. According to the NIH National Human Genome Research Institute, as of August 2021 genome sequencing costs are $0.006 per mega base pair (Mbp) which translates to about $562 per human genome[1]. Further costs decreases are expected, as multiple companies announced in 2022 that they will be able to sequence a human genome for $100 in the near future[2]. As microbial genomes are much smaller than human genomes, the cost to sequence a bacterial genome may become even lower.

These technological advancements and associated cost decreases now make it feasible to sequence collections of tens to thousands of microbial genomes. Whole genome sequencing (WGS) data from fluorescent pseudomonads could provide insight into biocontrol mechanisms for plant growth promoting strains or virulence and antibiotic resistance among strains which are pathogenic to plants and humans. This study sequences and assembles the genomes of a small collection of *Pseudomonas chlororaphis* isolates to better understand their abilities to produce phenazine compounds. Creating genome assemblies for these isolates, many of which are minimally characterized, could allow them to be developed as biotechnology hosts for phenazines or other products.

Assembling bacterial genomes from millions of raw WGS reads requires a pipeline of bioinformatics tools. Each next-generation sequencing technology has its own advantages and disadvantages for *de novo* bacterial genome assembly which should be

considered when selecting an assembly algorithm. Illumina technologies tend to generate reads with high quality scores and low error rates at relatively low costs. These short reads do not span long repeats often found in bacterial genomes, therefore assemblers cannot resolve those genomic regions & GC rich regions, yielding highly fragmented assemblies[3]. Oxford Nanopore technologies generate much longer reads on the order of 10,000 bp long, which can yield complete assemblies with much lower read coverage but require the assembler to account for their much higher error rates[4]. Using reads and assemblers that can utilize both sequencing technologies could address the concerns of each sequencing technology, but this requires both reads to be purchased, increasing costs. Selecting technologies with higher accuracy long reads would also address these concerns, but these reads may be much more expensive than the aforementioned technologies[5]. Sequencing technologies and bioinformatics tools should be selected due to their inherent advantages and disadvantages.

Here I report draft genome sequences for 34 *Pseudomonas* isolates (Table 2.1) purchased from international culture collection. I sequenced all isolates and assembled their genomes using a range of well-known bacterial genome assembly algorithms (i.e., SPAdes, Unicycler, and Flye) using 2nd generation (i.e., Illumina), 3rd generation (i.e., Nanopore), or both next-generation sequencing technologies. I created multiple sets of assemblies to test which combinations yielded the best assemblies for my applications and assembled a pangenome using the final genome assemblies.

Materials and methods

DNA isolation and sequencing

Genomic DNA was isolated using the Quick-DNA™ Fungal/Bacterial Miniprep Kit (Zymo Research; Irvine, CA) and sent to the Microbial Genome Sequencing Center (Pittsburgh, PA) for whole genome sequencing. All isolates were sequenced on the NextSeq 2000 (Illumina; San Diego, CA) with paired-end 150 base pair reads as well as with Oxford Nanopore technologies with an estimated genome coverage of at least 45x for each read set. Illumina read quality was assessed using FASTQC v.0.11.9[6] and Nanopore read statistics were assessed with NanoStats v.1.28.2[7] on the Galaxy servers before and after read filtering and trimming. Assembly summary statistics (i.e., total assembly length, number of contigs, N_{50}, L_{50}, % GC) were calculated using QUAST v.5.0.2[8]. Genome completeness was assessed by running BUSCO v.5.2.2 in genome mode using the pseudomonadales_odb10 (prokaryota, 2020-03-06) database[9]. Either the percentage of Complete and Single-copy BUSCOS (S) or Complete BUSCOS (C) present out of the 782 expected in the order Pseudomonadales was reported here as an assembly completeness metric.

Genome assembly

Flye genome assemblies were created with Flye v.2.8.3 and raw Nanopore reads as input (--nano-raw)[10]. All other genome assemblies used reads which were filtered and trimmed based on read quality or size. Raw Illumina reads were trimmed to remove adapters and low-quality ends using Trimmomatic v.0.38 (ILLUMINACLIP:NexteraPE :2:30:10:8, LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:15 MINLEN:36)[11].

Raw Nanopore reads were trimmed using the default setting of Porechop v.0.2.4 [https://github.com/rrwick/Porechop] then filtered with filtlong v.0.2.1 to remove reads shorter than 1000 bp and with average quality scores below 10 (--min_length 1000, --min_mean_q 10) [https://github.com/rrwick/Filtlong]. SPAdes genome assemblies were created using SPAdes v.3.12.0[12] on Galaxy's public servers at usegalaxy.org[13] with contigs shorter than 1000 bp later filtered out; Unicycler assemblies were created using Unicycler v.0.4.8 using both "normal" and "bold" bridging modes (--mode normal; --mode bold) and excluding contigs shorter than 1000 bp from the assemblies[14]. The short reads assemblies were created using only the paired end Trimmomatic output (i.e., the quality-trimmed Illumina reads). The long reads Nanopore assemblies were created using the output from filtlong (i.e., the trimmed then filtered Nanopore reads). The hybrid assemblies were assembled using both sets of aforementioned reads.

Genome annotation and pangenome construction

Genome assemblies were annotated with Prokka v1.14.6, using a minimum contig size of 1000 bp and 'Pseudomonas' as the genus name[15]. The pangenome was constructed using PEPPAN v.1.0.5 and the .gff files generated by Prokka as input [16]. A pangenome rarefaction curve, gene presence absence matrix, and accessory genome tree were created from the PEPPAN output excluding pseudogenes using the included PEPPAN_parser algorithm (--pseudogene). Statistics about core genome were calculated from the gene presence absence matrix using R v.4.2.1[17]. The resulting .nwk tree file was visualized using R v.4.2.1[17] and treeio package v.1.20.2[18]. All remaining figures were created using GraphPad Prism v.9.4.1 (GraphPad Software; San Diego, CA).

Results

I assembled all genomes using three established bacterial *de novo* genome assembly algorithms, SPAdes, Unicycler, and Flye, which specialize in short reads, hybrid, and long reads assemblies, respectively. The selected algorithms which utilize short reads function by creating simplified de Bruijn graphs from short sequences called k-mers due to the short length and high sequence accuracy of the Illumina reads. The hybrid approaches then use the long reads to close the gaps between contigs and resolve repeats in these assembly graphs.

SPAdes can create both short reads and hybrid assemblies, each run generating output in the form of contigs and scaffolds. Contigs are the sequence fragments which comprise a genome in incomplete assemblies, and scaffolds are made by joining likely-adjacent contigs using stretches of Ns as connectors. I included both the contigs and scaffolds when evaluating the assemblies. Unicycler also creates short and hybrid assemblies with 3 bridging modes available which merge contigs together based off different bridge quality thresholds. I selected the “normal” intermediate mode and the “bold” mode which more aggressively merges contigs with the goal of achieving complete assemblies at a higher risk of misassembly.

Due the much longer length and high intrinsic error rates of the long reads, the long read assemblers use different strategies than the short reads assemblers. Unicycler assembles its long reads assemblies with miniasm[19] which maps overlaps directly between reads then polishes the sequence with Racon to improve its accuracy. Flye creates repeat graphs using approximate matches among long uncorrected reads.

I calculated the genome summary statistics for all genome assemblies and compared them to evaluate which assemblers and read sets produced the best genome assemblies (Figure 3.1). The length of most *P. chlororaphis* assemblies ranged from 6.5-7.2 Mbp and had an average GC content of ~63% which seem reasonable compared to other previously published genomes. Strain DSM 6508 consistently had much larger assemblies than the other strains, around 7.65-7.94 Mbp for the short and hybrid assemblies. The long reads Unicycler assemblies had slightly lower GC content and a larger spread of total assembly lengths, with an average size of 7.7 Mbp and standard deviation of 0.96. For the *P. synxantha* strain ATCC 17413, all assemblers produced a smaller genome around 6.1 Mbp and approximately 60% GC content.

I compared the number of contigs, N_{50}, and L_{50} to assess assembly contiguity, which provide insight into the location and orientation of genomic features (Figure 3.1). The number of contigs describes the number of sequence fragments in an assembly, which should ideally approach one to accurately represent bacterial genomes with a singular circular chromosome. The N_{50} and L_{50} describe the lengths and number of the smallest number of contigs, respectively, which contain 50% of the genome assembly length. Thus, L_{50} should ideally approach 1 and N_{50} should approach the total assembly size in a contiguous assembly. All short reads assemblies were very fragmented, comprised of at least 25 contigs. The Flye assemblies and hybrid Unicycler assemblies had the best number of contigs, N_{50}, and L_{50} values. Overall, the hybrid SPAdes and long reads Unicycler assemblies had intermediate contiguity statistics with some assemblies more comparable to the short reads and others more comparable the Flye assemblies.

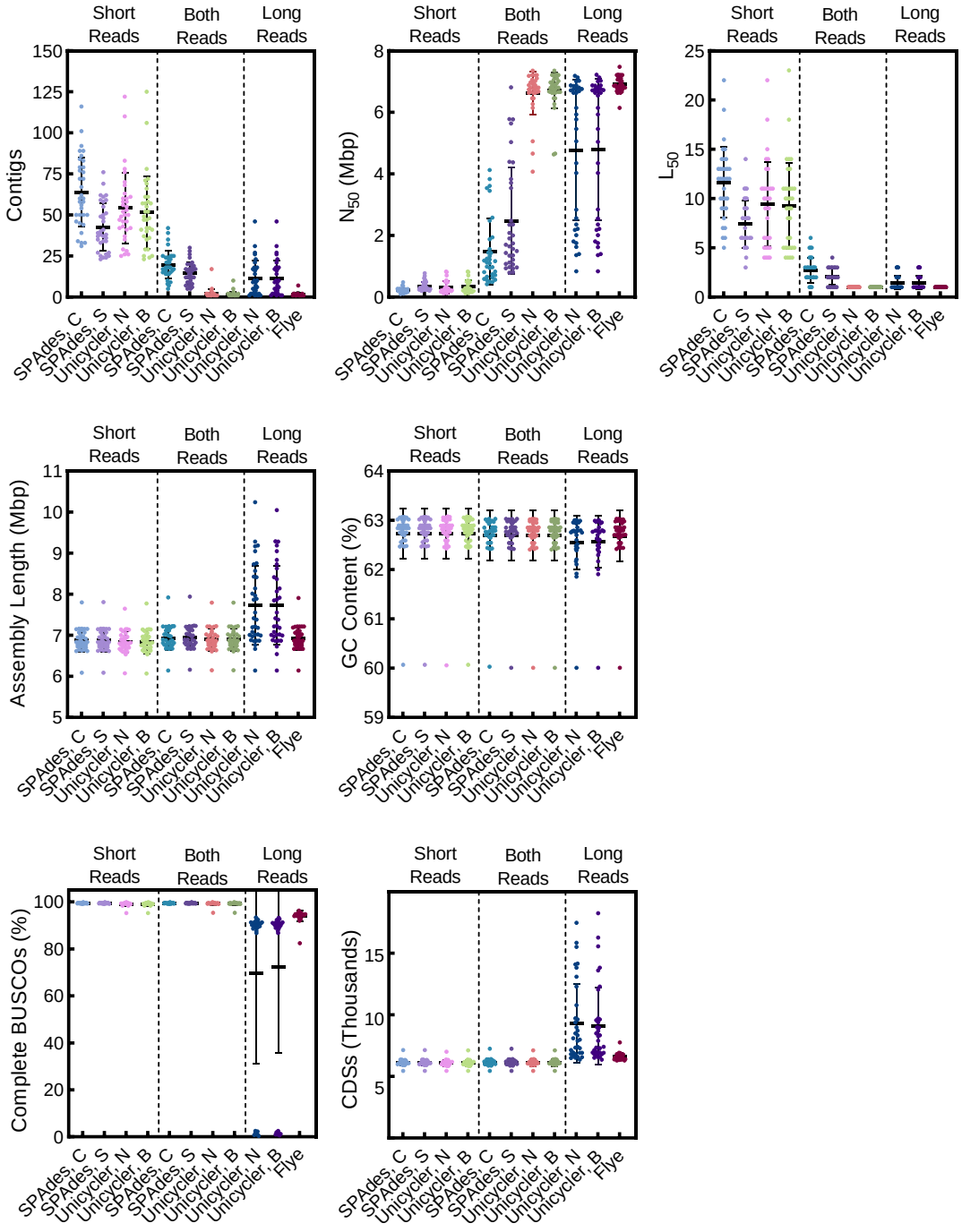

Figure 3.1. Graphical comparisons of the summary statistics for all genome assemblies. Genome assemblies were created with only short Illumina sequencing reads (left graph subsections), only long Oxford Nanopore sequencing reads (right graph subsection) or as hybrid assemblies with both reads sets (middle graph subsection), with contigs files (C) and scaffolds (S) generated by SPAdes, and genome assemblies generated by using either normal (N) or bold (B) bridging mode. Number of contigs, N_{50}, L_{50}, total sequence length, and GC content were generated using QUAST. The % complete BUSCOS was calculated using BUSCO, and number of CDSs from assembly annotations generated by Prokka.

To assess genome completeness, I implemented BUSCO and compared the percentage of complete BUSCOs (Benchmarking Universal Single-Copy Orthologs) present in each assembly (Figure 3.1) to predict how many common genes are present. Almost all the short reads assemblies have scores of at least 99%, indicating mostly complete genomes. Much fewer complete BUSCOs were present in long reads only assemblies, with median values of 94.3% and 90% for the Flye and Unicycler assemblies. Some Unicycler assemblies had only 0-3% BUSCOs present. Interestingly, in most of those cases either the normal or bold bridging mode would have this very low BUSCO score, but not both.

I also annotated all genomes with Prokka and compared the number of CDSs found in each assembly. The hybrid and short reads assemblies have an average of about 6,100 CDSs with most assemblies ranging from 5455-6422 CDSs and DSM 6508 ranging from 7022-7264 CDSs. The long reads assemblies have more CDSs than the others, where the Flye and Unicycler assemblies have averages of approximately 6,607 CDSs and 9,200 CDSs, respectively. The larger number of CDSs but lower percentage of complete BUSCOs in the long reads assemblies could indicate misassemblies which led to misannotations. The hybrid Unicycler assemblies were selected as the best assemblies due to their high contiguity and completeness metrics. While the Flye assemblies had slightly better contig numbers and N_{50} values, they had fewer complete BUSCOs present. For the final assemblies (Table 3.1), I selected a combination of hybrid Unicycler assemblies created with bold and normal bridging modes, preferencing the normal mode when summary statistics were not much better to minimize risk of misassembly.

Table 3.1. Summary statistics for final genome assemblies. The number of CDSs was tallied from the Prokka genome annotations, and the complete and single copy BUSCOS was calculated using BUSCO. All other statistics were generated from QUAST.

Strain	Total length (bp)	Number of contigs	N_{50}	L_{50}	GC (%)	Number of CDSs	Complete and single copy BUSCOS (%)
ATCC 13985	7 024 010	10	4 636 000	1	62.7	6251	99.1
ATCC 13986[1]	6 675 284	2	6 636 555	1	63.0	5926	99.5
ATCC 13986[2]	6 682 756	2	6 644 045	1	63.0	5949	99.5
ATCC 15926	6 763 921	1	6 763 921	1	62.9	5998	99.4
ATCC 17411	7 212 419	1	7 212 419	1	62.5	6366	99.2
ATCC 17413	6 147 644	1	6 147 644	1	60.0	5459	99.9
ATCC 17414	6 807 169	1	6 807 169	1	63.0	6048	99.5
ATCC 17415[1]	6 664 157	1	6 664 157	1	63.0	5887	99.4
ATCC 17415[2]	6 664 503	1	6 664 503	1	63.0	5884	99.2
ATCC 17417	6 746 536	1	6 746 536	1	62.9	5954	99.1
ATCC 17418[1]	6 883 267	1	6 883 267	1	62.8	6075	99.5
ATCC 17418[2]	6 881 643	1	6 881 643	1	62.8	6074	99.5
ATCC 17419	6 608 598	5	4 662 896	1	62.7	5919	99.0
ATCC 17809	7 020 903	1	7 020 903	1	62.4	6223	99.0
ATCC 17810	6 863 056	2	6 791 445	1	62.7	6074	99.4
ATCC 17811	7 189 114	1	7 189 114	1	62.4	6422	99.4
ATCC 17814	6 807 913	1	6 807 913	1	63.0	6050	99.4
ATCC 33663[1]	7 109 352	1	7 109 352	1	62.9	6281	99.1
ATCC 33663[2]	7 108 820	1	7 108 820	1	62.9	6284	99.2
ATCC 9446	6 637 791	1	6 637 791	1	63.0	5909	99.2
ATCC 9447	6 807 068	3	6 677 872	1	63.0	6048	99.4
DSM 21509	7 064 975	1	7 064 975	1	62.7	6246	99.1
DSM 29578[1]	7 216 947	1	7 216 947	1	62.5	6378	99.1
DSM 29578[2]	7 216 571	1	7 216 571	1	62.5	6380	99.1
DSM 6508	7 915 166	3	7 476 725	1	62.5	7222	99.4
NCCB 100368[1]	6 870 522	2	6 455 838	1	62.8	6010	99.2
NCCB 100368[2]	6 870 415	2	6 455 628	1	62.8	6010	99.1
NCCB 47033	7 221 530	1	7 221 530	1	62.4	6378	99.2
NCCB 60037	6 977 278	1	6 977 278	1	62.7	6209	99.1
NCCB 60038	6 979 353	1	6 979 353	1	62.7	6214	99.1
NCCB 82053[1]	6 763 242	2	6 274 577	1	62.9	6002	99.4
NCCB 82053[2]	6 762 156	1	6 762 156	1	62.9	6002	99.4
NCCB 88062[1]	7 025 460	2	6 660 177	1	62.8	6233	99.1
NCCB 88062[2]	6 923 225	1	6 923 225	1	62.8	6138	98.2

Even amongst members of the same species, there was quite the variation in genome summary statistics. Ignoring DSM 6508, the total assembly length range varies by over 500,000 bp for the *P. chlororaphis* strains which is a considerable portion of the genome. To get a better idea of how gene content varies among these strains, I also assembled the pangenome for all strains (Figure 3.2) from the final genome assemblies and their Prokka annotations listed in Table 3.1.

All CDSs present in a collection of strains make up the pangenome, and the CDSs which are present in every strain make up the core genome. My collection of all 34 strains has a core genome of 3471 CDSs and a pangenome of 12674 CDSs. This implies that 27.3% of the pangenome is comprised of core genes, and the remaining 9203 CDSs make up the accessory genome. CDSs which are found in only 4 or fewer strains, classifying them as cloud genes, comprise over half the accessory genome and 41.3% of the entire pangenome with 5,233 CDSs. The pangenome has 2914 singletons, or CDSs unique to a single isolate, most of which are from *P. synxantha* strain ATCC 17413 (1057 CDSs) and DSM 6508 (1134 CDSs).

As each genome has between 5459 and 7222 CDSs, core genes make up 48.1-63.6% of each genome. The 33 *P. chlororaphis* isolates alone have a core genome of 4396 CDS which comprise 60.1-74.7% of their genomes. When comparing which genes are present and absent among *P. chlororaphis* isolates, isolates cluster into groups based on whether they contain *phzO* or *phzH*. All isolates which originated from the same culture collection strain also clustered together and do not have any singletons, but they may have genetic variation in the form of single nucleotide polymorphisms (SNPs).

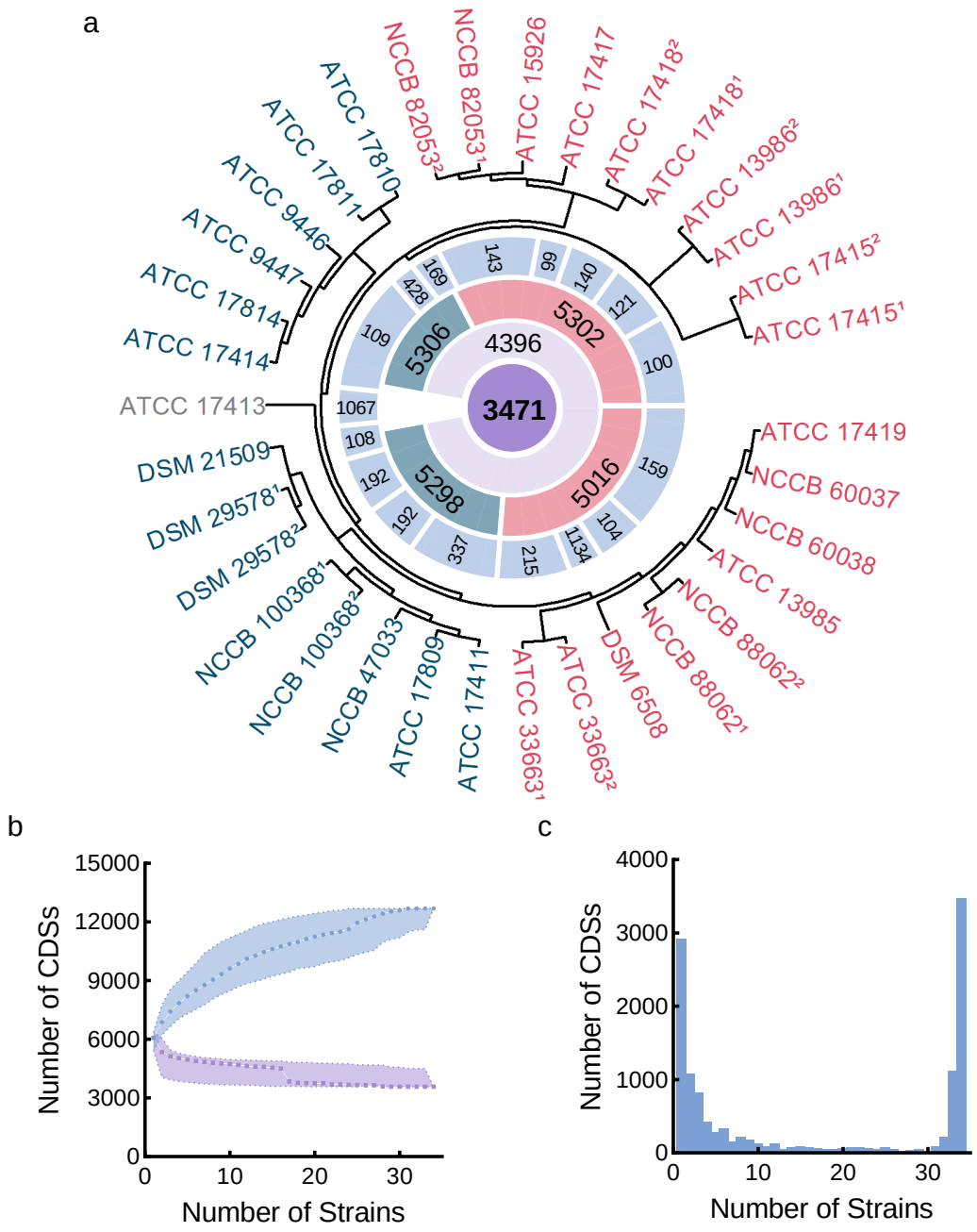

Figure 3.2. Summary of the pangenome constructed from the final hybrid genome assemblies. a) A tree which shows the similarity in gene content among all genomes. The 3 inner circles (from inside to outside) display the size of the core genomes for all strains (purple), only the *P. chlororaphis* strains (light purple), and the major branches which correspond to the presence of *phzO* (peach) or *phzH* (teal). The outer circle (blue) shows cloud genes, which are only present the respective group of isolates. b) The pangenome rarefaction curve shows how the total pangenome size (blue) increases and the core genome size (purple) decreases as isolates are added to the pangenome. c) Histogram showing how many CDSs are only found in a specific number of strains. Core genes present in almost all strains or cloud genes only found in a few are present with the highest frequencies.

Discussion

In this work, I assembled nearly complete genomes for 34 *Pseudomonas* isolates. As of September 17th, 2022, the Pseudomonas Genome Database has 80 assemblies of *P. chlororaphis* available with 47 marked as complete, and NCBI Genbank has 111 assemblies of *P. chlororaphis* available with 58 marked as complete. A few of these isolates have been sequenced previously[20] and deposited into these databases, but most members of this strain collection have never been previously sequenced.

I sequenced all strains with both Illumina and Oxford Nanopore technologies since each technology generates reads which vary in length, accuracy, and cost, which affect the quality of the genome assemblies and the economic feasibility of sequencing a small collection of strains. For this study, the Illumina and Nanopore sequencing services had cost $125 and $350, respectively, for a total of $475 per strain. These cost differences are small for few strains but add up for larger collections of strains. These sets of genome assemblies were assessed by comparing their genome summary statistics, which provide insight into the contiguity and completeness of each assembly.

The short reads only assemblies appear to be generally complete and correct, with genome size and number of CDSs comparable to other available complete assemblies and with >99% complete BUSCOS present, but with low contiguity statistics. For applications which do not require contiguous sequences, short reads assemblies could be the best option, as sequencing with only Illumina reads costs almost 4 times less than using both technologies in this study.

Depending on the assembly algorithm and specific sequencing technologies selected, long reads assemblers create highly contiguous assemblies with varying accuracy. While the Flye assemblies had the best contiguity metrics among all assemblies and seemingly correct genome size, Unicycler produced assemblies varying from reasonable sizes to larger than expected by millions of base pairs. The assemblies with large genome sizes and numbers of CDSs appear to be misassemblies, as the largest complete *P. chlororaphis* genome assembly available on NCBI has a length of 7.31 Mbp with 6,446 annotated CDSs [NCBI Accession: GCA_017922975.1]. Possibly the wide range in Unicycler assembly quality could be a result of the input data, either issues intrinsic to the genome itself (e.g., repeats, other difficult to resolve regions) or read sets (e.g., coverage, read length distribution).

There are new opportunities to produce better long reads assemblies as 3rd generation sequencing technologies improve for greater accuracy and accessibility. New portable sequencing devices have been developed which allow individual lab groups to affordably create long sequencing read sets in real time. As error rates decrease for long reads sequencing technologies, error correction steps with short reads are no longer necessary[21]. As each long reads assembler has different strengths and weaknesses, algorithms which combine multiple long reads assemblies could improve accuracy of assemblies created with only long reads[22-24].

Our best and final assemblies were created using both Illumina and Oxford Nanopore reads, which was the most expensive of the options. Other studies also had

success creating genome assemblies and pangenomes using Unicycler with Illumina and Nanopore reads in *P. aeuruginosa* and other bacteria[25, 26].

One strain, DSM 6508, consistently had much a much larger total assembly length than the other strains. With a final assembly length of 7.9 Mbp and largest contig length of 7.48 Mbp, DSM 6508 is larger than the largest reported complete *P. chlororaphis* genome assembly. This strain was isolated from soil in Japan and identified and deposited into DSMZ [original strain designation: 305-STR-1-4] as a styrene degradation strain[27, 28]. Possibly the 1 larger contig could be the genomic chromosome while the other 2 smaller contigs could comprise large plasmid(s) for aromatics degradation. The 1134 CDSs which are unique to this strain could be involved in the utilization of styrene, benzene, or toluene as a carbon source.

The pangenome assembled in this study has a similar structure to the largest pangenome study of phenazine production in fluorescent pseudomonads. For their study of 63 fluorescent pseudomonads, Biessy et al. calculated a core genome size of 3143 CDSs, and their 38 *P. chlororaphis* isolates had a core genome of 4469 CDSs[20]. Their strains also clustered into groups based on presence of *phzO* or *phzH*, which also corresponded to common subspecies designations for *P. chlororaphis,* which have ~5,450 common CDSs for each of these groups. There have been some other comparative genomics and pangenomics studies of phenazine-producing pseudomonads as well, but fewer *P. chlororaphis* isolates are included[29-32].

As expected, all sequenced strains had a single copy of the phenazine biosynthesis operon, and all strains of *P. chlororaphis* had a copy of either *phzO* or *phzH* immediately

downstream of *phzABCDEFG*. Thus, the inability for some strains to produce various phenazines is not due to missing *phz* genes, although there still could be loss of function mutations. Many of the isolates which originated from the same culture collection had different phenotypes recorded in Chapter 2. Since many of these isolates do not have any CDSs unique to them, the presence or absence of a particular gene is not responsible for their differing phenotypes. Other genetic variation, such as SNPs or differences in noncoding regions, or non-genetic differences in gene expression could be responsible for these phenotype differences. It is unsurprising that most of the isolates originating from the same culture collection strain have few to no singletons, as they may be closely related clonal isolates.

This study provides nearly complete genome assemblies for 34 Pseudomonas isolates, many of which have never been previously sequenced. The assembled pangenome has 12674 CDSs from which metabolic engineering targets can be selected for a population genomics-guided metabolic engineering approach in the next chapter. Selecting targets from a collection of isolates doubles the amount of potential targets versus a tradition genome-wide approach which can select targets from the ~6000 CDSs present in a single genome.

References

1. KA, W. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. 9/25/2022]; Available from: www.genome.gov/sequencingcostsdata.

2. Pennisi, E., *A $100 genome? New DNA sequencers could be a 'game changer' for biology, medicine.* Science, 2022.

3. Pop, M. and S.L. Salzberg, *Bioinformatics challenges of new sequencing technology.* Trends Genet, 2008. **24**(3): p. 142-9.

4. Delahaye, C. and J. Nicolas, *Sequencing DNA with nanopores: Troubles and biases.* PLoS One, 2021. **16**(10): p. e0257521.

5. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC Genomics, 2012. **13**: p. 341.

6. Andrews, S., *FastQC: A Quality Control tool for High Throughput Sequence Data.*

7. De Coster, W., et al., *NanoPack: visualizing and processing long-read sequencing data.* Bioinformatics, 2018. **34**(15): p. 2666-2669.

8. Mikheenko, A., et al., *Versatile genome assembly evaluation with QUAST-LG.* Bioinformatics, 2018. **34**(13): p. i142-i150.

9. Manni, M., et al., *BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.* Mol Biol Evol, 2021. **38**(10): p. 4647-4654.

10. Kolmogorov, M., et al., *Assembly of long, error-prone reads using repeat graphs.* Nat Biotechnol, 2019. **37**(5): p. 540-546.

11. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

12. Prjibelski, A., et al., *Using SPAdes De Novo Assembler.* Curr Protoc Bioinformatics, 2020. **70**(1): p. e102.

13. Afgan, E., et al., *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update.* Nucleic Acids Res, 2018. **46**(W1): p. W537-W544.

14. Wick, R.R., et al., *Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads.* PLoS Comput Biol, 2017. **13**(6): p. e1005595.

15. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-2069.

16. Zhou, Z., J. Charlesworth, and M. Achtman, *Accurate reconstruction of bacterial pan- and core genomes with PEPPAN.* Genome Res, 2020. **30**(11): p. 1667-1679.

17. RCoreTeam, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria., 2022.

18. Wang, L.G., et al., *Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data.* Mol Biol Evol, 2020. **37**(2): p. 599-603.

19. Li, H., *Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.* Bioinformatics, 2016. **32**(14): p. 2103-10.

20. Biessy, A., et al., *Diversity of phytobeneficial traits revealed by whole-genome analysis of worldwide-isolated phenazine-producing Pseudomonas spp.* Environ Microbiol, 2019. **21**(1): p. 437-455.

21. Sereika, M., et al., *Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing.* Nat Methods, 2022. **19**(7): p. 823-826.

22. Wick, R.R., et al., *Trycycler: consensus long-read assemblies for bacterial genomes.* Genome Biol, 2021. **22**(1): p. 266.

23. Boostrom, I., et al., *Comparing Long-Read Assemblers to Explore the Potential of a Sustainable Low-Cost, Low-Infrastructure Approach to Sequence Antimicrobial Resistant Bacteria With Oxford Nanopore Sequencing.* Front Microbiol, 2022. **13**: p. 796465.

24. Wick, R.R. and K.E. Holt, *Benchmarking of long-read assemblers for prokaryote whole genome sequencing.* F1000Res, 2019. **8**: p. 2138.

25. Molina-Mora, J.A., et al., *High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers.* Scientific Reports, 2020. **10**(1).

26. Chen, Z., D.L. Erickson, and J. Meng, *Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing.* BMC Genomics, 2020. **21**(1): p. 631.

27. Shirai, K. and K.-i. Hisatsuka, *Isolation and Identification of Styrene Assimilating Bactiria.* Agricultural and Biological Chemistry, 2014. **43**(7): p. 1595-1596.

28. Lang, E., *Diversity of bacterial capabilities in utilizing alkylated benzenes and other aromatic compounds.* Lett Appl Microbiol, 1996. **23**(4): p. 257-60.

29. Garrido-Sanz, D., et al., *Genomic and Genetic Diversity within the Pseudomonas fluorescens Complex.* Plos One, 2016. **11**(2).

30. Loper, J.E., et al., *Comparative Genomics of Plant-Associated Pseudomonas spp.: Insights into Diversity and Inheritance of Traits Involved in Multitrophic Interactions.* Plos Genetics, 2012. **8**(7).

31. Gross, H. and J.E. Loper, *Genomics of secondary metabolite production by Pseudomonas spp.* Nat Prod Rep, 2009. **26**(11): p. 1408-46.

32. Chen, Y.W., et al., *Comparative genomic analysis and phenazine production of Pseudomonas chlororaphis, a plant growth-promoting rhizobacterium.* Genomics Data, 2015. **4**: p. 33-42.

Chapter 4: Population genomics-guided metabolic engineering of phenazine biosynthesis in *Pseudomonas chlororaphis*

Abstract

With the development of next generation sequencing and new bioinformatics tools, large genomic datasets can be now generated, analyzed, and used for metabolic engineering studies. This study aims to employ a population genomics-guided metabolic engineering approach which identifies engineering targets from a small collection of whole genome sequencing data (WGS) to improve phenazine production in *P. chlororaphis*. We provided WGS and phenazine production data from 34 strains to a microbial genome-wide association studies (GWAS) tool called DBGWAS, which identified 330 genomic loci which were significant for phenazine production. I overexpressed the 7 coding sequences (CDS) associated with the top phenotype scores in PCN-production strain DSM 21509. 3 of these CDSs significantly affected phenazine production: ProY_1, a histidine transporter, and hypothetical protein PS__04251 which increased PCN production by 77 mg/L and 57 mg/L, respectively, and hypothetical protein and putative transcriptional regulator PS__04252 which decreased PCN by 40 mg/L. This study successfully identified some non-traditional metabolic engineering targets and demonstrated the potential of this metabolic engineering approach

Introduction

The development of next-generation sequencing and CRISPR genome editing has enabled entire microbial genomes to be sequenced and manipulated, resulting in genome-wide metabolic engineering approaches often within non-traditional hosts. With further advancements in DNA sequencing technologies, it is now economically feasible for a single research group to sequence small collections of tens to hundreds of microbial isolates, sometimes even in-house with portable real-time sequencing devices. New metabolic engineering strategies could take advantage of this increasing accessibility of microbial whole-genome sequencing data by identifying metabolic engineering targets from a collection of genomes in a "pangenome"-wide or population genomic approach.

This work seeks to identify non-intuitive metabolic engineering targets within a collection of *Pseudomonas chlororaphis* isolates to improve phenazine production as part of a population genomics approach to metabolic engineering. Phenazines are redox-active, often colorful, secondary metabolites with applications in agriculture as antifungal agents and potential applications as redox meditators in flow cell batteries and bioelectrochemical devices[1-3]. *P. chlororaphis* is a commercially available biocontrol strain which would make a good potential phenazine production host, as it is non-pathogenic to humans and plants, can utilize the inexpensive carbon source glycerol, and natively produces multiple phenazine derivatives. Even so, many strains have traits which are detrimental to industrial bioprocessing, such as biofilm formation and low growth temperatures. We sequenced the genomes of 34 *Pseudomonas* isolates and characterized

their bioprocess-relevant phenotypes (i.e., phenazine production, biofilm formation, and growth temperature) to select the optimal host strain for phenazine production.

Pseudomonas chlororaphis has already been successfully engineered for phenazine production, with metabolic engineering works often pursuing rational design strategies and targets. Replacing genes within the phenazine biosynthesis operon can modulate final phenazine composition and allow non-native phenazines to be produced, including 1-hydroxyphenazine[4] and phenazine-1,6,-dicarboxylic acid derivatives iodinin and 1,6-dimethoxyphenazine[5]. Regulation of the phenazine biosynthesis operon provides many opportunities to improve phenazine production, including *phzR* and *phzI* which directly regulate expression through quorum sensing[6] and components of the Gac/Rsm pathway (e.g., *rpeA*, *rsmE*, *lon* protease, *psrA*, *parS, gacA*[4, 7, 8]) which indirectly interact with the PhzR/PhzI in response to other environmental factors. Increasing carbon flux through the shikimate pathway, such as by overexpressing *aroB*, *aroD*, *aroE, ppsA, and tktA*[7, 8], also improves phenazine production by increasing flux through phenazine biosynthesis.

Our population genomics strategy uses microbial genome-wide association studies (mGWAS) to identify metabolic engineering targets from our genomic and phenotypic data. GWAS correlate genomic and phenotypic datasets to identify causal genetic variants[9]. While GWAS are traditionally used to identify human disease risk factors, recent bioinformatics tools have been developed to adapt these studies to bacteria[10-12]. This approach requires no prior knowledge of relevant biosynthetic pathways and could identify previously unknown targets for metabolic engineering

throughout the genome. We further sought to experimentally validate the top mGWAS hits by overexpressing or deleting associated genes and measuring phenazine production with respect to the wildtype control.

Materials and methods

mGWAS analysis

We performed mGWAS analysis using DBGWAS v. 0.5.4[12] for all 34 isolates listed in Table 2.1. The final hybrid genome assemblies listed in Table 3.1 were used as the input genotypic data. The phenazine production data from Figure 2.3 was manipulated to create 7 different phenotypes that were input as phenotype data:

(i) PCA production in KMB

(ii) PCA production in KMB+Fe

(iii) Effect of Fe on PCA production

(iv) PCN production in KMB

(v) PCN production in KMB+Fe

(vi) Effect of Fe on PCN production

(vii) Total phenazine production in KMB.

For phenotypes (i), (ii), (iv) and (v), concentrations of PCA and PCN (mg/L) obtained from experiments were used directly. Values of phenotypes (iii) and (vi) were estimated by subtracting the concentration of the phenazine compound in KMB from that in KMB+Fe. If the difference was negative, it was replaced by 0. Total phenazine production in KMB was obtained by simply adding together the concentrations of all phenazine compounds (i.e., PCA, 2-OHPCA, PCN and 2-HP) that were produced in

KMB for each isolate. The genome sequences of strains were also provided as a BLAST database to DBGWAS for genome mapping of significant unitigs, which were identified based on a q-value cutoff, and a minor allele frequency greater than 1% (default).

Even though DBGWAS maps significant unitigs to genomes by BLAST, we chose to independently perform unitig alignment to genomes by exact matching to avoid any tolerance to mismatches during alignment. Custom Python3 scripts were used for this purpose with the 34 genome sequences as the mapping database. To ensure that each unitig finds a match, both the significant unitig and its reverse-complement were used. Further, genome annotations were used to determine the genomic regions of the mapped unitigs (i.e., whether the unitig falls within a gene or an intergenic region).

The lists of GWAS hits from the 7 phenotypes were concatenated into a single list called the 'preliminary list'. In this list, each occurrence of a significant unitig constituted a single entry, creating separate entries for each phenotype and strain. Custom Python3 scripts were used to remove redundancies and collapse the preliminary list so that each significant unitig has a single entry in the final list. For each strain, entries for identical unitigs that were significant for multiple phenotypes were collapsed together in the 'phenotype-collapsed' list. Each entry was assigned a phenotype score, which represents the number of phenotypes (out of 7) where each unitig was significant. If a unitig had a phenotype score greater than 1, its q-value was taken to be the minimum of q-values for all phenotypes it is found in. Similarly, the effect of that unitig was taken to be the one with the highest magnitude across all phenotypes. The 'phenotype+strain-collapsed list' combined entries where the same unitig mapped to the same genomic region in multiple

strains. Redundancies where the reverse complement of a significant unitig shows up as a separate entry were then collapsed to create the 'final list' (also called as the 'phenotype+strain+rev comp'-collapsed list) of GWAS hits.

Genes associated to GWAS hits were determined based on the overlap of unitigs to genes. If the overlap to a gene was partial or complete, that gene was considered associated to the unitig. In case of no overlap, i.e., when the unitig appeared completely between two genes, both the neighboring genes were considered linked to the unitig.

The hits from the final list with a phenotype score of 3 or higher which were significant for PCN production phenotypes were selected as top hits for experimental validation. The CDS immediately upstream and downstream of each significant unitig were chosen as metabolic engineering targets to be overexpressed in the top PCN-producing strain. Any CDS which encoded ribosomal RNA was discarded from the list. If the significant unitig sequence was completely contained within a CDS, only the unitig-containing CDS was studied rather than the 2 adjacent CDSs.

Cloning and genetic manipulation

Each target was PCR-amplified from the genomic DNA of the strain listed on the top hits file then inserted into the backbone of plasmid pBb(RK2)1k-GFPuv using either restriction digest cloning or NEBuilder® HiFi DNA Assembly (New England Biolabs; Ipswich, MA) and using the primers listed in Table 4.1. pBb(RK2)1k-GFPuv is a broad-range expression vector with an IPTG-inducible promoter which was gifted by Brian Pfleger at the University of Wisconsin, Madison, and used as the empty vector control[13]. All plasmids used in this study are listed in Table 4.2.

Table 4.1. List of primers used in this study.

Name	Sequence (5'→3')	Description
PS__04252-pBb(RK2)1k_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGGTCAAACGCACAAGC	Amplify PS__04252 from DSM 21509 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
PS__04252-pBb(RK2)1k_R	CCTTACTCGAGTTTGGATCCCTATTGCACCGGCACCC	Amplify PS__04252 from DSM 21509 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
PS__04251-pBb(RK2)1k_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGACCGTGGCTCAAAGC	Amplify PS__04251 from DSM 21509 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
PS__04251-pBb(RK2)1k_R	ATCCTTACTCGAGTTTGGATCCTCAGCGCAGGATGCCGA	Amplify PS__04251 from DSM 21509 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
RhtA-pBb(RK2)1k_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGAATGACCAGCCCCG	Amplify *rhtA* from DSM 29578[2] gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
RhtA-pBb(RK2)1k_R	ATCCTTACTCGAGTTTGGATCCTCAATCAGCTGCAACCAAAG	Amplify *rhtA* from DSM 29578[2] gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
ProY1-pBb(RK2)1k_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGCAACAGCAAGCTCAA	Amplify *proY_1* from ATCC 9447 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
ProY1-pBb(RK2)1k_R	ATCCTTACTCGAGTTTGGATCCTTATCGATGGGACAAAGAAGG	Amplify *proY_1* from ATCC 9447 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
UctC-pBb(RK2)1k_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGGGCGCGTTATCTCAT	Amplify *uctC* from DSM 29578[2] gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
UctC-pBb(RK2)1k_R	ATCCTTACTCGAGTTTGGATCCTCACAGCACGCCCGAG	Amplify *uctC* from DSM 29578[2] gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
HutH2-pBb(RK2)1k_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATGTGACTGCGCTAAATCTG	Amplify *hutH2* from ATCC 9447 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
HutH2-pBb(RK2)1k_R	ATCCTTACTCGAGTTTGGATCCTTACAGGCTCGGCAGC	Amplify *hutH2* from ATCC 9447 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
pBb(RK2)1k_F	GGATCCAAACTCGAGTAAG	Amplify pBb(RK2)1k-GFPuv backbone for HiFi assembly, forward primer
pBb(RK2)1k_R	ATGTATATCTCCTTCTTAAAAGATCT	Amplify pBb(RK2)1k-GFPuv backbone for HiFi assembly, reverse primer
YbhH-HiFi_F	TTTAAGAAGGAGATATACATATGTCTTTTGAACTGGACCTTCCC	Amplify *ybhH* from DSM 21509 gDNA for HiFi assembly, forward primer
YbhH-HiFi_R	CCTTACTCGAGTTTGGATCCTTAGCCCCGCCCTTTCAAC	Amplify *ybhH* from DSM 21509 gDNA for HiFi assembly, reverse primer
Seq-pBb(Rk2)1k_F	CAATTAATCATCCGGCTCG	Forward sequencing primer for overexpression plasmids
Seq-pBb(Rk2)1k_R	GACTCTAGTAGAGAGCGTTC	Reverse sequencing primer for overexpression plasmids

Table 4.2. List of plasmids used in this study.

Plasmid name	Description	Source
pBb(RK2)1k-GFPuv	IPTG-inducible trc promoter expressing *gfpuv*, kanamycin resistance, RK2 origin of replication	Cook et al. 2018 [13]
pBb(RK2)1k-PS__04252	IPTG-inducible trc promoter expressing PS__04252 amplified from DSM 21509, kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-PS__04251	IPTG-inducible trc promoter expressing PS__04251 amplified from DSM 21509, kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-YbhH	IPTG-inducible trc promoter expressing *ybhH* amplified from DSM 21509, kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-RhtA	IPTG-inducible trc promoter expressing *rhtA* amplified from DSM 29578[2], kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-UctC	IPTG-inducible trc promoter expressing *uctC* from amplified from DSM 29578[2], kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-HutH2	IPTG-inducible trc promoter expressing *hutH_2* amplified from ATCC 9447, kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-ProY1	IPTG-inducible trc promoter expressing *proY_1* amplified from ATCC 9447, kanamycin resistance, RK2 origin of replication	This study

Plasmids were transformed into the respective *Pseudomonas* strain via electroporation based on the method of Choi et al. [14]. Electroporations were performed by pulsing either 1.8 or 2.5 kV through a 0.1 or 0.2 cm electroporation cuvette using a MicroPulser™ Electroporation Apparatus (Bio-Rad) then recovering the reaction for 2-3 hours at 30 °C. Culturing and phenazine quantification were performed as described in Chapter 2. A final concentration of 1mM sterile-filtered isopropyl β-D-1-thiogalactopyranoside (IPTG), was added to all cultures about 4 hours after inoculation. Cultures were supplemented with 50 μg/mL kanamycin sulfate in strains containing plasmids. Luria Bertani (LB) broth and TOP10 chemically competent *E. coli* cells were used for cloning.

Results

We input the phenazine production data and final genome assemblies from Chapters 2 and 3 into the mGWAS tool DBGWAS[12] to identify genomic loci that are significantly associated with phenazine production. This generated a preliminary list of 2493 significant hits across all phenotypes. To remove redundancies in the preliminary list, we collapsed the list in 3 stages: phenotype-collapsing, strain-collapsing, and reverse-complement collapsing. The resulting final list contained 330 significant hits, which were associated with 158 genes, including 78 hypothetical proteins.

We used our phenotype score metric to select which hits to experimentally validate *in vivo*. In the 'phenotype-collapsing' stage, we assigned each hit a 'phenotype score' which is the number of phenotypes where a hit was significant. We targeted hits with higher phenotype scores because they were significant for more phenotypes and thus we may observe effects on multiple phenotypes (PCN production in KMB, KMB+Fe, and the difference between production in both media) *in vivo*. 6 significant unitigs had phenotype scores of 3 or greater. Of those unitigs, 4 were significant for PCN production phenotypes and 2 was important for PCA production phenotypes. We chose to focus on the hits which were significant for PCN production, as our top phenazine-producing strain DSM 21509 primarily produces PCN.

To assess whether these genomic regions were significant for phenazine production, I overexpressed the CDSs associated with the top phenotype-scoring unitigs in DSM 21509, the top producer of PCN and total phenazines in both media conditions (Table 4.3). Only one CDS had a significant effect on PCN production in KMB (Figure

4.1). Significant unitig 1 partially overlapped a CDS annotated in DSM 21509 as PS__04252, a hypothetical protein. PS__04252 overexpression led to 19.1 ± 3.5 mg/L PCN in KMB compared to the 58.7 ± 12.3 mg/L produced by the empty vector control, which is a 67.5% decrease. PS__04252 has 57% sequence identify and 97% coverage to helix-turn-helix (HTH) transcriptional regulator PA1607 (NCBI Reference Sequence NP_250298.1)) from *P. aeruginosa* belonging to the multiple antibiotic resistance regulator (MarR) family[11]. MarR regulators typically function as transcriptional repressors for antibiotic efflux pump or stress response genes and derepress in response to an antibiotic ligand or oxidative stress[15].

Table 4.3. Top hits identified from mGWAS analysis.

	Unitig length	Phenotype score and effect		Major phenotype	Associated gene(s)	Min. q-value	Genomic region	Strains
1	31	+	4	PCN Total phz	Hypothetical protein (PS__04252) Hypothetical protein (PS__04251)	2.6×10^{-14}	non-coding	DSM 21509 DSM 29578[1] DSM 29578[2] NCCB 100368[1]
2	48	+	4	PCN Total phz	Putative isomerase (YbhH)	2.6×10^{-14}	CDS	DSM 21509 DSM 29578[1] DSM 29578[2] NCCB 100368[1]
3	40	-	3	PCN	Threonine/homoserine exporter (RhtA) Acetyl-CoA oxalate CoA-transferase (UctC)	6.7×10^{-10}	non-coding	DSM 29578[1] DSM 29578[2] NCCB 100368[2]
4	33	+	3	PCN	Proline-specific permease (ProY_1) Histidine ammonia-lyase (HutH_2)	2.16×10^{-9}	non-coding	DSM 21509 DSM 29578[1] DSM 29578[2] NCCB 100368[1] NCCB 47033 ATCC 17411 ATCC 17809 ATCC 17414 ATCC 9446 ATCC 17814 ATCC 9447 ATCC 17811 ATCC 17810

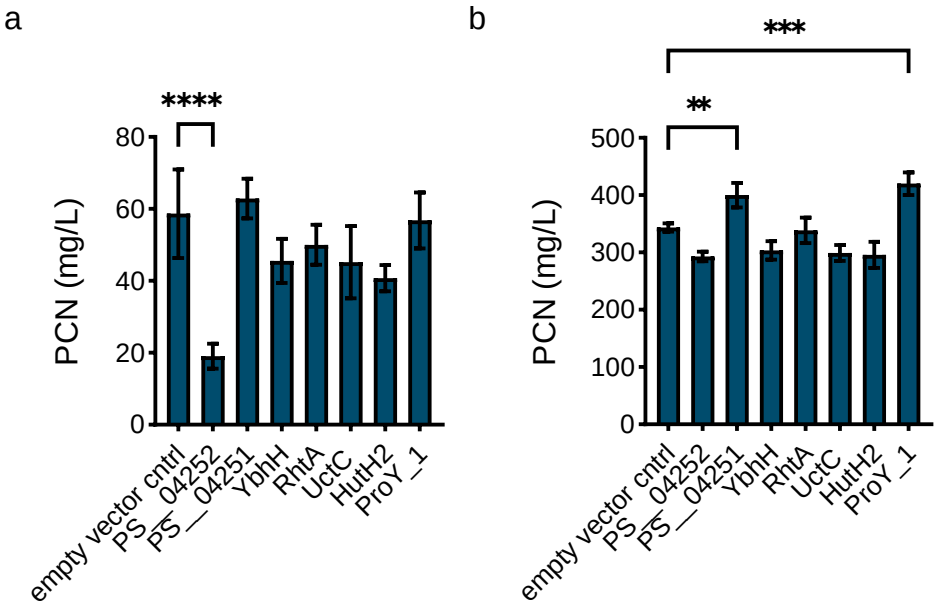

Figure 4.1. PCN production while overexpressing top mGWAS hits in DSM 21509. CDSs associated with top phenotype-scoring significant unitigs for PCN production were overexpressed in the top PCN-producing strain DSM 21509. PCN was quantified after 48 hr of culture in a) King's Media B and b) King's Media B + Fe. Asterisks denote p-values <0.01 (**), <0.001 (***) and <0.0001 (****) when performing an ordinary one-way ANOVA comparison to the empty vector control.

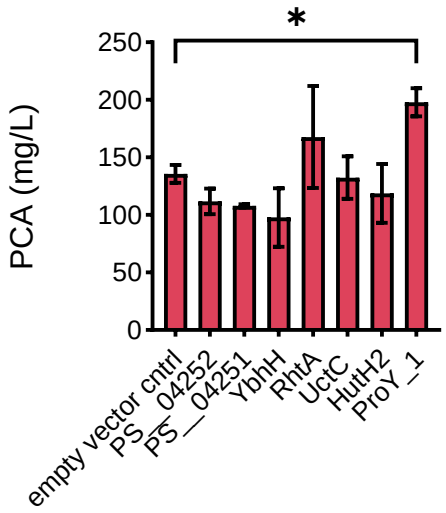

Figure 4.2. PCA production in King's Media B + Fe while overexpressing top mGWAS hits in NCCB 88062[1]. CDSs associated with top phenotype-scoring significant unitigs for PCN production were overexpressed in the top PCA-producing strain NCCB 88062[1]. PCA was quantified after 48 hr of culture in King's Media B + Fe. Asterisks denote p-values <0.05 (*), when performing an ordinary one-way ANOVA comparison to the empty vector control. Because all samples had similar 2-HPCA and 2-HP production, the total phenazine production graph has the same profile as the PCA graph, so only the PCA graph is pictured. There was no significant effect in KMB, so those results are also not pictured.

The other CDS adjacent to significant unitig 1, labeled as PS__04251 in our DSM 21509 annotations, is a hypothetical protein with unknown function. PS__04251 has 86% identity and 100% coverage to putative M14-type zinc cytosolic carboxypeptidase PSF113_3889 (NCBI Reference Sequence: WP_041476041.1) from *Pseudomonas ogarae* whose function is also unknown[16]. PS__04251 was also significant for phenazine production, as it produced 400.1 ± 21.5 mg/L PCN in KMB+Fe compared to 343.6 ± 7.3 mg/L PCN produced by the empty vector control.

The other hit which improved PCN production in KMB+Fe was ProY_1, associated with significant unitig 4. ProY_1 overexpression in DSM 21509 significantly increased PCN production to 420.2 ± 19.7 mg/L in KMB+Fe, which is a 22% increase over the empty vector control. Significant unitig 4 occurred within a non-coding region of the *hut* operon, which is responsible for histidine catabolism. Due to its position in the highly conserved operon and its sequence similarity, the hit annotated as ProY_1 is likely the histidine permease HutT which imports histidine and is required for its utilization[17]. The other adjacent CDS is HutH_2, the histidine ammonia lyase (HutH2) which catalyzes the first step of histidine catabolism: the conversion of L-histidine into urocanate[18]. While HutH_2 is part of the same operon, its overexpression did not improve phenazine production.

ProY_1 also significantly increased PCA production and total phenazine production in KMB+Fe when overexpressed within NCCB 88062[1], the top PCA/2-HPCA/2-HP producer (Figure 4.2). Overexpressing ProY_1 produced 198.0 ± 12.2 mg/L PCA (290.2 ± 20.2 mg/L total phenazines) compared to 135.8 ± 7.7 mg/L PCA (210.2 ±

18.1 mg/L total phenazines) produced by the empty vector control. This average titer increase of 62.2 mg/L represents approximately a 46% increase in PCA production. I overexpressed the other top phenotype-scoring hits in NCCB 88062[1] as well, but there were no significant effects for any hit in KMB or for any other hit in KMB+Fe. This is not surprising as none of these hits were significant for PCA production phenotypes.

Discussion

Many of the hits generated by the microbial GWAS analysis are connected to oxidative stress responses, which seems logical as phenazines generate reactive oxygen species. Many MarR-family regulators respond to oxidative stress, including PA1607 which has some similarity to PS__04252 as well as conserved cysteine residues which suggests its role as an oxidative stress response protein. One study suggests that histidine catabolism in *Pseudomonas fluorescens* is connected to oxidative stress response, as it could increase intracellular pools of the antioxidant α-ketoglutarate[19]. Possibly overexpressing ProY_1 (HutT) could increase flux through the histidine catabolism pathway by increasing intracellular histidine levels and lead to oxidative stress response in this manner. While it did not significantly affect phenazine production, YbhH has 34% identity and 92% coverage with *E. coli* YbhH, an uncharacterized protein which is upregulated during stress response by σ^E sigma factor[20]. Some of the other untested hits on the final list have connections to oxidative stress as well, including RscC[20], glutathione synthase[21, 22] and glucose-6-phosphate 1-dehydrogenase (zwf)[23].

It would be interesting to see how our final hit list for phenazine production would compare to the hits that would be generated by mGWAS of the other phenotypes

measured in Chapter 2: growth temperature and biofilm formation. With the known links among phenazine production, biofilm formation, growth temperature, and stress responses there could be some hits which appear on both lists.

Novel transcriptional regulators could make good metabolic engineering targets for phenazine production, due to the apparent success with PS__04252 and due to the literature showing the importance of transcriptional regulators for phenazine production. Our final hit list included other HTH-type transcriptional regulators including ArgP, BenM, CynR, MtrA, and RhaS, which could be targeted in future efforts. The list also includes GacA, a global transcriptional regulator known to impact phenazine production[8, 24].

Overall, we were successful in identifying and confirming 3 CDSs associated with 2 genomic loci which significantly affected phenazine production. Out of the 7 CDSs which were overexpressed in DSM 21509 cultured in KMB+Fe, 2 significantly increased PCN titers over the empty vector control: ProY_1 and PS__04251, and one hit, PS__04252, decreased PCN production in KMB. While these hits make relatively small titer changes overall (<100 mg/L), these are decent improvements for a single simple modification. These hits could be combined with each other or other known metabolic engineering targets to achieve even higher titers. These results demonstrate that this is a promising new approach to metabolic engineering.

References

1. Hollas, A., et al., *A biomimetic high-capacity phenazine-based anolyte for aqueous organic redox flow batteries.* Nature Energy, 2018. **3**(6): p. 508-514.

2. Rabaey, K., et al., *Microbial phenazine production enhances electron transfer in biofuel cells.* Environ Sci Technol, 2005. **39**(9): p. 3401-8.

3. Clifford, E.R., et al., *Phenazines as model low-midpoint potential electron shuttles for photosynthetic bioelectrochemical systems.* Chemical Science, 2021. **12**(9): p. 3328-3338.

4. Wan, Y., et al., *Biosynthesis and metabolic engineering of 1-hydroxyphenazine in Pseudomonas chlororaphis H18.* Microb Cell Fact, 2021. **20**(1): p. 235.

5. Guo, S.Q., et al., *Metabolic Engineering of Pseudomonas chlororaphis for De Novo Production of Iodinin from Glycerol.* Acs Sustainable Chemistry & Engineering, 2022.

6. Yu, J.M., et al., *An upstream sequence modulates phenazine production at the level of transcription and translation in the biological control strain Pseudomonas chlororaphis 30-84.* PLoS One, 2018. **13**(2): p. e0193063.

7. Liu, K.Q., et al., *Genetic engineering of Pseudomonas chlororaphis GP72 for the enhanced production of 2-Hydroxyphenazine.* Microbial Cell Factories, 2016. **15**.

8. Li, L., et al., *Metabolic Engineering of Pseudomonas chlororaphis Qlu-1 for the Enhanced Production of Phenazine-1-carboxamide.* Journal of Agricultural and Food Chemistry, 2020. **68**(50): p. 14832-14840.

9. Lees, J.A. and S.D. Bentley, *Bacterial GWAS: not just gilding the lily.* Nat Rev Microbiol, 2016. **14**(7): p. 406.

10. Brynildsrud, O., et al., *Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary.* Genome Biol, 2016. **17**(1): p. 238.

11. Lees, J.A., et al., *Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes.* Nat Commun, 2016. **7**: p. 12797.

12. Jaillard, M., et al., *A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events.* PLoS Genet, 2018. **14**(11): p. e1007758.

13. Cook, T.B., et al., *Genetic tools for reliable gene expression and recombineering in Pseudomonas putida.* J Ind Microbiol Biotechnol, 2018. **45**(7): p. 517-527.

14. Choi, K.H., A. Kumar, and H.P. Schweizer, *A 10-min method for preparation of highly electrocompetent Pseudomonas aeruginosa cells: application for DNA fragment transfer between chromosomes and plasmid transformation.* J Microbiol Methods, 2006. **64**(3): p. 391-7.

15. Housseini, B.I.K., G. Phan, and I. Broutin, *Functional Mechanism of the Efflux Pumps Transcription Regulators From Pseudomonas aeruginosa Based on 3D Structures.* Front Mol Biosci, 2018. **5**: p. 57.

16. Rimsa, V., et al., *High-resolution structure of the M14-type cytosolic carboxypeptidase from Burkholderia cenocepacia refined exploiting PDB_REDO strategies.* Acta Crystallogr D Biol Crystallogr, 2014. **70**(Pt 2): p. 279-89.

17. Zhang, X.X., et al., *Variation in transport explains polymorphism of histidine and urocanate utilization in a natural Pseudomonas population.* Environ Microbiol, 2012. **14**(8): p. 1941-51.

18. Zhang, X.X. and P.B. Rainey, *Genetic analysis of the histidine utilization (hut) genes in Pseudomonas fluorescens SBW25.* Genetics, 2007. **176**(4): p. 2165-76.

19. Lemire, J., et al., *Histidine is a source of the antioxidant, alpha-ketoglutarate, in Pseudomonas fluorescens challenged by oxidative stress.* FEMS Microbiol Lett, 2010. **309**(2): p. 170-7.

20. Bury-Mone, S., et al., *Global analysis of extracytoplasmic stress signaling in Escherichia coli.* PLoS Genet, 2009. **5**(9): p. e1000651.

21. Nikel, P.I., et al., *Reconfiguration of metabolic fluxes in Pseudomonas putida as a response to sub-lethal oxidative stress.* ISME J, 2021. **15**(6): p. 1751-1766.

22. Wongsaroj, L., et al., *Pseudomonas aeruginosa glutathione biosynthesis genes play multiple roles in stress protection, bacterial virulence and biofilm formation.* PLoS One, 2018. **13**(10): p. e0205815.

23. Kim, J., C.O. Jeon, and W. Park, *Dual regulation of zwf-1 by both 2-keto-3-deoxy-6-phosphogluconate and oxidative stress in Pseudomonas putida.* Microbiology (Reading), 2008. **154**(Pt 12): p. 3905-3916.

24. Wang, D., et al., *Roles of the Gac-Rsm pathway in the regulation of phenazine biosynthesis in Pseudomonas chlororaphis 30-84.* Microbiologyopen, 2013. **2**(3): p. 505-24.

Conclusion

This dissertation starts out by illustrating the opportunities for developing non-traditional microbes with stress-tolerant phenotypes as novel biotechnology hosts in Chapter 1. Chapter 2 screens 34 *Pseudomonas* (33 *P. chlororaphis*) isolates available from the American Type Culture Collection, German Collection of Microorganisms and Cell Cultures, and the Westerdijk Fungal Biodiversity Institute's Netherlands Culture Collection of Bacteria for biofilm formation, growth temperature, and phenazine production to identify the best strain to develop as a non-traditional host for phenazine compounds. Chapter 3 sequences and assembles genomes for all these strains using both 2nd and 3rd generation sequencing technologies and multiple assembly algorithms to arrive at the optimal genome assemblies as resources for metabolic engineering. Chapter 4 used the datasets from Chapters 2 and 3 to identify metabolic engineering targets to improve phenazine production using a population genomics-guided approach.

Strain DSM 21509 was identified as the best strain for phenazine production due to its low biofilm formation and naturally high phenazine titers (~0.5 g/L). The next steps for assessing DSM 21509 as a biotechnology host for phenazine production would be scaling up from the 96-well plate cultures used for screening to a small bioreactor fermentation. Not only would a bioreactor more accurately portray industrial conditions and increase total phenazine yield, but it would allow one to control and monitor process parameters that are important for phenazine production and cell growth such as dissolved oxygen levels, pH, and cell density. Additionally, phenazines induce the formation of reactive oxygen species causing oxidative stress in cells. Trying an extractive

fermentation process to remove the phenazines which accumulate in the media could promote healthy cell growth and encourage continued phenazine production.

I would also like to measure phenazine production over a longer time scale, such as 5-7 days rather than only after 48 hours of culture. As seen in Figure 2.4, pigment accumulates over time so having phenazine production data over a longer period could lead to higher phenazine yields. We could also do additional phenotyping that would be useful for selecting a microbiology host, such as antibiotic resistance profiles. Pseudomonads tend to have resistance to many common antibiotic resistance markers, and an example of a molecular mechanism for this is shown in Figure 1.3.

I assembled the genomes of all 34 strains using the assembler Unicycler and both Illumina and Oxford Nanopore reads as input. Most of the assemblies are almost complete but could use some manual curation to complete them and check for inaccuracies. Some assemblies have a single chromosome that has not yet been completely circularized and others have other smaller contigs that may be plasmids or evidence of misassemblies. I would also like to annotate these assemblies more accurately. During the microbial GWAS analysis, about half the hits mapped to hypothetical proteins and many of the CDSs which were annotated had annotations from *E. coli* or bacteria from other genera which did not seem to be accurate. I think it would be more useful if we had annotations based on a similar *Pseudomonas* species, such as *Pseudomonas fluorescens* or *Pseudomonas aeruginosa* if such annotations were available.

This work successfully identified 330 short sequences significantly associated with phenazine production. Out of the 7 CDSs which were overexpressed in DSM 21509, 2 significantly increased PCN titers over the empty vector control: ProY_1 and PS__04251, and one decreased phenazine production: PS__04252. While this work successfully identified 3 novel metabolic engineering targets for phenazine production, it employed a rather small collection of 34 isolates. For this study I was limited to strains that were available to our institution for purchase. Through research partnerships with other laboratories, larger studies can take place with more strains. Increasing the number of strains would increase the statistical power of the hits.

Since overexpressing PS__04252 decreased phenazine production in King's media B, phenazine production could be quantified in a knock-out strain to test whether it would increase phenazine production. The coding sequences which improved PCN production (ProY_1 and PS__04251) could be overexpressed simultaneously as well to see if we observe a combinatorial affect and further improvements in phenazine-1-carboxamide production. Since ProY_1 improved phenazine production in both PCN and 2-HP producing strains, this hit could also be interesting to pursue in other strains. Some other hits from the list of 330 could be tested as well, maybe selected based off a metric other than phenotype score or by cross referencing that list with other data, such as transcriptomics. The biofilm formation and growth temperature datasets could also be input into the microbial GWAS to identify targets to go after which affect those phenotypes with and without iron. These are but a few possible paths to pursue to build upon the work of this dissertation.