# Pathway and Protein Engineering for Improved Glucaric Acid Production in *Escherichia coli*

by

Lisa Marie Guay

B.S. Chemical Engineering
B.A Economics
University of Arizona, 2013

Submitted to the Department of Chemical Engineering
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2019

Signature redacted

Signature of Author ..........................................................................

Lisa Marie Guay
Department of Chemical Engineering
January 11, 2019

Signature redacted

Certified by ..........................................................................

Kristala L. J. Prather
Professor of Chemical Engineering
Thesis Supervisor

Signature redacted

Accepted by ..........................................................................

Patrick S. Doyle
Professor of Chemical Engineering
Chairman, Committee for Graduate Students

# Pathway and Protein Engineering for Improved Glucaric Acid Production in *Escherichia coli*

by

Lisa Marie Guay

Submitted to the Department of Chemical Engineering
on January 11, 2019 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Chemical Engineering

## ABSTRACT

Microbial fermentation is an attractive method for the renewable production of chemicals. Glucaric acid was identified as a "top value added chemical from biomass" by the Department of Energy in 2004, and a biological route for its production from glucose in *E. coli* was developed in our lab in 2009. Two of the pathway enzymes, *myo*-inositol phosphate synthase (MIPS) and *myo*-inositol oxygenase (MIOX), appear to control flux. This work addressed several limitations of these reactions.

One approach was the relief of reactive oxygen species (ROS) to improve MIOX performance. MIOX converts *myo*-inositol (MI) to glucuronic acid. Overexpression of native catalase and superoxide dismutases led to significantly higher titers of glucuronic acid from MI. This result corresponded to better maintenance of MIOX activity and expression over the course of the fermentation. A reduction in labile iron levels, which are linked to ROS formation, was also shown to improve glucuronic acid titers.

A second approach was the examination of natural MIPS diversity. MIPS competes with central carbon metabolism for its substrate, glucose-6-phosphate. Thirty-one representative MIPS homologs were selected using a sequence similarity network. Nineteen variants produced detectible *myo*-inositol (MI) from glucose, and *H. contortus* MIPS performed equally well or better than the current *S. cerevisiae* MIPS. Interesting differences in stability were identified between the variants, and further work to explore the network may yield more information about important sequence features.

A third approach was the evaluation of screening methods for glucuronic and glucaric acid to support protein engineering. We attempted to extend a previous screen to growth from glucose, but while growth was achieved from MI, low flux appeared to prevent growth from glucose. A previously-developed biosensor based on the regulator CdaR was also tested. We discovered that the biosensor does not respond to glucaric acid but instead to a downstream metabolite, likely glycerate, and that the biosensor is affected by catabolite repression. While a reliable screen was not realized, our improved understanding of native regulation aids in the identification of alternative strategies.

This work overall produced significant improvements in the glucaric acid pathway and helped to identify opportunities for further development.

Thesis Supervisor: Kristala L. J. Prather
Title: Professor of Chemical Engineering

# Acknowledgements

To my advisor, Kris, thank you for taking a chance on me and providing the opportunity to try my hand at metabolic engineering research. Joining the Prather Lab was the best decision I made during my entire time at MIT. I really appreciated your perspective, thoughtfulness, and ability to provide the right level of support when I needed it. Thank you also for your encouragement of my outside interests in policy and leadership. I definitely feel that I took full advantage of all that MIT had to offer. I also want to thank my thesis committee members Cathy Drennan, Hadley Sikes, and Greg Stephanopoulos for all the support and advice over the years. It was wonderful to have you as mentors throughout my graduate school journey.

I am incredibly grateful to all the members of the Prather Lab. You were my day-to-day companions, and I always looked forward to seeing you even when my experiments were failing. Irene Brockman Reizman, thank you for patiently teaching me biology lab skills. Michael Hicks, you taught me everything I know about bioinformatics. Shawn Manchester, Amita Gupta, and Lisa Anderson, I really enjoyed working alongside you on the glucaric acid pathway. Aditya Kunjapur, thank you for making me feel welcome in the Prather Lab, especially when I first joined. Kat Tarasova, thank you for introducing me to the Science Policy Initiative and for dragging me unwillingly to social events. Sue Zanne Tan, I am so grateful for your steady support throughout my years in the lab and beyond. Stephanie Doong, thank you for listening to my venting about lab and life and for being an awesome friend. Cynthia Ni, I really appreciate all our discussions about community and society. Kevin Fox, thank you for putting up with the women of the Prather Lab and for taking the initiative to set up lab outings. Jennifer Kaczmarek, thank you for always being willing to lend a helping hand.

I was initially apprehensive about coming to MIT, but classmates, colleagues, and friends made the experience bearable, rewarding, and fun. Kristen, Kathryn, Leia, Brooke, Brinda, Shannon, Leslie, Lina, Orpheus, Garrett, Amos, Camille, and Yamini, thank you so much for everything. To everyone in the ChemE Communication Lab, thank you for helping me become a better communicator. Thank you to Ashdown House Executive Committee members Orpheus, Chris, Drew, Calvin, and Sai. Thank you to Rachel, Kat, Alec, Abigail, Jack, Peter, and other leaders in the Science Policy Initiative. Lastly, I am so glad I had the opportunity to work with my fellow Graduate Student Council Officers, Arolyn, Angie, Orpheus, Sarah, and Krithika, and ExComm members to help improve the graduate student experience at MIT.

To my family, I do not even have the words to express how thankful I am for your tremendous support. Mom and Dad, thank you for sending me off to UA, the U of A, and MIT to go after my education and interests. To my brother Andrew, thank you for providing some much-needed perspective and levity along my academic journey. And to my cats, Maverick, Merlin, Oliver, and Hemingway, for their unconditional snuggles and purrs.

Lisa Guay

# Table of Contents

# List of Figures

# List of Tables

# 1. Introduction

## 1.1. Metabolic Engineering Tools

Biomanufacturing is an attractive method for the sustainable production of fuels and commodity chemicals. However, fermentation processes often require extensive strain engineering and bioprocess optimization to reach the high yields and selectivity required for economic viability.[1] This kind of manipulation is difficult because metabolic networks are complex, and we have incomplete knowledge of the important interactions that affect overall phenotype and productivity. Moreover, the addition of heterologous enzymes and metabolites to a system further reduces our understanding of its overall behavior. However, metabolic engineers have developed many tools to help address these problems, and commercial processes have been developed for several commodity and specialty chemicals, including 1,4-butanediol, succinic acid, isoprene, isobutanol, acetic acid, polyethylene, and artemisinin.[2–5] An overview of some of these tools is provided below.

## 1.2. Strain and systems engineering

An organism's native metabolism is complex and employs many regulatory mechanisms to maintain homeostasis and respond to environmental fluctuations. However, the introduction of new pathways into an organism can lead to unexpected interactions between native metabolism and the introduced proteins and metabolites. A few methods to mediate these interactions to improve production are described below.

### 1.2.1. Improving flux

New pathways often produce low titers when they are first constructed. Pathway balancing, which involves tuning the relative expression levels of pathway enzymes, can help improve pathway flux and reduce the overall protein expression burden.

In addition, native metabolism may limit flux and pathway yield through competing side reactions. Metabolic databases such as KEGG,[6] as well as genome scale models[7] and computational tools like OptKnock[8] and PROPER,[9] can help identify native enzymes that may affect the pathway. Nonessential genes can be knocked out, and essential genes can be knocked down. Knockdown can be achieved via transcriptional, translational, and posttranslational control. Common mechanisms for implementing these types of control are CRISPR interference (CRISPRi) using dCas9 and targeted sgRNA,[10,11] RNA interference (RNAi),[12] and protein degradation tags,[13] respectively. Dynamic knockdown strategies have recently been developed that allow for additional flexibility in flux optimization.[14–16]

### 1.2.2. Alleviating toxicity

A frequent problem in bioprocesses is toxicity of the end product or a side or co-product. Product tolerance is often a complex phenotype determined by multiple genes. One successful approach is whole-cell evolution for growth in the presence of increasing concentrations of the toxic compound of interest. Following evolution, genome sequencing can help reveal the mutations responsible for the improvement,[12] which may suggest ways to further improve tolerance. In addition, comparing the transcriptome for organisms exposed and not exposed to the compound can also provide clues about how the cellular response may be improved.[17] Finally, overexpression or introduction of efflux pumps for a toxic product is a complementary approach that can reduce toxicity, simultaneously reducing the elevated intracellular concentration and enhancing product concentration in the supernatant.[12,17,18] In addition to the strategies for product toxicity, it may be possible to consume nonessential side or co-products using specific catabolic or scavenging enzymes.[19]

### 1.2.3. Overcoming regulation

Native regulation also poses challenges for bioprocess development. Organisms employ an extensive set of control systems to modulate metabolism. In engineered systems, the desired pathway may be subject to downregulation. This regulation is often achieved through allosteric control of enzymes or transcriptional control by protein regulators.[20,21] Transcriptional repression can be alleviated by knocking out regulator proteins or by constitutive expression of the regulated genes.[12] Relief of allosteric control has been achieved using enzyme engineering at the binding interface to prevent binding and render the target protein always active or inactive.[12,20,22] In addition, substitution of a homologous enzyme or alternative pathway from another organism can help circumvent native regulation.[20]

One global regulation system of considerable interest to metabolic engineers is carbon catabolite repression (CCR). CCR is common in bacteria and allows for the preferential utilization of available carbon sources. However, in metabolic engineering applications, it may downregulate necessary pathways and preclude efficient co-utilization of carbon sources.[23]

Glucose is the preferred carbon source in many bacteria, and import and catabolism of many other carbon sources are only activated in its absence.[24] In *E. coli*, the presence or absence of glucose is reflected in the phosphorylation state of EIIA in the phosphotransferase system (PTS). When glucose is absent, phosphorylated EIIA activates adenylate cyclase to produce cyclic AMP (cAMP), which binds to the cAMP receptor protein (CRP).[24] The CRP-cAMP complex is an important transcriptional activator, controlling expression of hundreds of genes.[25] In the presence of glucose, dephosphorylated EIIA can also bind to some transporters to prevent import of alternative carbon sources, a phenomenon known as inducer exclusion.[26,27] Another contributor to CCR is the catabolite repressor activator (Cra), which senses glycolytic flux through the relative levels of fructose-1,6-bisphosphate and fructose-1-phosphate.[28] Some relief of CCR has been achieved through knockouts of PTS system components (*ptsG, ptsHIcrr*) and glycolysis (*pgi*).[23,29,30] Engineering of CRP has also shown promise.[23,25,29]

## 1.3. Protein Engineering

Protein engineering comprises several methods that yield proteins with better stability, selectivity, and activity. Protein engineering is extensively used in metabolic engineering to optimize bioprocesses. The majority of enzymes have $k_{cat}/K_m$ values that are several orders of magnitude below the diffusion limit, and the most efficient enzymes tend to be involved in central carbon metabolism.[31] Less-efficient enzymes are unlikely to have experienced the same degree of selective pressure and may prove successful targets for engineering.[31] In addition, the introduction of an enzyme into a new organism or the overexpression of a native enzyme inherently changes its fitness landscape, further increasing the potential benefit for heterologous enzymes used in bioprocesses. Engineering has also been used to adapt enzymes to different substrates and temperatures, improving selectivity and stability.

### 1.3.1. Natural protein diversity

Naturally-occurring protein diversity is the starting point for much of protein engineering. Methods that involve modifying a template, including rational engineering and directed evolution described below, typically start from a sequence derived from nature. Until relatively recently, little was known about the extent of natural diversity within classes of sequences. However, as the cost of DNA sequencing has fallen, the amount of sequence information has accumulated exponentially (Figure 1.1).[32] Nevertheless, making effective use of the large amount of sequence information is challenging, because functional information and experimental characterization lag well behind sequencing. This disparity is illustrated in Figure 1.1 by the gap between the blue line representing total sequences and the orange line representing reviewed sequences. A number of databases now attempt to classify sequences by motifs, domains, and homology into putative enzyme families or superfamilies. Two common ones are InterPro[33] and one of its component databases, Pfam.[34]

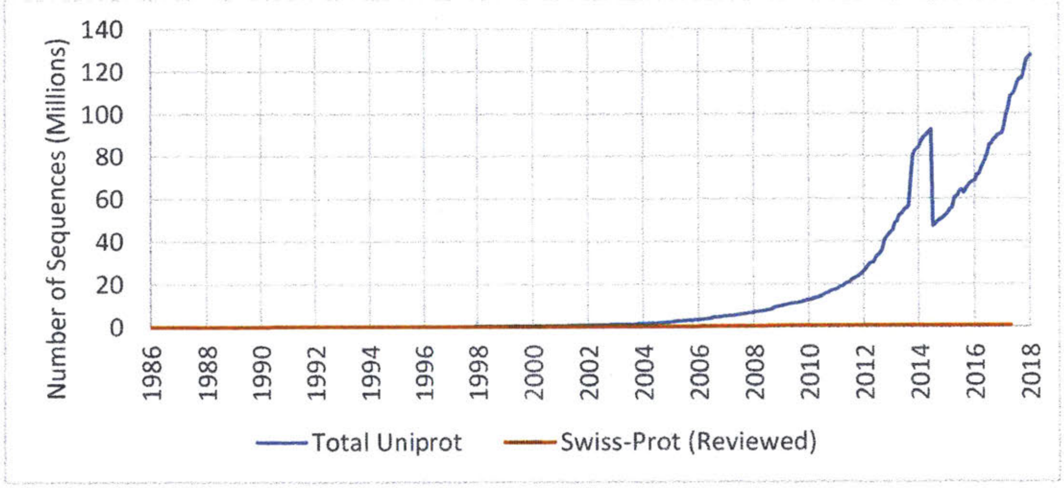

Figure 1.1. Number of sequences in UniProt databases, 1986-2018. The blue line represents the total number of sequences in both Swiss-Prot (reviewed) and TrEMBL (unreviewed) databases that comprise UniProt. The orange line represents only Swiss-Prot sequences.

In general, bioinformatics tools are most powerful where distinguishing information or features exist between proteins in a class or between related classes of proteins. However, these tools often require as input experimental or functional information about individual proteins. Proteins within a single class are likely to exhibit differences in stability, and approaches using consensus and correlated residues have proven effective.[35,36] Bioinformatics can also aid in determining sequence differences in enzyme function and allosteric regulation between larger families or superfamilies.[35,37] Selectivity, on the other hand, has proven more challenging because sequences alone do not provide reliable information about spatial interactions.[35]

Bioinformatics tools in combination with other protein engineering methods can help to address some of these limitations. Different evolutionary trajectories may be accessible from different sequence templates, so using homologs may allow for additional exploration of sequence space[38]. In addition, homologous recombination of related sequences, with techniques such as DNA shuffling, is an effective library generation method for directed evolution.[36,39,40]

### 1.3.2. Rational engineering

Rational engineering usually involves creating and testing a small library of targeted ("rational") mutations. The approach relies on knowledge about the protein of interest to identify amino acid mutations that may improve the property of interest.[41] For this reason, the availability of information about the overall structure and mechanism, as well as residues in the active site and binding pockets, is often crucial for effective rational engineering. Molecular modeling tools based on molecular dynamics and quantum mechanics are often used to guide prediction of beneficial mutations.

However, rational engineering remains challenging. First, many enzymes are not well-characterized. Second, even when structural and functional information is available, it is difficult to choose the best locations for mutagenesis, as residues far away from the active site and binding pockets have been found to be important for overall function.[42–44] Since these regions have typically not been well-studied even in well-characterized proteins, molecular modeling approaches also struggle.[42]

### 1.3.3. Directed evolution

In contrast, directed evolution is a powerful tool to change an enzyme's activity, specificity, and stability without *a priori* knowledge of its structure or catalytic mechanism. Directed evolution relies on the creation of a diverse library of protein sequences followed by screening or selection to identify the top performers. It can also be used iteratively to allow the accumulation of beneficial mutations. Beneficial mutations are rare,[45] so directed evolution relies on large libraries and high-throughput screens.[46]

Directed evolution is a very general method, and the results of a particular experiment depend on the details of both library generation and screening or selection. How libraries are generated determines which sequence variants may be detected. Sequence space is vast – for any given protein of N amino acids, there are $20^N$ possible sequence variants. For a relatively small 100 amino acid protein, this translates to approximately 1.3 x $10^{130}$ possible sequences, far larger than the estimated number of atoms in the universe. Clearly, generating and testing all sequences is impossible. However, library generation fundamentally determines the portion of sequence space available in a directed evolution study. Moreover, fitness landscapes, which define the relationship of sequence to fitness, often contain epistatic sequence interactions, limiting the accessible evolutionary trajectories.[47] Advances have been made to reduce the bias in random mutation methods,[48,49] but these inherent limitations remain.

The particular screen or selection method used also has consequences for the results of directed evolution. The context-dependence of screens and selections is memorably captured in the First Law of Directed Evolution: "You get what you screen for."[50] Mutations that improve the screen output but do not improve the protein of interest as intended are common, and these undesired mutations may well obscure the detection of desired mutations. This type of problem is common in metabolic engineering applications because different experimental conditions are often used for production and for screening or selection.

### 1.3.4. Development of screens and selections

The importance of the detection method to the results of directed evolution studies has led to significant work to develop and improve screens and selections. Any successful detection method must connect a sequence to a phenotype. Many different methods exist, but the most common phenotypes used are growth or production of a colored substance or fluorescent reporter.[40] Growth-based methods are often used for selections because growth phenotypes are relatively binary. Only the cells that are able to grow survive the selection and can be further characterized. In contrast, colored or fluorescent phenotypes are useful for screens. All cells must be examined to determine which ones are the most colored or fluorescent, and high-throughput screens such as fluorescence-activated cell sorting (FACS) are frequently used for this purpose.[40]

Many pathways and enzymes of interest in metabolic engineering do not directly produce an easily-detectible phenotype and therefore require screen or selection development. In some cases, the phenotype can be linked to growth under certain conditions, possibly with the use of strain engineering to knock out other growth pathways. In other cases, regulators may be used to create fluorescent or growth-associated biosensors. While details of biological control systems are still being elucidated, naturally-occurring or engineered transcription factors and riboswitches that bind to a metabolite of interest are increasingly being used to develop biosensors.[51–56] These are commonly used to drive production of a fluorescence or antibiotic resistance gene.

## 1.4. Glucaric Acid

Glucaric acid is a six-carbon aldaric acid that was named a "top value added chemical from biomass" in 2004 by the U.S. Department of Energy.[57] Glucaric acid and other aldaric acids can be used to produce lactone solvents, esters, metal-chelating surfactants, and a wide range of polymeric materials, including hydroxylated nylons and branched polyesters.[57] These wide-ranging applications make it an attractive target for replacing petroleum-based chemicals.

Conventional production involves selective oxidation of the aldehyde and terminal alcohol groups of glucose with nitric acid or other oxidizing agents. However, the oxidation produces low yields and a large range of difficult to separate glucose derivatives. Metal catalysts have been developed to help improve selectivity, but these processes are expensive.[58] Glucaric acid is also naturally produced in fruits, vegetables, and mammals, though the amounts are small and the pathways are lengthy.[6,59,60] Taken together, these limitations have so far precluded large-scale production.

### 1.4.1. Glucaric acid pathway in *E. coli*

A novel heterologous pathway was introduced in *E. coli* in 2009 and is shown in Figure 1.2.[59] The pathway uses three heterologous enzymes, *myo*-inositol-1-phosphate synthase (MIPS) from *S. cerevisiae*, *myo*-inositol oxygenase (MIOX) from *Mus musculus*, and uronate dehydrogenase (Udh) from *Pseudomonas syringae*. Glucose is first imported as glucose-6-phosphate (G6P) using *E. coli's* phosphotransferase system (PTS). MIPS converts G6P to *myo*-inositol-1-phosphate, using $NAD^+$ as a catalyst. The product is then dephosphorylated to *myo*-inositol (MI) by an endogenous phosphatase. Next, *myo*-inositol is oxidized to glucuronic acid by MIOX using molecular oxygen. Finally, glucaric acid is produced through a second oxidation by Udh, which consumes $NAD^+$. Titers of up to 2 g/L of glucaric acid have been produced from glucose using this pathway, with yields of 10-20%.[59,61]

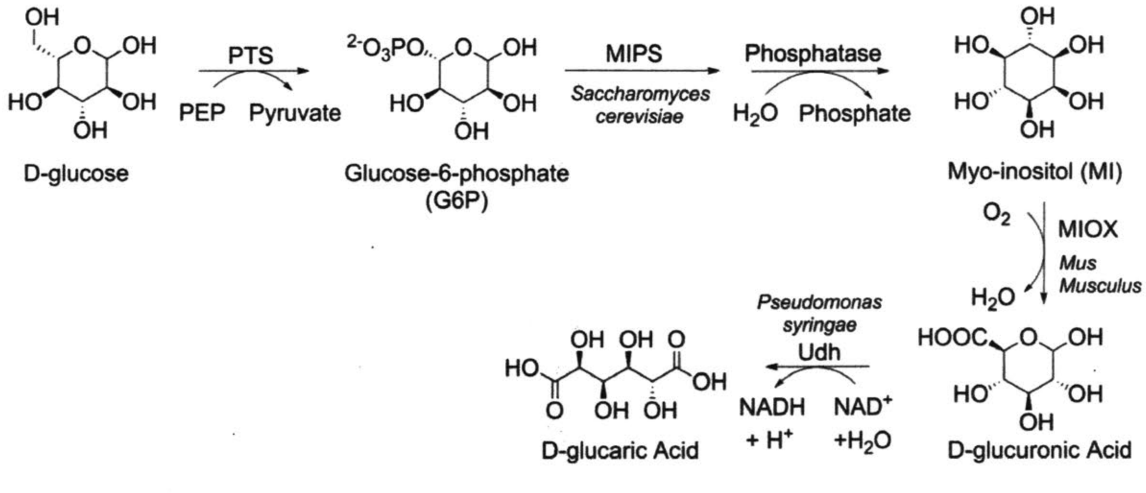

Figure 1.2. Heterologous pathway from glucose to glucaric acid in *E. coli*.

MIOX is an unusual oxidase and an unstable enzyme.[62,63] Like many monooxygenases, MIOX contains a non-heme diiron cluster in its active site. However, the mixed-valent Fe(II)-Fe(III) state is catalytically active instead of the more common Fe(II)-Fe(II) state.[62,64,65] This unusual redox state enables MIOX to perform the four-electron oxidation of MI to glucuronic acid using a single equivalent of molecular oxygen as the co-substrate.[62] It has been suggested that MIOX turnover may generate reactive oxygen species (ROS) through incomplete reduction of oxygen,[62] and hydrogen peroxide has been shown to inactivate the enzyme.[66,67] However, evidence that ROS is associated with MIOX expression or activity is mixed.[66,68–70] Nevertheless, MIOX activity declines significantly over the course of a typical fermentation experiment,[63] and MI accumulation has sometimes been observed in the context of the full glucaric acid pathway.[59]

MIPS catalyzes the first step in inositol biosynthesis and is essential in many organisms for generating cell membrane components and signaling molecules.[71] However, MIPS must compete for its substrate, glucose-6-phosphate, against major enzymes in central carbon metabolism, namely glucose-6-phosphate isomerase (encoded by *pgi*) of glycolysis and glucose-6-phosphate dehydrogenase (encoded by *zwf*) of the pentose phosphate pathway.[72] This competition limits glucaric acid titers from glucose, as much higher titers have been achieved from MI than from glucose.[59,63] In addition, *S. cerevisiae* MIPS currently limits pathway operation to 30°C because its activity falls at higher temperatures, whereas *M. musculus* MIOX performs better at 37°C.[59]

### 1.4.2. Previous engineering of the glucaric acid pathway

Pathway improvement has focused on the MIPS and MIOX enzymes because each appears to control pathway flux and overall titers under some conditions. Initial pathway characterization showed low *in vitro* activity for both enzymes relative to Udh, with MIOX activity an order of magnitude lower than MIPS activity.[59]

Several approaches have already been taken to improve the *M. musculus* (Mm) MIOX enzyme. First, the addition of an N-terminal small ubiquitin-like modifier (SUMO) fusion protein was shown to boost glucuronic and glucaric acid titers from MI by increasing soluble expression.[63] Second, colocalization of MIPS and MIOX led to an increase in MIOX specific activity and in product titers, possibly due to a stabilizing effect of higher local substrate concentrations.[73] Third, directed evolution was undertaken using a growth screen for the one-step conversion of MI to glucuronic acid, which resulted in the identification of a mutant with a partial gene insertion that increased the rate of MI import but did not improve production from glucose.[63] Fourth, dynamic regulation to delay expression of MIOX until MI accumulated in the culture led to increased glucaric acid production.[61] Finally, our lab has undertaken an effort to use bioinformatics to probe MIOX homologs for improved pathway performance in *S. cerevisiae* and *E. coli*.

Unlike MIOX, little protein engineering work has been completed for MIPS. However, strain and pathway engineering have enabled MIPS to better compete for its G6P substrate. Knocking out both *pgi* and *zwf* and co-feeding glucose with another sugar allowed for the separation of glucaric acid production (from glucose) and cell growth (from the additional sugar substrate), leading to improved yield.[72] In addition, dynamic downregulation of phosphofructokinase (*pfk*), which catalyzes the first committed step in glycolysis, led to improved titers and yield by improving the balance of growth and production.[74]

## 1.5. Thesis Scope

Building on previous work in our lab, we sought to further improve the productivity of the glucaric acid pathway while developing or evaluating additional metabolic engineering tools. We focused primarily on improving the reactions catalyzed by MIPS and MIOX due to their apparent role in controlling flux through the pathway.

Here, we show that the performance of MIOX is significantly impacted by reactive oxygen species. While the problem of oxidative stress has been discussed in the metabolic engineering literature, and a variety of solutions have been offered for particular situations, a general approach is lacking. In order to alleviate oxidative stress and improve conversion of MI to glucuronic acid by MIOX, we overexpress native catalase *katE* and superoxide dismutases *sodA* and *sodB*. We also show a connection between reactive oxygen species and labile iron pools.

Additionally, we employ sequence similarity networks to explore natural MIPS enzyme sequence diversity. Relatively little work has been done to directly improve MIPS for glucaric acid production, and MIPS is conserved across most branches of life. Thirty-one sequences are evaluated for MI production, and efforts to improve stability and activity are discussed.

Finally, we evaluate two different screens for glucuronic or glucaric acid production. Protein evolution of MIPS and MIOX is likely to benefit pathway productivity, but a previous growth screen from MI did not result in an improved MIOX enzyme. A growth screen from glucose is assessed, and its limitations are discussed. In addition, a previously characterized biosensor for glucaric acid is evaluated, and native regulation of glucaric acid catabolism in *E. coli* is clarified.

## 1.6. Thesis Organization

This thesis is organized into five chapters. Chapter 1 provides background on strain and protein engineering strategies to support bioprocess development. It also introduces the glucaric acid pathway in *E. coli* and outlines previous pathway optimization efforts. Chapter 2 describes work to alleviate oxidative stress and improve MIOX performance. Chapter 3 discusses a search for improved MIPS homologs guided by sequence similarity networks. Chapter 4 reports on efforts to develop a growth screen and a fluorescent screen for glucuronic or glucaric acid detection. Finally, Chapter 5 contains conclusions and future directions.

## 2. Alleviation of Reactive Oxygen Species

## Abstract

It has been suggested that the MIOX mechanism may produce reactive oxygen species (ROS). Endogenous scavenging systems are typically sufficient to reduce ROS to safe levels, but introduction or amplification of metabolic pathways through genetic engineering can exhaust this natural antioxidant capacity. We verified that ROS affect the conversion of MI to glucuronic acid by MIOX and then alleviated the damage using catalase and superoxide dismutases. Overexpression of native catalase *katE* increased overall glucuronic acid titers (up to 1.9-fold) as well as soluble MIOX levels and activity (up to 10.8-fold at 72 hours). Overexpression of superoxide dismutases *sodA* or *sodB* in combination with *katE* further increased titers, suggesting endogenous hydrogen peroxide and superoxide scavenging are insufficient in this system. The performance benefit observed with overexpression of catalytically inactive versions of iron-binding enzymes *katE* and *sodB* and with addition of chemical iron chelating agents also indicated a link between labile iron and ROS damage. The strategies used here to alleviate oxidative stress significantly improved performance of the glucaric acid pathway and may also be applied in other biological systems.

## 2.1. Introduction

Oxidative stress, the systemic cellular damage associated with elevated levels of reactive oxygen species (ROS), is a common problem in biological systems. Three major biologically-relevant ROS are superoxide ($O_2^{\bullet-}$), hydrogen peroxide ($H_2O_2$), and the hydroxyl radical ($OH^{\bullet}$).[75] Cells continuously generate superoxide and hydrogen peroxide during normal metabolism.[76] In addition, hydrogen peroxide is a common weapon in cellular warfare because it freely crosses cell membranes.[75,77] Important biomolecules are damaged by ROS, and cells employ scavenging systems to mitigate this damage (Figure 2.1). Superoxide and hydrogen peroxide oxidize iron in iron-sulfur cluster and mononuclear iron proteins, leading to iron release and protein inactivation.[76,78] The hydroxyl radical is an even more potent oxidant and reacts with most biomolecules at the diffusion limit, catalyzing lipid peroxidation cascades, creating DNA lesions and breaks, and oxidizing proteins and sugars.[79,80] A hydroxyl radical is produced when hydrogen peroxide acts upon intracellular free or labile iron via the Fenton reaction ($H_2O_2 + Fe^{2+} \rightarrow OH^- + OH^{\bullet} + Fe^{3+}$).[81,82] Under oxidizing conditions, superoxide may be able to recycle the iron ($O_2^{\bullet-} + Fe^{3+} \rightarrow Fe^{2+} + O_2$), completing the Haber Weiss reaction (overall: $H_2O_2 + O_2^{\bullet-} \rightarrow O_2 + OH^- + OH^{\bullet}$) and allowing net iron-catalyzed hydroxyl generation.[80,83–85] As a group, ROS promote growth defects, enzyme inactivation, mutations, and cell death.[86]

Because of the damage potential of ROS, cells have developed sophisticated defense systems. Hydrogen peroxide present at low concentrations is parimarily reduced by peroxidases ($RH_2 + H_2O_2 \rightarrow R + 2\ H_2O$; reducing power often provided by NAD(P)H), while hydrogen peroxide present at high concentrations is largely disproportionated by catalases ($2\ H_2O_2 \rightarrow O_2 + 2\ H_2O$).[76] Superoxide is disproportionated by superoxide dismutases (SODs; $2\ O_2^{\bullet-} + 2\ H^+ \rightarrow O_2 + H_2O_2$).[87] Cells also use antioxidants and thiol proteins to preferentially react with hydrogen peroxide and superoxide.[79,88] The more reactive hydroxyl radical reacts too quickly and nonspecifically for enzymatic scavengers to be effective, and cells instead reduce its formation via the Fenton reaction by sequestering labile (chelatable and redox-active) iron.[75,79,89] Cells commonly employ both basal and transcriptionally-activated defense systems,[75] which are typically sufficient to protect cells in their native environments.

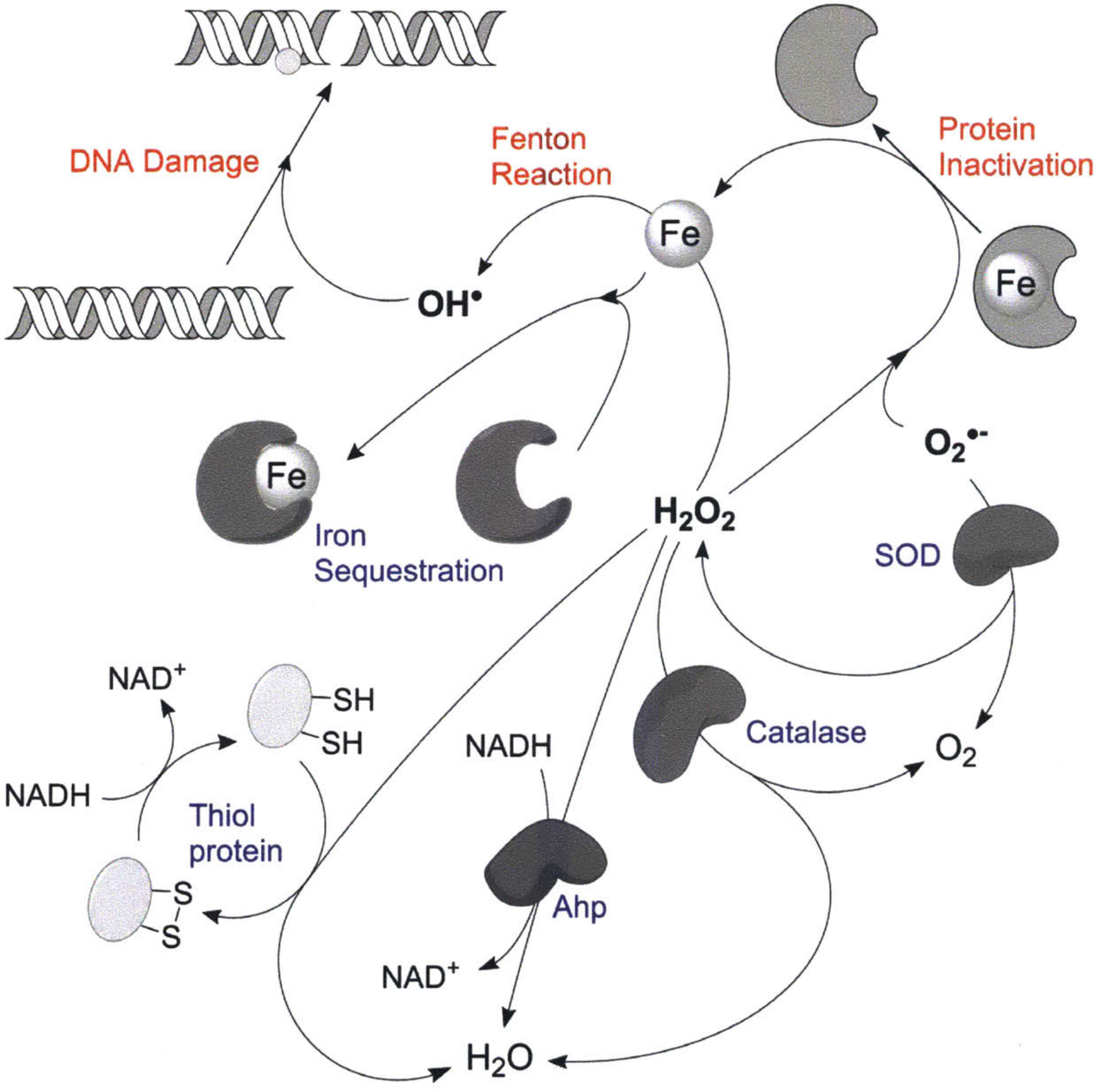

Figure 2.1. Overview of ROS damage and scavenging pathways in E. coli. The major ROS species hydrogen peroxide, superoxide radical, and hydroxyl radical are shown in bold. Methods of ROS damage are indicated in red, and methods of ROS scavenging are indicated in blue. Note that processes involving free and labile iron are simplified, and redox state and cycling steps are not shown.

Metabolic engineers have recently observed oxidative stress in several engineered pathways, which suggests that the native pathways to scavenge ROS may be insufficient in these contexts. Bioproduction of a wide range of products, including alkanes,[90] lipids,[91,92] acids,[93] and alcohols,[19,94–96] has been affected in bacterial, yeast, and algal hosts. Common factors in these pathways are incomplete reduction of oxygen by overexpressed oxygenases,[90,97] generation of ROS side products,[19] and production of unstable or toxic intermediates and products.[91–96] Approaches for alleviating oxidative stress have included overexpressing catalases,[90,93,98] peroxidases,[19,91] SODs,[92,98] thiol proteins,[94,99–101] and disulfide reductases,[91] as well as by adding antioxidants,[96,102] and iron chelators[98] to culture media. While these approaches have yielded positive results, little work has been done to evaluate or compare them, and a general framework for relieving oxidative stress has not yet been reported.

As discussed in Section 1.4.1, hydrogen peroxide has been shown to inactivate *myo*-inositol oxygenase (MIOX),[66,67] and MIOX turnover may generate ROS through incomplete reduction of oxygen.[62] However, it is unclear whether these issues are significant *in vivo.* Overexpression of native *Miox* has been associated with elevated levels of ROS in mice[68,69] and of ROS-scavenging enzymes in rice.[70] However, MIOX purified from hog kidney did not show increased hydrogen peroxide generation in the presence of its substrate, MI.[67] Thus, it is unclear how MIOX may affect overall ROS levels in an engineered microbial host.

Here, we demonstrate that ROS significantly reduce the performance of heterologous *Miox* expressed in two different strains of *E. coli*, suggesting limitations in the native scavenging systems. We then take a general and systematic approach to alleviating the damage, focusing on overexpression of native catalase and SODs.

## 2.2. Materials and Methods

### 2.2.1. Strains & plasmids

The *E. coli* strains and plasmids used in this study are listed in Table 2.1. Primers used for construction are listed in Table 2.2. *E. coli* strain DH5α was used for molecular cloning and plasmid preparation. The *E. coli* strains used for production were derived from either MG1655 (DE3) or BL21Star (DE3). Knockouts of *gudD* and *uxaC* were performed by sequential P1 transduction using Keio collection donor strains JW2258-5 and JW3603-2, respectively.[103] FLP recombinase expressed from plasmid pCP20 was used to cure the kanamycin resistance cassette after each transduction.[104] Transduction and curing were verified by PCR amplification and sequencing using primer pairs IB185 and IB186 for *gudD* and LMG1 and LMG2 for *uxaC*. The resulting double knockout strains used for glucuronic acid production are LG1458 (derived from MG1655) and LG1460 (derived from BL21Star).

Integration of *udh* from *A. tumefaciens* into the *E. coli* genome was performed via "clonetegration" (See Appendix A.1).[105] Primers LG49 and LG55 were used to amplify the coding sequence of *udh* from plasmid pTATudh2[106] and place it under the control of constitutive Anderson promoter BBa_J23100 (1.0 measured relative promoter strength).[107] This insert and the pOSIP-CH backbone were each digested with BamHI and SpeI then ligated. The ligation product was used to transform LG1458 and LG1460 for integration at the HK022 locus. The phage integration and chloramphenicol antibiotic resistance genes were cured from the MG1655 strain using FLP recombinase expressed from pE-FLP as previously described.[105] After difficulty transforming the BL21 strain with pE-FLP, we constructed an anhydrotetracycline (aTc)-inducible version. The pE-FLP plasmid backbone (excluding the constitutive pE promoter) was amplified using primers LG29 and LG30. This insert and plasmid pKVS45 (containing *tetR-P$_{tet}$*) were each digested with AvrII and XhoI and ligated. The resulting pE-Ptet-FLP plasmid was used to first transform the BL21 strain and then express FLP recombinase (induced with 50 ng/mL aTc) in a second step to remove the integration cassette. Integration and curing were verified by PCR amplification and sequencing with HK022 primers 1-4[105] and LG77 and LG78. The resulting glucaric acid production strains are MG1655 (DE3) Δ*gudD* Δ*uxaC* HK022::1.0-AtUdh (LG2477) and BL21 (DE3) Δ*gudD* Δ*uxaC* HK022::1.0-AtUdh (LG2512).

Plasmids used to express glucaric acid pathway genes and ROS scavenging genes were constructed from Duet vectors (Novagen, Darmstadt, Germany). PCR and sequencing primers LG73-76 were used for all vectors described below. For this study, we used *M. musculus* MIOX fused to an N'-terminal small ubiquitin-like modifier protein (SUMO) tag that was previously shown to increase soluble expression and overall pathway flux.[63] Catalase *katE* was amplified from *E. coli* strain MG1655 genomic DNA using primers LG69 and LG70. The insert and pRSFD-SUMO-MIOX were each digested with MfeI and AvrII and then ligated to create pRSFD-SUMO-MIOX-katE. pRSFD-SUMO-MIOX-katE(H128A) was created by amplifying pRSFD-SUMO-MIOX-katE with primers LG83 and LG84 designed to introduce the H128A mutation into *katE* using Agilent's QuikChange primer design web tool.[108]

Plasmids to evaluate SOD, alone and in combination with catalase, were derived from the pETDuet-1 backbone. Mn- and Fe- superoxide dismutases *sodA* and *sodB* were amplified from *E. coli* strain MG1655 genomic DNA using primer pair LG115 and LG116 and primer pair LG117 and LG118, respectively. The SOD inserts and the pETDuet-1 backbone were each digested with AscI and NotI then ligated to produce pET-sodA and pET-sodB. Analogous catalase plasmids were constructed by digesting pETDuet-1 as well as pRSFD-SUMO-MIOX-katE and pRSFD-SUMO-MIOX-katE(H128A) with MfeI and AvrII, followed by ligation to produce pET-katE and pET-katE(H128A), respectively. Four SOD-catalase combination plasmids (pET-sodA-katE, pET-sodA-katE(H128A), pET-sodB-katE, and pET-sodB-katE(H128A)) were assembled by digestion with MfeI and AvrII and ligation, using pET-katE or pET-katE(H128A) with pET-sodA or pET-sodB as appropriate. A second set of four SOD-catalase combination plasmids with SOD mutations (pET-sodA(Q147E)-katE, pET-sodA(Q147E)-katE(H128A), pET-sodB(Q70E)-katE, and pET-sodB(Q70E)-katE(H128A)) were constructed by amplifying pET-sodA-katE and pET-sodA-katE(H128A) with primers LG119 and LG120 and by amplifying pET-sodB-katE and pET-sodB-katE(H128A) with primers LG121 and LG122 designed using Agilent's QuikChange primer design web tool.[108]

For glucaric acid production, pRSFD-IN-SUMO-MIOX was created by digesting pRSFD-IN and pRSFD-SUMO-MIOX with MfeI and AvrII and ligating. Low-copy plasmids expressing catalase were constructed by digesting pACYCDuet-1 as well as pRSFD-SUMO-MIOX-katE and pRSFD-SUMO-MIOX-katE(H128A) with AscI and AvrII and ligating to produce pACYC-katE and pACYC-katE(H128A), respectively.

### 2.2.2. Culture conditions

For glucuronic and glucaric acid production, strains were grown in 250 mL baffled flasks containing 50 mL Luria-Bertani (LB) medium supplemented with either 60 mM *myo*-inositol (MI; 10.8 g/L; Sigma-Aldrich, St. Louis, MA) or 10 g/L glucose (Sigma-Aldrich), respectively. Working cultures were inoculated to an optical density at 600 nm ($OD_{600}$) of 0.01 from overnight cultures grown in LB at 37°C without MI or glucose. Cultures were induced with 100 μM isopropyl β-D-1-thiogalactopyranoside (IPTG) and supplemented with kanamycin (50 μg/mL), carbenicillin (100 μg/mL), and chloramphenicol (34 μg/mL) as required. For the iron chelator study, cultures were also supplemented with deferoxamine mesylate (Sigma-Aldrich), 2,2'-bypyridine (2,2'-bipyridyl, Sigma-Aldrich), diethylenetriaminepentaacetic acid (DTPA; Sigma-Aldrich), and 1,10-phenanthroline (Sigma-Aldrich) at the indicated concentrations. Cultures were incubated at 30°C, 250 rpm, and 80% relative humidity for 72 hours, with samples taken periodically for measurements of biomass, enzyme activity, and extracellular metabolites.

### 2.2.3. Measurement of MIOX activity and expression level

Cell pellets were taken from 1.5 mL of culture media at 24, 48, and 72 hr after inoculation, washed twice in sodium phosphate buffer (0.1 M, pH 7.2), and resuspended in 200 μL B-PER (supplied in sodium phosphate buffer; Thermo Fisher Scientific, Waltham, MA) supplemented with an EDTA-free protease inhibitor cocktail (Roche Applied Science). Lysates were prepared by shaking at room temperature for 15 min followed by centrifugation, and total soluble protein was measured using a BCA assay kit (Thermo Fisher Scientific). MIOX activity was measured as previously described[109] and normalized by the total protein concentration. To compare the effect of exogenous catalase and superoxide dismutase on activity, 4.3 μg/mL purified bovine liver catalase (MP Biomedicals, Santa Ana, CA) and/or 7.5 μg/mL purified *E. coli* Mn superoxide dismutase (Sigma-Aldrich) were added to the assay reaction.

*Miox* expression was visualized by SDS-PAGE and Coomassie staining using a 10% polyacrylamide gel with 15 μg of total protein per lane (Bio-Rad Laboratories, Hercules, CA).

### 2.2.4. Measurement of extracellular metabolites

MI, glucuronic acid, glucaric acid, glucose, and acetate concentrations in culture supernatant samples were quantified by high performance liquid chromatography (HPLC) on an Agilent 1200 series instrument (Santa Clara, CA) with an Aminex HPX-87H anion exchange column (300 mm by 7.8 mm; Bio-Rad Laboratories) using 5 mM sulfuric acid at a flow rate 0.6 mL/min as the mobile phase. The column and refractive index detector temperatures were held at 45°C and 35°C, respectively. Compounds were quantified from 10 μL injections using the refractive index signal.

Supernatant hydrogen peroxide concentrations were quantified using the Amplex Red kit (Thermo Fisher Scientific) per manufacturer instructions.

### 2.2.5. Statistics

Reported values are the average of at least three replicates, and error bars denote one standard deviation above and below the mean value. P-values were calculated using paired or unpaired two-tailed student's t-tests with unequal variance.

Table 2.1. *E. coli* strains and plasmids used in this chapter

| Name | Genotype | Source |
|---|---|---|
| **Strains** | | |
| MG1655(DE3) | F-, λ-, ilvG-, frb-50, rph-1, (DE3) | Tseng, Martin, Nielsen, & Prather, 2009 |
| BL21Star(DE3) | F-, ompT, hsdSB (rB- mB-), gal, dcm, rne131, (DE3) | Thermo Fisher (Waltham, MA) |
| JW2258-5 | F-, Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ-, rph-1, Δ(rhaD-rhaB)568, hsdR514CGSC, ΔgudD785::kan$^R$ | CGSC #10161; Baba et al., 2006 |
| JW3603-2 | F-, Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ-, rph-1, Δ(rhaD-rhaB)568, hsdR514CGSC, ΔuxaC782::kan$^R$ | CGSC #10338, Baba et al., 2006 |
| LG1458 | MG1655 (DE3) ΔgudD ΔuxaC | This study |
| LG1460 | BL21 (DE3) ΔgudD ΔuxaC | This study |
| LG2477 | MG1655 (DE3) ΔgudD ΔuxaC HK022::1.0-AtUdh | This study |
| LG2512 | BL21 (DE3) ΔgudD ΔuxaC HK022::1.0-AtUdh | This study |
| **Plasmids** | | |
| pOSIP-CH | pUC *ori*, RK6γ *ori*, Cm$^R$, attP HK022, ccdB, HK022 integrase expressedby λ p$_r$ under control of λ cI857 | St-Pierre et al., 2013 |
| pE-FLP | R101 *ori*, *repA*101ts, Amp$^R$, FLP recombinase expressed by *pE* | St-Pierre et al., 2013 |
| pKVS45 | p15A *ori*, Amp$^R$, tetR, P$_{Tet}$ | Solomon, Sanders, & Prather, 2012 |
| pE-Ptet-FLP | *ori*R101, *repA*101ts, Amp$^R$, TetR, FLP recombinase expressed by P$_{tet}$ | This study |
| pTATudh2 | pTrc99SE, *udh* from *A. tumefaciens* | Yoon et al., 2009 |
| pCP20 | Rep$^a$, Amp$^R$, Cm$^R$, FLP recombinase expressed by λ p$_r$ under control of λ cI857 | CGSC #7629 |
| pRSFDuet-1 | pRSF1030 *ori*, lacI, Kan$^R$ | Novagen (Darmstadt, Germany) |

Table 2.1. *E. coli* strains and plasmids used in this chapter (cont.)

| Name | Name | Name |
|---|---|---|
| **Plasmids** | | |
| pETDuet-1 | ColE1(pBR322) *ori*, lacI, Amp$^R$ | Novagen (Darmstadt, Germany) |
| pACYCDuet-1 | p15A *ori*, lacI, Cm$^R$ | Novagen (Darmstadt, Germany) |
| pTrc-SUMO-MIOX | pTrc99A with SUMO-MIOX inserted into the NcoI and HindIII sites | Shiue & Prather, 2014 |
| pRSFD-SUMO-MIOX | pRSFDuet-1 with SUMO-MIOX inserted into the NcoI and HindIII sites | Shiue & Prather, 2014 |
| pRSFD-SUMO-MIOX-katE | pRSFD-SUMO-MIOX with *E. coli* katE inserted into the MfeI and AvrII sites | This study |
| pRSFD-SUMO-MIOX-katE(H128A) | pRSFD-SUMO-MIOX-katE with katE His-128 mutated to Ala | This study |
| pET-katE | pETDuet-1 with *E. coli* katE inserted into the MfeI and AvrII sites | This study |
| pET-katE(H128A) | pETDuet-1 with katE(H128A) inserted into the MfeI and AvrII sites | This study |
| pET-sodA | pETDuet-1 with *E. coli* sodA inserted into the AscI and NotI sites | This study |
| pET-sodB | pETDuet-1 with *E. coli* sodB inserted into the AscI and NotI sites | This study |
| pET-sodA-katE | pET-sodA with katE inserted into the MfeI and AvrII sites | This study |
| pET-sodA-katE(H128A) | pET-sodA with katE(H128A) inserted into the MfeI and AvrII sites | This study |
| pET-sodB-katE | pET-sodB with katE inserted into the MfeI and AvrII sites | This study |
| pET-sodB-katE(H128A) | pET-sodB with katE(H128A) inserted into the MfeI and AvrII sites | This study |
| pET-sodA(Q147E)-katE | pET-sodA-katE with sodA Gln-147 mutated to Glu | This study |
| pET-sodA(Q147E)-katE(H128A) | pET-sodA-katE(H128A) with sodA Gln-147 mutated to Glu | This study |
| pET-sodB(Q70E)-katE | pET-sodB-katE with sodB Gln-70 mutated to Glu | This study |
| pET-sodB(Q70E)-katE(H128A) | pET-sodB-katE(H128A) with sodB Gln-70 mutated to Glu | This study |
| pRSFD-IN | pRSFDuet-1 with S. cerevisiae INO1 inserted into the EcoRI and HindIII sites | Moon, Yoon, Lanza, et al., 2009 |

Table 2.1. *E. coli* strains and plasmids used in this chapter (cont.)

| Name | Name | Name |
|---|---|---|
| **Plasmids** | | |
| pRSFD-IN-SUMO-MI | pRSFD-IN with SUMO-MIOX inserted into the NcoI and HindIII sites | This study |
| pACYC-katE | pACYCDuet-1 with *E. coli* katE inserted into the MfeI and AvrII sites | This study |
| pACYC-katE(H128A) | pACYCDuet-1 with *E. coli* katE(H128A) inserted into the MfeI and AvrII sites | This study |

Table 2.2. Oligonucleotides used in this chapter.

| Primer | Sequence[a] |
|---|---|
| IB185 | gctatcgatacccactggatttgg |
| IB186 | aaccggagctgctggaact |
| LMG1 | ctaattcggcttccgtaccggt |
| LMG2 | acttcacgatctgccgcttg |
| LG29 | taagcaCCTAGGatgtactaaggaggttgtatgccac |
| LG30 | taagcaCTCGAGcaggtggcacttttcggg |
| LG49 | taagcaGGATCCttgacggctagctcagtcctaggtacagtgctagcggataacaatttcacacagg |
| LG55 | tgcttaACTAGTccgggtaccgagctctta |
| IB140 | ggaatcaatgcctgagtg |
| IB141 | acttaacggctgacatgg |
| IB142 | acgagtatcgagatggca |
| IB143 | ggcatcaacagcacattc |
| LG77 | ccgccataaactgccaggaattg |
| LG78 | cagtttaggttaggcgccatgc |
| LG69 | tgcttaCAATTGatgtcgcaacataacgaaaagaaccc |
| LG70 | tgtaacCCTAGGtcaggcaggaattttgtcaatcttagga |
| LG83 | gatccgcgtgcagcaacaatacgttccggaatgcgct |
| LG84 | agcgcattccggaacgtattgttgctgcacgcggatc |
| LG115 | tgcttaGGCGCGCCatgagctataccctgccatc |
| LG116 | tgcttaGCGGCCGCttattttttcgccgcaaaacg |
| LG117 | tgcttaGGCGCGCCatgtcattcgaattacctgcac |
| LG118 | tgcttaGCGGCCGCttatgcagcgagatttttcgc |
| LG119 | atcagcggagaatcctcgttagcagtagaaacca |
| LG120 | tggtttctactgctaacgaggattctccgctgat |
| LG121 | tatggttccagacctcagctgcgttgttgaatac |
| LG122 | gtattcaacaacgcagctgaggtctggaaccata |
| LG73 | ggcgctatcatgccataccg |
| LG74 | gattatgcggccgtgtacaatacg |
| LG75 | cgtattgtacacggccgcataatc |
| LG76 | gctagttattgctcagcggtgg |

[a] Capital letters indicate restriction enzyme cut sites; underlining designates promoter sequence.

## 2.3. Results

### 2.3.1. ROS and MIOX

As previously noted, the impact of ROS on MIOX is unclear given the conflicting reports between endogenous and *in vitro* systems. To begin to understand the relationship between MIOX and ROS in our system, we examined the effect of scavenging enzymes on MIOX activity. Purified catalase and SOD were added to cell lysates and *in vitro* MIOX activity was measured (Table 2.3). The addition of catalase led to a 60% increase in activity, while SOD did not produce a significant change in the one hour assay period.

Table 2.3. Effect of exogenous catalase and SOD on MIOX activity

Measured MIOX Activity (nmol/min/mg)

| Condition[†] | Average | SD | p-value[‡] |
|---|---|---|---|
| Control | 45.6 | 2.14 | |
| + Catalase | 72.9 | 2.77 | <0.00005 |
| + SOD | 43.9 | 2.11 | |
| + Catalase + SOD | 69.8 | 3.38 | <0.00005 |

[†] Strain LG1460 harboring plasmid pTrc-SUMO-MIOX was grown in LB with 60 mM MI, and quintuplicate cell pellet samples were taken at 13 hr. Lysates were supplemented with water, commercial purified bovine catalase, and/or *E. coli* Mn SOD prior to the one hour incubation step of the MIOX assay.

[‡] P-values relative to the control were calculated using paired two-tailed student's t-tests with unequal variance. The same sample lysates were evaluated under all four conditions.

We also measured hydrogen peroxide levels in the culture supernatants of strains with and without *Miox* expression (denoted MIOX and EV, respectively) and in the presence and absence of the substrate MI (indicated by +MI and –MI). In general, hydrogen peroxide levels were slightly higher in LG1458 (K strains) than in LG1460 (B strain), and they fell over the course of the fermentation in both strains (Figure 2.2). Cultures expressing *Miox* typically had lower levels than the empty vector (EV) control early in the fermentation but showed higher levels by the end of the fermentation. However, EV had lower cell density than MIOX beyond 6 hours, and the EV cell density decreased by the end of the experiment, while MIOX cell density remained constant (Figure 2.3). There were no consistent differences between the +MI and –MI samples for LG1458, but LG1460 MIOX +MI showed higher hydrogen peroxide levels than –MI at both 48 and 72 hours.

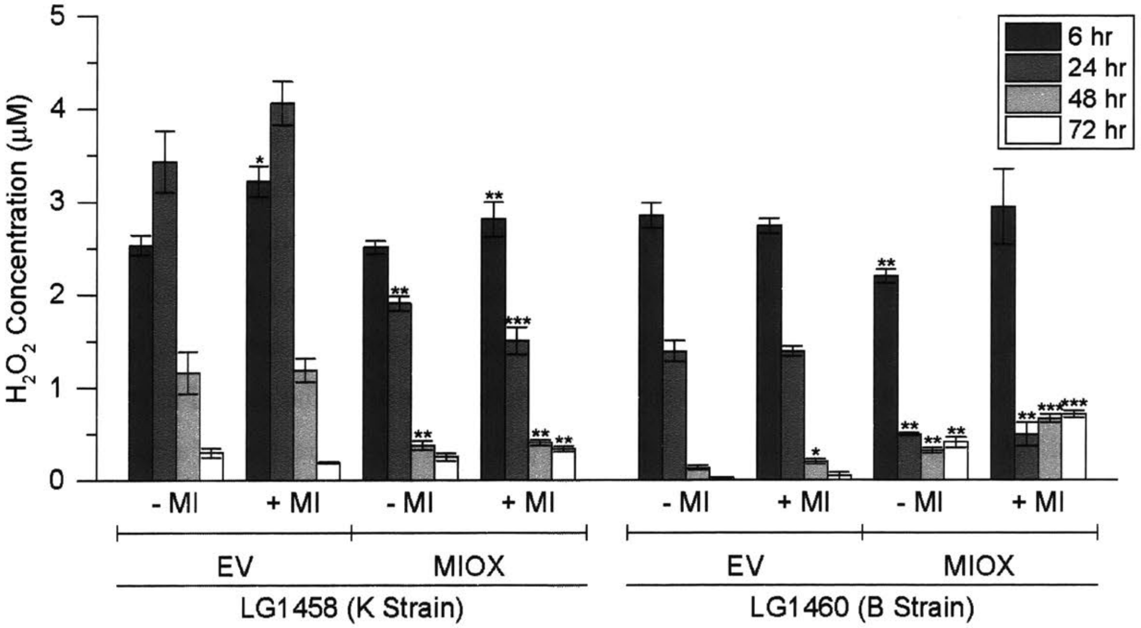

Figure 2.2. Supernatant hydrogen peroxide levels in *E. coli* expressing *Miox*. Strains LG1458 and LG1460 harboring pRSFDuet-1 ("EV") or pRSFD-SUMO-MIOX ("MIOX") were grown in LB with or without 60 mM MI ("-MI" and "+MI," respectively). Hydrogen peroxide levels were measured in the supernatant, and mean values ± SD for triplicate samples are shown. P-values were calculated for unpaired two-tailed student's t-tests with unequal variance. In all cases, * denotes $p < 0.05$ for +MI samples relative to -MI samples for the same strain and time point, ** denotes $p < 0.05$ for MIOX samples relative to EV samples, and *** denotes $p < 0.05$ for both comparisons.

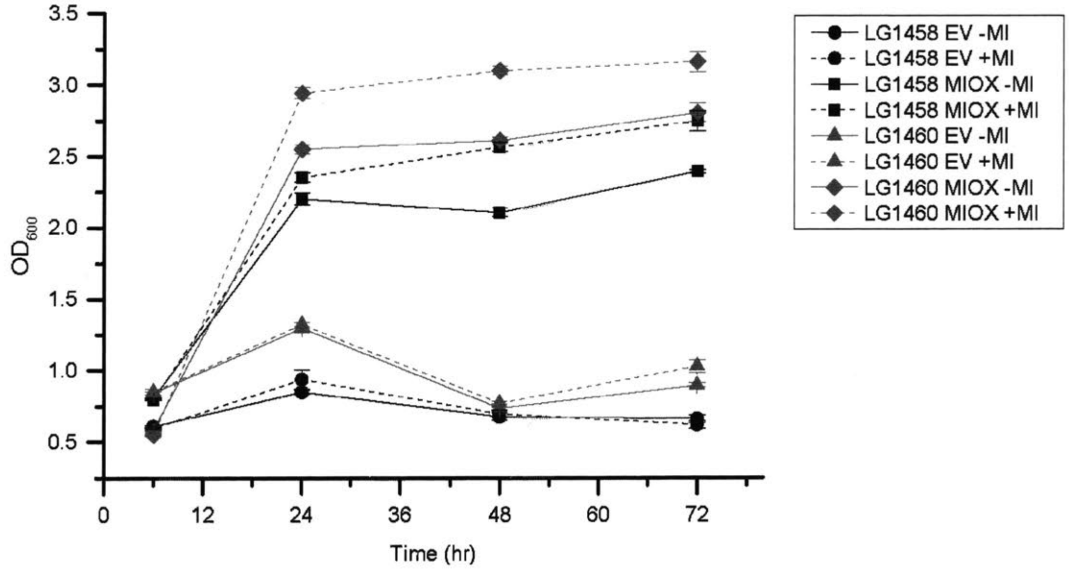

Figure 2.3. Effect of *Miox* expression on biomass as measured by $OD_{600}$. Strains LG1458 and LG1460 harboring pRSFDuet-1 ("EV") or pRSFD-SUMO-MIOX ("MIOX") were grown in LB with or without 60 mM MI ("-MI" and "+MI," respectively). Optical density was measured at 600 nm ($OD_{600}$), and mean values ± SD for triplicate samples are shown.

### 2.3.2. Overexpression of catalase

Given the higher hydrogen peroxide levels observed at late times in fermentations with *Miox* expression and substrate conversion, we then considered the impact of catalase overexpression *in vivo* on the production of glucuronic acid from MI. *E. coli katE* (KatE) and a catalytically inactive version *katE(H128A)* (KatE Mut; Obinger, Maj, Nicholls, & Loewen, 1997) were compared. The strains with KatE had significantly increased glucuronic acid titers (Figure 2.4a). While LG1458 produced higher absolute titers under all conditions, increasing KatE levels resulted in similar overall titer improvements relative to the control in both strains (1.9-fold in LG1458 and 1.8-fold in LG1460). Surprisingly, overexpression of a catalytically inactive mutant *katE(H128A)* (KatE Mut) also improved titers, though the effect was larger in LG1458 than in LG1460 (1.6-fold vs. 1.04-fold). These titer enhancements increased over the course of the fermentation for 1458 KatE, 1458 KatE Mut, and 1460 KatE.

These improvements in glucuronic acid titer were accompanied by similar enhancements in MIOX activity (Figure 2.4b). MIOX activity decreased over the course of the fermentation but was higher in LG1460 under all conditions. MIOX activity was higher when *katE* was overexpressed compared to the control in both strains at all time points, and the effect was largest at 72 hours (10.8-fold for LG1458 and 3.8-fold for LG1460). LG1460 KatE retained an impressive 53% of its 24 hour activity at the end of the fermentation, while LG1458 KatE retained 12%. LG1460 KatE also showed increased soluble protein levels of MIOX at 72 hours relative to the control and KatE Mut (Figure 2.5). KatE Mut had higher activity than the control in LG1458 (up to 3.1-fold at 72 hours), but had a negative or neutral effect in LG1460 after 24 hours.

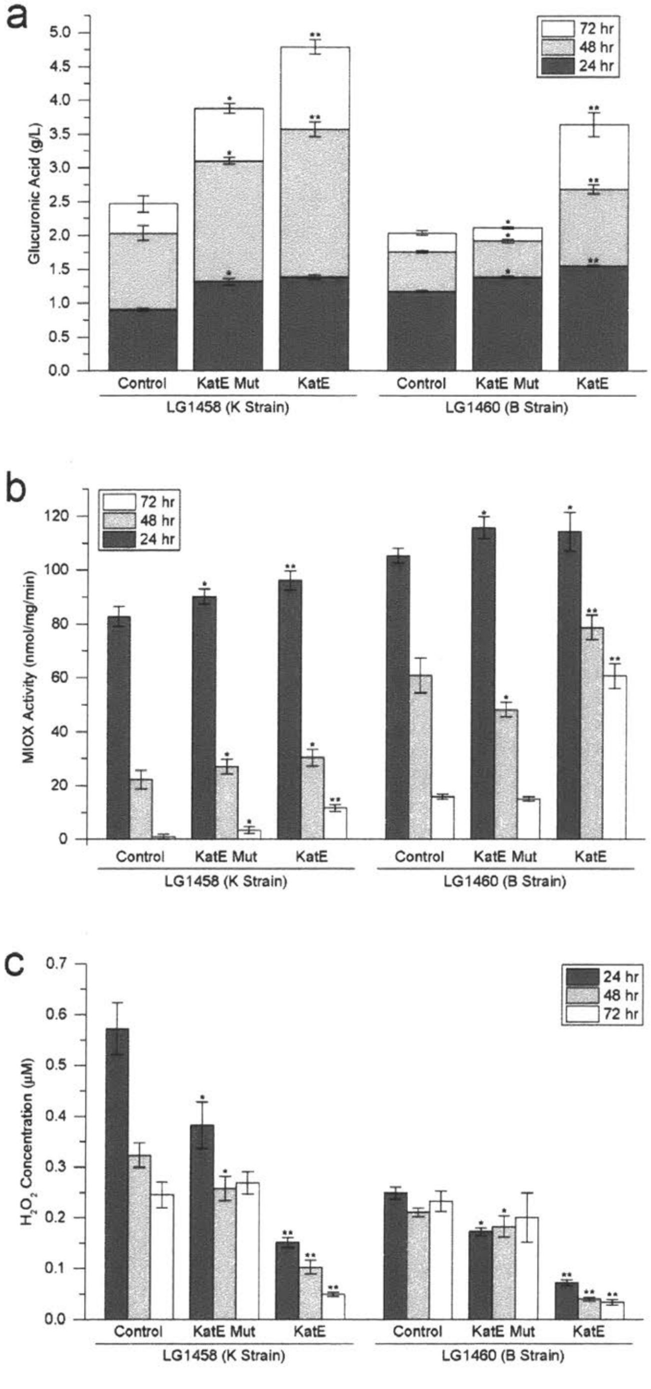

Figure 2.4. *katE* overexpression improves one-step conversion of MI to glucuronic acid. Strains LG1458 and LG1460 harboring pRSFD-SUMO-MIOX ("Control"), pRSFD-SUMO-MIOX-katE(H128A) ("KatE Mut"), or pRSFD-SUMO-MIOX-katE ("KatE") were grown in LB with 60 mM MI. Mean values ± SD for quintuplicate samples are shown, and p-values were calculated for unpaired two-tailed student's t-tests with unequal variance. In all cases, * denotes $p < 0.05$ relative to Control for the same strain and time point, and ** denotes $p < 0.05$ relative to both KatE Mut and Control. (a) Glucuronic acid titers. (b) MIOX activity in crude cell lysates. (c) Hydrogen peroxide concentrations in the supernatant.

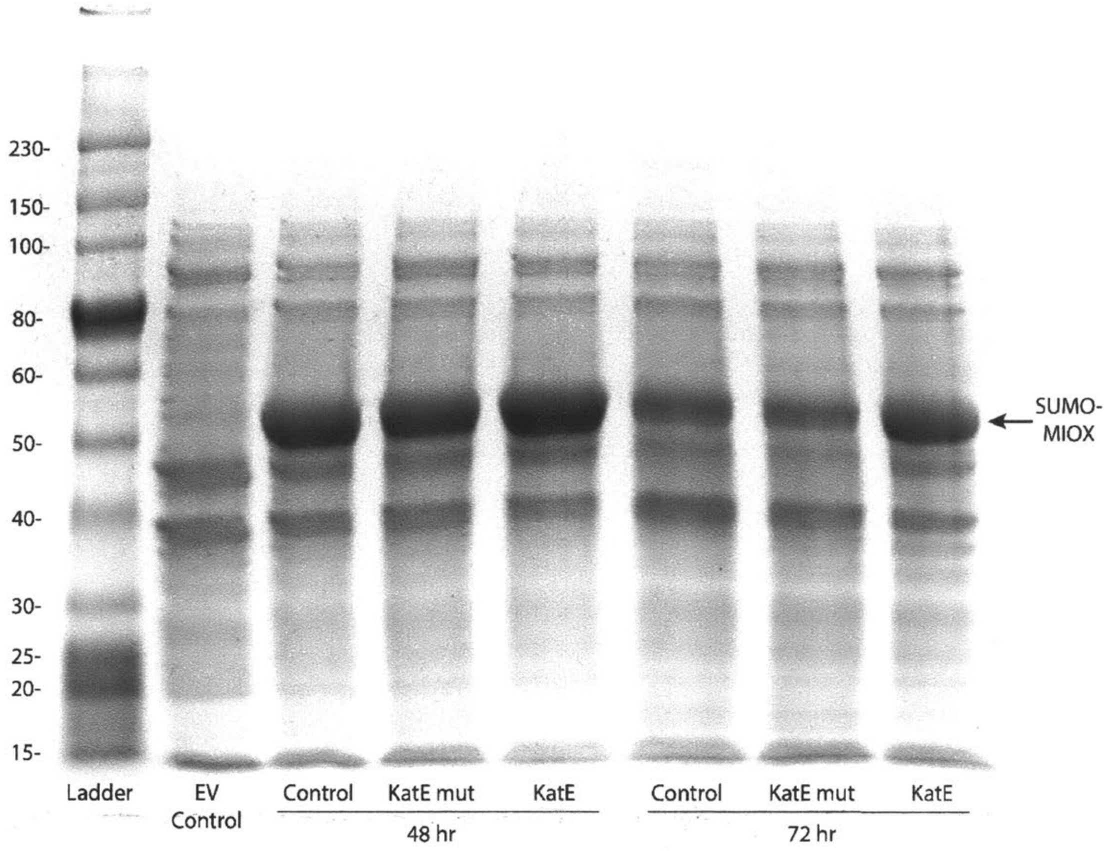

Figure 2.5. *katE* overexpression increases MIOX soluble expression. Strain LG1460 harboring pRSFD-SUMO-MIOX ("Control"), pRSFD-SUMO-MIOX-katE(H128A) ("KatE Mut"), or pRSFD-SUMO-MIOX-katE ("KatE"), as well as LG1460 harboring pRSFDuet-1 and pETDuet-1 ("EV Control"), were grown in LB with 60 mM MI. Crude lysates for one sample each of the Control, KatE Mut, and KatE strains at 48 and 72 hours and one sample of the EV Control strain at 24 hours were run on an SDS-PAGE gel. The band corresponding to SUMO-MIOX (46 kDa) is indicated by the arrow. The band corresponding to KatE (84 kDa) was not easily distinguishable.

To verify that overexpression of catalase reduced hydrogen peroxide in the system, we measured hydrogen peroxide levels in the supernatant. LG1458 had higher hydrogen peroxide levels that generally fell over the course of the fermentation, while LG1460 had lower levels that were more stable with time (Figure 2.4c). Overexpression of *katE* dramatically reduced hydrogen peroxide for both strains and at all time points compared to the control (68-85% reduction), with the relative effect generally increasing with time. Interestingly, KatE Mut also reduced hydrogen peroxide concentrations at 24 and 48 hours, though the effect was smaller (13-

33% reduction) and diminished with time. These effects were accompanied by modest increases in stationary phase $OD_{600}$ for KatE and KatE Mut (Figure 2.6).

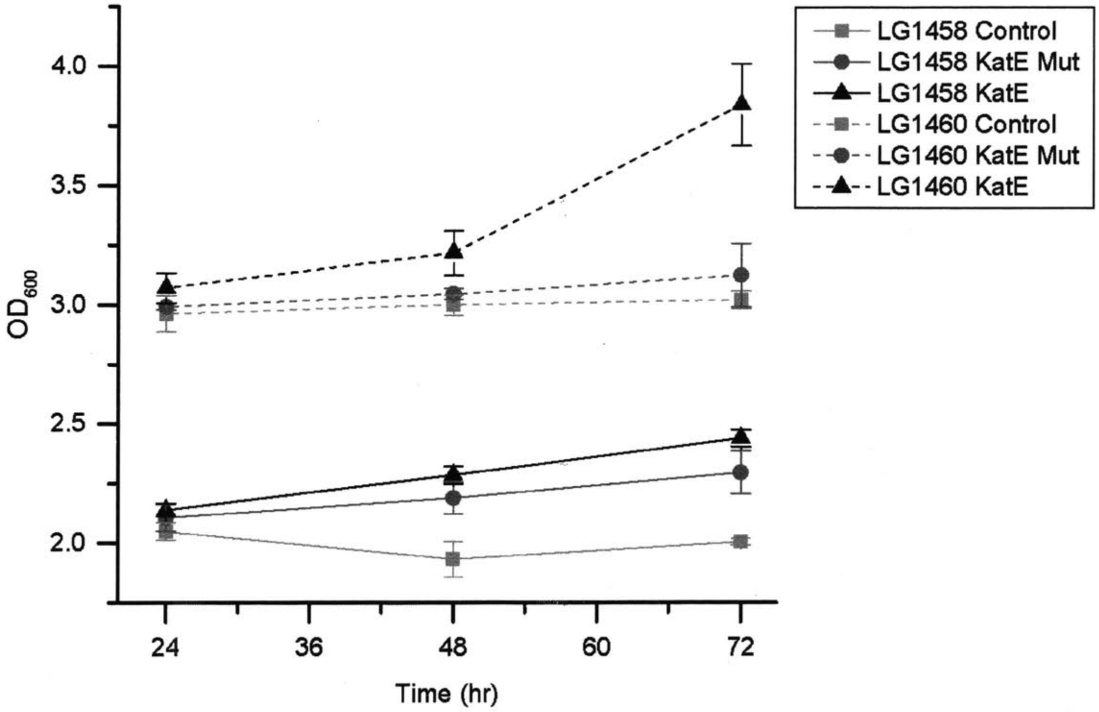

Figure 2.6. *katE* overexpression increases biomass as measured by $OD_{600}$. Strains LG1458 and LG1460 harboring pRSFD-SUMO-MIOX ("Control"), pRSFD-SUMO-MIOX-katE(H128A) ("KatE Mut"), or pRSFD-SUMO-MIOX-katE ("KatE") were grown in LB with 60 mM MI. Optical density was measured at 600 nm ($OD_{600}$), and mean values ± SD for quintuplicate samples are shown.

We also tested the effect of KatE in the context of the full glucaric acid pathway (Figure 2.7). Titers are significantly lower from glucose than from MI because MIPS competes with central carbon metabolism for glucose-6-phosphate, limiting substrate availability to MIOX. However, when expressed from low-copy pACYCDuet plasmids, *katE* and *katE(H128A)* corresponded to small but significant increases in glucaric acid titers. The negative effect of expression from high-copy pETDuet plasmids suggests a tradeoff between hydrogen peroxide scavenging and metabolic burden associated with gene expression.

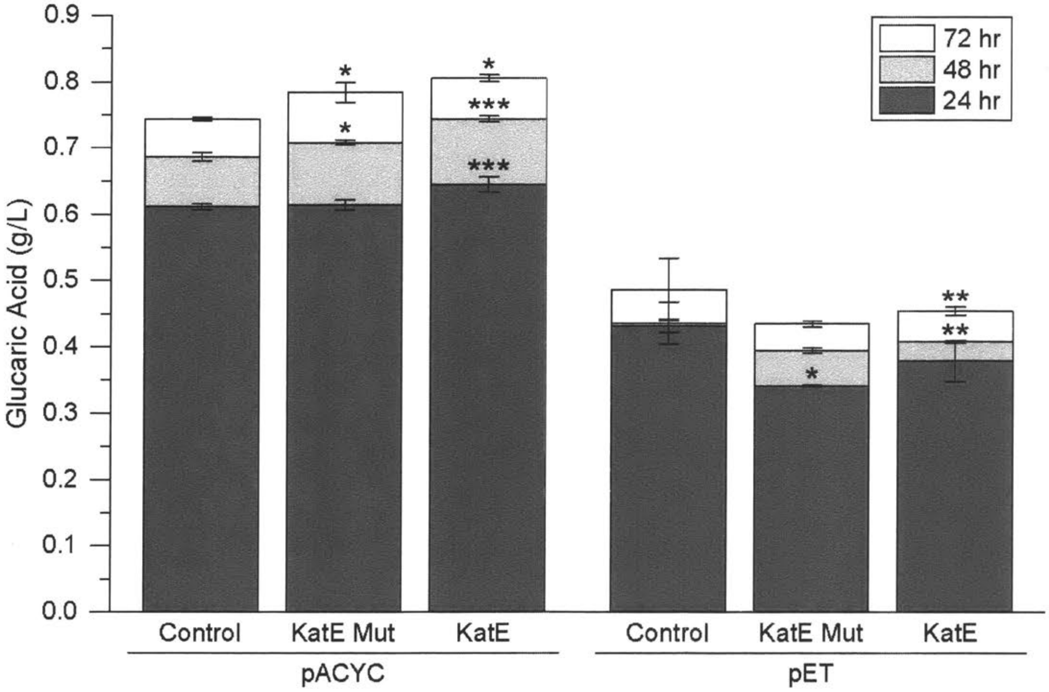

Figure 2.7. Effect of *katE* overexpression on glucaric acid production. Strain LG1460 harboring pACYCDuet-1 or pETDuet-1 ("Control"), pACYC-katE(H128A) or pET-katE(H128A) ("KatE Mut"), or pACYC-katE or pET-katE ("KatE") were grown in LB with 60 mM MI. Mean values ± SD for triplicate samples are shown. P-values were calculated for unpaired two-tailed student's t-tests with unequal variance, and * denotes $p < 0.05$ relative to Control for the same plasmid backbone and time point, ** denotes $p < 0.05$ relative to KatE Mut, and *** denotes $p < 0.05$ relative to both Control and KatE Mut.

### 2.3.3. Overexpression of SODs

While we did not observe an increase in MIOX activity with exogenous SOD addition, we still proceeded to evaluate the impact of overexpression of *sodA* and *sodB in vivo* in LG1460. The MIOX mechanism likely includes both superoxo and hydroperoxo intermediates,[62] and negative effects of superoxide could be present in the cell without particular damage to MIOX itself. SodA and SodB are both cytoplasmic SODs, but SodA employs a manganese cofactor whereas SodB uses an iron cofactor.[112] Expression of catalytically inactive versions of each gene, *sodA(Q147E)* and *sodB(Q70E)*,[113,114] was also included. Consistent with our previous findings, the presence of KatE substantially improved titers (Figure 2.8). While the effect of SOD was less pronounced than that of catalase, all strains overexpressing either *sodA* or *sodB* outperformed the empty vector control. Moreover, strains overexpressing *sodA* or *sodB* outperformed their counterparts expressing *sodA(Q147E)* or *sodB(Q70E)*, though the effect was more pronounced for *sodA*. However, overexpression of both SodB and KatE resulted in the highest titers of all cases tested, achieving a 2.6-fold increase over the control as well as a 7% increase over KatE alone. As before, titer improvements generally increased over the course of the fermentation.

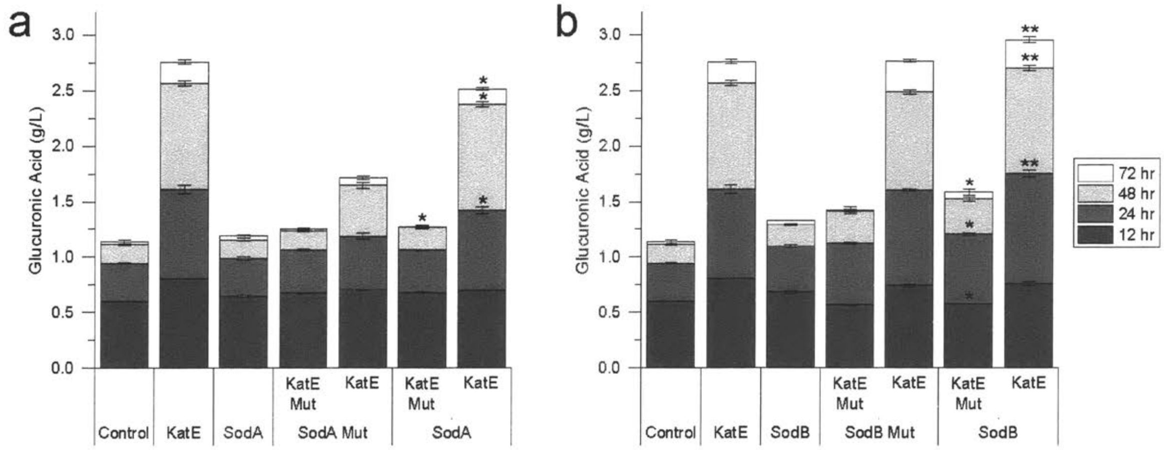

Figure 2.8. SOD overexpression further improves glucuronic acid production. Strain LG1460 harboring pRSFD-SUMO-MIOX and pETDuet-1 or a derivative thereof was grown in LB with 60 mM MI. Glucuronic acid titers were measured, and mean values ± SD for triplicate samples are shown. P-values were calculated for unpaired two-tailed student's t-tests with unequal variance. (a) Glucuronic acid titers for *sodA* overexpression, both with and without *katE* overexpression. "Control" refers to pETDuet-1, "KatE" to pET-katE, "SodA" to pET-sodA, "KatE Mut SodA Mut" to pET-sodA(Q147E)-katE(H128A), "KatE SodA Mut" to pET-sodA(Q147E)-katE, "KatE Mut SodA" to pET-sodA-katE(H128A), and "KatE SodA" to pET-sodA-katE. All strains at all time points yielded higher titers than the control ($p < 0.05$), and strains with plasmids containing both *sodA* and *katE* genes, whether active or mutant, performed better than that with *sodA* alone ($p < 0.05$). In addition, strains with plasmids containing *katE* performed better than their counterparts with *katE(H128A)* ($p < 0.01$). On the plot, * denotes $p < 0.05$ for comparisons of *sodA* to *sodA(Q147E)* (SodA KatE Mut vs. SodA Mut KatE Mut and SodA KatE vs. SodA Mut KatE) at the same time point. (b) Glucuronic acid titers for *sodB* overexpression, both with and without *katE* overexpression. "Control" refers to pETDuet-1, "KatE" to pET-katE, "SodB" to pET-sodB, "KatE Mut SodB Mut" to pET-sodB(Q70E)-katE(H128A), "KatE SodB Mut" to pET-sodB(Q70E)-katE, "KatE Mut SodB" to pET-sodB-katE(H128A), and "KatE SodB" to pET-sodB-katE. All strains at time points later than 12 hours yielded higher titers than the control ($p < 0.005$), and strains with plasmids containing both *sodB* and *katE* genes, whether active or mutant, performed better than that with *sodB* alone ($p < 0.05$). In addition, strains with plasmids containing *katE* performed better than their counterparts with *katE(H128A)* ($p < 0.005$). On the plot, * denotes $p < 0.05$ for comparisons of *sodB* to *sodB(Q70E)* (SodB KatE Mut vs. SodB Mut KatE Mut and SodB KatE vs. SodB Mut KatE) at the same time point, and ** indicates titers above that of the strain containing *katE* alone with $p < 0.01$.

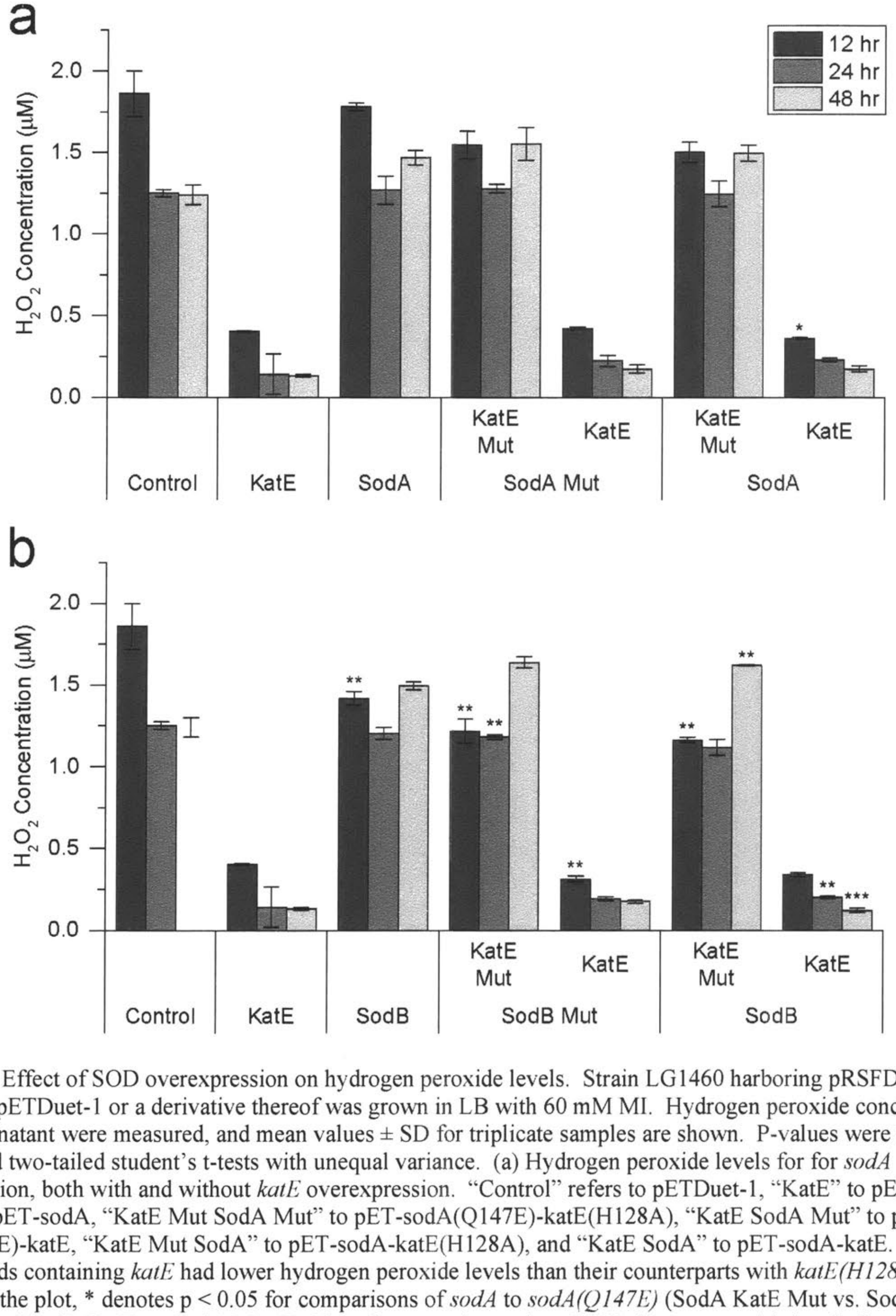

Figure 2.9. Effect of SOD overexpression on hydrogen peroxide levels. Strain LG1460 harboring pRSFD-SUMO-MIOX and pETDuet-1 or a derivative thereof was grown in LB with 60 mM MI. Hydrogen peroxide concentrations in the supernatant were measured, and mean values ± SD for triplicate samples are shown. P-values were calculated for unpaired two-tailed student's t-tests with unequal variance. (a) Hydrogen peroxide levels for for *sodA* overexpression, both with and without *katE* overexpression. "Control" refers to pETDuet-1, "KatE" to pET-katE, "SodA" to pET-sodA, "KatE Mut SodA Mut" to pET-sodA(Q147E)-katE(H128A), "KatE SodA Mut" to pET-sodA(Q147E)-katE, "KatE Mut SodA" to pET-sodA-katE(H128A), and "KatE SodA" to pET-sodA-katE. Strains with plasmids containing *katE* had lower hydrogen peroxide levels than their counterparts with *katE(H128A)* ($p < 0.005$). On the plot, * denotes $p < 0.05$ for comparisons of *sodA* to *sodA(Q147E)* (SodA KatE Mut vs. SodA Mut KatE Mut and SodA KatE vs. SodA Mut KatE) at the same time point. (b) Hydrogen peroxide levels for *sodB* overexpression, both with and without *katE* overexpression. "Control" refers to pETDuet-1, "KatE" to pET-katE, "SodB" to pET-sodB, "KatE Mut SodB Mut" to pET-sodB(Q70E)-katE(H128A), "KatE SodB Mut" to pET-sodB(Q70E)-katE, "KatE Mut SodB" to pET-sodB-katE(H128A), and "KatE SodB" to pET-sodB-katE. Strains with plasmids containing *katE* performed better than their counterparts with *katE(H128A)* ($p < 0.005$). On the plot, * denotes $p < 0.05$ for comparisons of *sodB* to *sodB(Q70E)* (SodB KatE Mut vs. SodB Mut KatE Mut and SodB KatE vs. SodB Mut KatE) at the same time point, ** indicates $p < 0.05$ for comparisons of *sodB* or *sodB(Q70E)* to *sodA* or *sodA(Q147E)*, respectively, and *** denotes $p < 0.05$ for both comparisons.

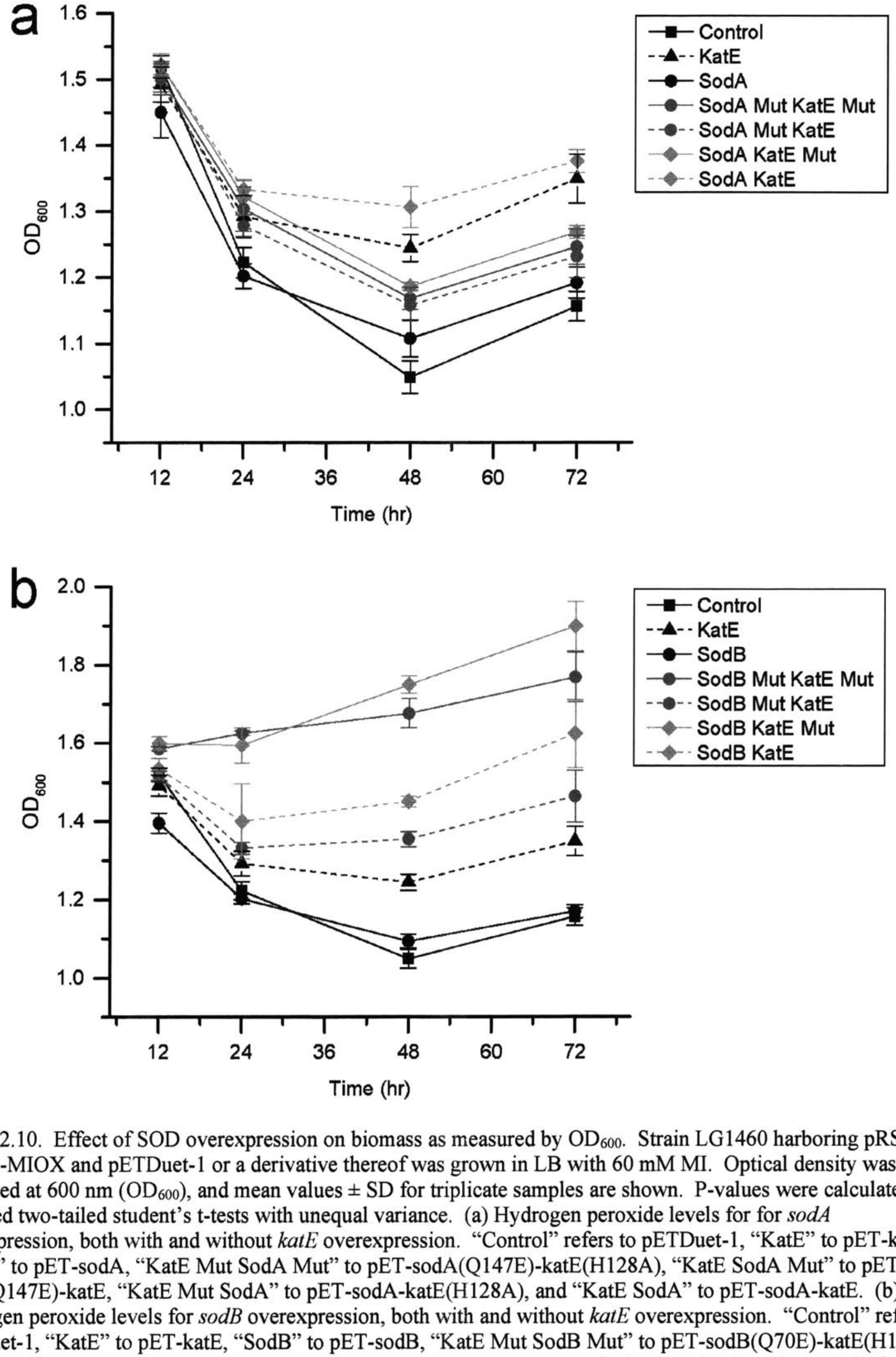

Figure 2.10. Effect of SOD overexpression on biomass as measured by $OD_{600}$. Strain LG1460 harboring pRSFD-SUMO-MIOX and pETDuet-1 or a derivative thereof was grown in LB with 60 mM MI. Optical density was measured at 600 nm ($OD_{600}$), and mean values ± SD for triplicate samples are shown. P-values were calculated for unpaired two-tailed student's t-tests with unequal variance. (a) Hydrogen peroxide levels for for *sodA* overexpression, both with and without *katE* overexpression. "Control" refers to pETDuet-1, "KatE" to pET-katE, "SodA" to pET-sodA, "KatE Mut SodA Mut" to pET-sodA(Q147E)-katE(H128A), "KatE SodA Mut" to pET-sodA(Q147E)-katE, "KatE Mut SodA" to pET-sodA-katE(H128A), and "KatE SodA" to pET-sodA-katE. (b) Hydrogen peroxide levels for *sodB* overexpression, both with and without *katE* overexpression. "Control" refers to pETDuet-1, "KatE" to pET-katE, "SodB" to pET-sodB, "KatE Mut SodB Mut" to pET-sodB(Q70E)-katE(H128A), "KatE SodB Mut" to pET-sodB(Q70E)-katE, "KatE Mut SodB" to pET-sodB-katE(H128A), and "KatE SodB" to pET-sodB-katE.

Hydrogen peroxide levels were also measured for these cultures (Figure 2.9). While increasing KatE again substantially reduced hydrogen peroxide levels, the effects of the SODs were comparatively small. Cultures without overexpressed *katE* had higher hydrogen peroxide levels than the control at 48 and 72 hours, and SodB and SodB Mut samples had slightly higher levels than SodA and SodA Mut ones at the end of the fermentation. Higher hydrogen peroxide levels generally corresponded to higher cell densities (Figure 2.10), but the differences noted above largely persisted in normalized data (not shown). We also attempted to measure MIOX activity, but activities were low and became undetectable by 48 hours (data not shown). Overall, SodB improved titers more than SodA, and SodB Mut KatE performed markedly better than SodA Mut KatE.

### 2.3.4. Addition of iron chelators

The improved performance seen with both KatE Mut and SodB Mut led us to suspect that labile iron levels may also be important in our system. Both mutant enzymes have been shown to retain their bound iron cofactor.[111,114] To test this hypothesis, chemical iron chelator supplementation was used to assess the effect of reducing labile iron levels on glucuronic acid titers. We considered four chelators with different cell permeability and metal binding selectivity characteristics: one cell-permeable and favoring $Fe^{3+}$ binding (deferoxamine),[115] two cell-permeable and favoring $Fe^{2+}$ binding (2,2'-bipyridine and 1,10-phenanthroline),[78,116,117] and one cell-impermeable (DTPA).[118] We determined appropriate concentration ranges for each chelator by serial dilution until growth impairment was no longer evident. We observed a significant increase in glucuronic acid titers for deferoxamine and 1,10-phenanthroline, and this benefit was most pronounced later in the fermentation (Figure 2.11). Addition of deferoxamine resulted in the largest titer increases at the end of the fermentation, but it also significantly decreased titers at 12 and 24 hours.

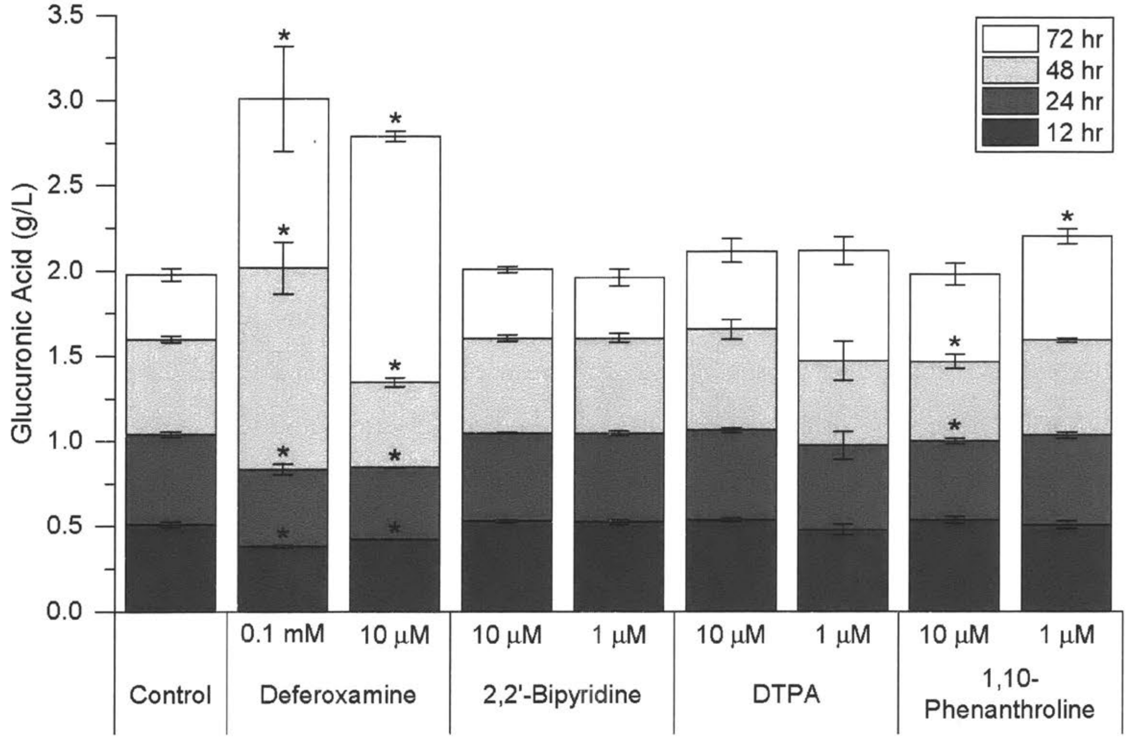

Figure 2.11. Iron chelator supplementation improves glucuronic acid titers. Strain LG1460 harboring pRSFD-SUMO-MIOX was grown in LB with 60 mM MI and supplemented with iron chelators at the indicated concentrations. Glucuronic acid titers were measured, and mean values ± SD for triplicate samples are shown. P-values were calculated for unpaired two-tailed student's t-tests with unequal variance, and * denotes $p < 0.05$ relative to the no chelator control at the same time point.

## 2.4. Discussion

### 2.4.1. ROS measurement methods

Studying ROS in cells is challenging because many detection and diagnostic methods are nonspecific and subject to interference by unrelated phenomena.[119–121] Here, we employed overexpression of enzymes specific to particular reactive oxygen species and selective measurement of hydrogen peroxide via horseradish peroxidase and Amplex red. Direct measurement of ROS levels was restricted to measurement of extracellular hydrogen peroxide due to limitations of ROS probes. Unlike superoxide and hydroxyl radicals, hydrogen peroxide is relatively stable in culture media, and elevated intracellular concentrations that exceed a cell's scavenging capacity are reflected in elevated extracellular concentrations.[121–124]

### 2.4.2. Selection of scavenging strategies

The addition of purified catalase improved *in vitro* MIOX activity of crude lysates, consistent with reports that hydrogen peroxide inhibits the enzyme[66,67] and suggesting that hydrogen peroxide levels present in the system may affect performance. The lower initial supernatant hydrogen peroxide levels observed in the presence of overexpressed *Miox* could be a result of induction of *E. coli* ROS scavenging systems, similar to the induction seen in rice.[70] However, this potential scavenging appears to be less effective at late times in the fermentation, particularly in LG1460. It is also notable that the measured hydrogen peroxide levels are so high. Growth defects in *E. coli* are evident at hydrogen peroxide concentrations of 0.4 µM,[86] and we measured a maximum concentration of 4.1 µM for EV samples and 2.9 µM for MIOX samples. While the measured supernatant concentrations are not necessarily equivalent to intracellular concentrations, the hydrogen peroxide levels observed both in the presence and absence of MIOX are clearly a potential cause for concern.

In selecting strategies to improve ROS scavenging capacity in *E. coli*, we focused on catalases and SODs to directly address elevated levels of hydrogen peroxide and superoxide. We also hoped that scavenging these two species would help reduce the formation of the especially damaging hydroxyl radical by limiting the Fenton reaction and the Haber Weiss cycle. Catalases are efficient scavengers of hydrogen peroxide at high concentrations and are thus well-suited to supplement native antioxidant capacity. While cells also use several other hydrogen peroxide-specific scavenging enzymes and systems, they ultimately require reducing power, consuming additional cellular resources and potentially upsetting redox balance.[76,91] SODs are the only known enzymes in *E. coli* that scavenge superoxide, and they also do not require reducing power.[125] While antioxidant supplementation of culture media has shown some promise in mitigating oxidative stress, we did not consider that strategy here due to the expense, possible prooxidant rather than antioxidant effects,[102] and likely limited potential benefit for bacteria.[119]

*E. coli* has at least two catalases, encoded by *katE* and *katG*. KatE is a typical monofunctional catalase, while KatG is a bifunctional catalase-peroxidase.[126] Both enzymes have high catalytic efficiencies ($k_{cat}/K_M \sim 10^6$ $M^{-1}$ $s^{-1}$).[76] *katG* is part of the OxyR regulon that is induced under oxidative stress, while *katE* is commonly expressed in stationary phase.[126] We chose to overexpress *katE* to minimize disruption to native metabolism and regulation.

*E. coli* contains at least three SODs, encoded by *sodA*, *sodB*, and *sodC*. SodA and SodB are cytoplasmic, while SodC is periplasmic.[125] SodA and SodB are highly homologous, share the same active site sequence, and have similar kinetics.[127–129] The genes are also differentially regulated. Fur represses *sodA* but activates *sodB* when iron is available, and SoxR upregulates *sodA* in response to redox stress.[119,130] The cytoplasmic SODs *sodA* and *sodB* were both selected for this work because each is likely to impact native regulation differently, and it has been suggested that the two enzymes may not be functionally equivalent.[129] The SODs were expressed both alone and in combination with catalase since the SOD reaction generates hydrogen peroxide.

### 2.4.3. Effect of ROS scavengers on MIOX performance

In our system, catalase overexpression appears to improve production of glucuronic acid largely by helping to maintain soluble expression and activity of MIOX over time. While MIOX activity still decreased over the course of the fermentation, as previously observed,[63] strains overexpressing *katE* produced the largest gains in titers, MIOX soluble protein, and MIOX activity at later time points. These results are consistent with the activity benefit seen from exogenous addition of catalase to crude lysates and the known hydrogen peroxide sensitivity of the enzyme.

Slight differences in behavior were observed between the two strains tested. LG1458 (K strain) produced higher glucuronic acid titers, and LG1460 (B strain) generally reaped more benefit from KatE, as reflected in titer, activity, and biomass data. However, LG1458 was associated with higher hydrogen peroxide concentrations and showed more benefit from KatE Mut. Literature results conflict with respect to strain differences in antioxidant capacity, with some reports suggesting *E. coli* K strains have higher capacity than B strains[131,132] and another showing the opposite.[133] Our results do not fully support either conclusion and instead suggest that scavenging capacity may differ between ROS species.

When catalase overexpression was tested in the context of the full glucaric acid pathway, a small but significant benefit was detected when *katE* was expressed from low-copy pACYCDuet vectors, but no improvement was seen for *katE* expressed from high-copy pETDuet vectors. This suggests that metabolic burden is significant when many genes are overexpressed.

In addition, this result indicates that tuning the expression level of *katE* is likely required to produce optimal results.

The smaller effect of SOD overexpression relative to catalase overexpression suggests that hydrogen peroxide impacts performance more than superoxide. However, SODs do still boost glucuronic acid titers, indicating that their benefit outweighs their protein production cost. Because the SOD reaction produces hydrogen peroxide, we expected that SODs would perform best when catalase was also overexpressed. This is indeed what is observed for SodA. However, SodB performs similarly in both the presence and absence of catalase overexpression. Because there was no observed benefit from exogenous SOD addition for *in vitro* MIOX activity, the small increase in titers from overexpression of SOD *in vivo* may be due to differences in superoxide between the *in vitro* and *in vivo* conditions or to systemic effects of superoxide that do not directly impact MIOX.

Overexpression of genes for catalytically inactive enzymes was intended to help correct for the increased burden associated with protein overexpression, and the significant positive effect of KatE(H128A) and SodB(Q70E) was unexpected. However, while these mutations destroy activity, iron cofactor binding is retained. This suggests that KatE and SodB – both catalytically active and inactive versions – may function to sequester labile iron. The similar boost in glucuronic acid titers observed upon addition of cell-permeable chemical iron chelators deferoxamine and 1,10-phenanthroline confirmed that iron sequestration is effective in the system. Both $Fe^{2+}$-selective and $Fe^{3+}$-selective chelators improved performance. Iron is tightly regulated in living systems because it is essential for life but also has the potential to promote hydroxyl radical formation.[89,134] Both catalase and iron-sequestering proteins are upregulated via the OxyR regulon in response to oxidative stress in *E. coli*,[75] and this native response appears to be insufficient for optimal production of glucuronic acid. However, directly tuning iron levels can also trigger iron starvation, and we indeed see negative effects from deferoxamine early in the fermentation. Further work to optimize labile iron levels may yield valuable tools to reduce damage from hydroxyl radicals.

Oxidative stress has become a common problem in metabolic engineering, and the strategy outlined here for its relief is general and could be applied to other pathways and organisms. Catalases and SODs are efficient enzymes that can be employed to address hydrogen peroxide and superoxide stress from any source. Moreover, the approach is applicable to other organisms. The well-characterized *E. coli* enzymes used here may be suitable for use in other hosts, but homologous scavengers of ROS are also present across all kingdoms of life. Among the many hydrogen peroxide scavenging enzymes, typical catalases like KatE are the most abundant in nature.[126] Similarly, Fe-SODs like SodB are the most abundant scavengers of superoxide.[135]

## 2.5. Conclusions

The performance of MIOX in *E. coli* was shown to be affected by ROS, and a systematic approach was used to alleviate oxidative stress. Catalase and SOD overexpression led to increased biomass, MIOX activity, and glucuronic acid titers. The beneficial effect of ROS scavenging increased with fermentation time and corresponded to maintenance of soluble MIOX expression and activity. Alone, catalase had a larger impact than SODs, but the highest titers were produced when both were overexpressed. The addition of iron chelators and overexpression of iron-binding proteins also improved performance, suggesting labile iron levels contribute to ROS damage. The strategies used here to supplement native ROS scavenging capacity substantially improved glucuronic acid production and are in principle adaptable to a wide range of other metabolic pathways and organisms.

# 3. Leveraging Sequence Networks to Identify Improved MIPS Enzymes

## Abstract

The MIPS enzyme (INO1 in *S. cerevisiae*) appears to limit glucaric acid pathway flux due to its competition with central carbon metabolism for its substrate, glucose-6-phosphate. Many putative MIPS enzymes have been identified, and we aimed to leverage this natural diversity to help identify improved homologs. Thirty-one diverse MIPS enzymes were selected from a sequence similarity network for Pfam family PF01658. Of these 31 sequences, 19 produced detectible MI production when expressed with an N-terminal polyhistidine tag. One homolog, *H. contortus* (Hc31) MIPS, performed as well as or better than INO1 under most experimental conditions. Several eukaryotic and prokaryotic enzymes also had significantly higher activity than INO1. However, stable enzyme expression and thermostability appears to be a challenge. While statistical power to determine important sequence features was limited because of the small number of experimentally validated sequences, this work provides guidance for further exploration of the MIPS network.

### 3.1. Introduction

Making bioprocesses economically competitive often requires developing better enzymes to catalyze reactions of interest. Directed evolution is a powerful and well-utilized tool in metabolic engineering to improve an enzyme from a template sequence. However, sequence spaces are vast and can be difficult to navigate, even by directed evolution.[44] Utilizing natural protein diversity can provide an alternative or complementary approach. Nature has already produced many sequences that can perform the same reactions, and using this information wisely can reduce the screening effort and allow for more exploration in sequence space.

However, extracting useful information from protein databanks is challenging. As discussed in Section 1.3.1, databases are growing exponentially, and the vast majority of sequences have not been functionally validated. In addition, deposited protein sequences may contain errors from sequencing or miscalled introns, and automatic protein classification algorithms are imperfect. Moreover, making effective use of large or diverse sets of proteins is difficult without testing large numbers of sequences. An inherent tradeoff exists between leveraging sequence diversity and obtaining useful structural information. Proteins that are diverse have many amino acid differences, which makes pinpointing particular structure-function relationships challenging.

While most engineering work in the glucaric pathway to date has focused on the MIOX enzyme due to its low activity and stability, MIPS also appears to limit pathway performance. MIOX has already been the subject of a directed evolution study,[63] and a bioprospecting effort to identify improved homologs is ongoing.[136] However, there has been comparatively little focus on MIPS, which also has relatively low activity in the pathway.[59] MIPS competes with glycolysis and the pentose phosphate pathway for its substrate, glucose-6-phosphate, and elimination of flux to the pentose phosphate pathway and dynamic downregulation of glycolysis improves glucaric acid production.[14,74,137] Improvement of MIPS may thus allow it to better compete with endogenous pathways.

MIPS is not naturally present in *E. coli*, but it is widely conserved throughout all branches of life.[138] Its mechanism and key catalytic residues have been well-studied [139,140], a few homologs have been crystallized,[139,141–143] and conserved sequence stretches have been identified.[144,145] Eukaryotic sequences are relatively similar, while prokaryotic and archaeal sequences have significantly more variability.[144] In addition, eukaryotic sequences are longer than their prokaryotic and archaeal counterparts, though the lengths of these sequence insertions vary, and their function is not well known.[140] In general, while there has been interest in the phylogeny of MIPS enzymes, the functional differences between homologs are still not well understood.

Sequence similarity networks (SSNs) are a relatively new tool that display pairwise alignments (edges) between sequences (nodes), grouping more similar sequences into clusters. A simple example SSN is shown in Figure 3.1. Increasing the stringency of the threshold value applied to pairwise alignments prunes edges in the network, breaking apart existing clusters into subclusters containing more similar sequences. SSNs reduce sequence information to an intuitive two-dimensional format and allow orthogonal information to be overlaid on the network through node and edge properties.[146] In addition, they are faster to generate, can accommodate larger sets of sequences, and are easier to visualize than more traditional tools like multiple alignments and phylogenetic trees.[146] SSNs have been used to clarify differences in specificity and function within large superfamilies of proteins.[37,147,148] They have also been used to provide helpful context for identifying the function of unknown proteins and prospecting for new functions.[146]

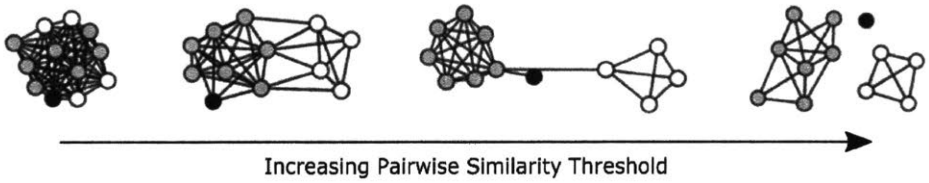

Figure 3.1. Effect of pairwise alignment threshold value on example SSN. As the threshold for similarity increases (i.e. as alignment score increases or the E-value decreases), the network breaks apart into smaller subclusters of more similar sequences. This figure was adapted from Atkinson et al., 2009.

SSNs may also be helpful in identifying improved enzyme homologs for metabolic engineering applications. In this work, we employed SSNs as a primary tool to aid in categorizing and grouping putative MIPS sequences to efficiently explore natural sequence diversity.

## 3.2. Materials and Methods

### 3.2.1. Sequence Similarity Network Generation and Visualization

Sequence similarity networks were generated using the University of Illinois Enzyme Function Initiative's Enzyme Similarity Tool (EFI-EST)[149] using the MIPS Pfam family PF01658, sometimes supplemented with additional user-supplied sequences (see Section 3.3.4). The resulting networks were visualized in Cytoscape.[150]

### 3.2.2. Strains and Plasmids

The *E. coli* strains and plasmids used in this study are listed in Table 3.1. Primers used for construction are listed in Table B.1. *E. coli* strain DH5α was used for molecular cloning and plasmid preparation. The *E. coli* strain used for all MIPS screening was LG1460, constructed as described in Chapter 2.

The plasmids pRSFD-IN, pRSFD-SUMO-MIOX, and pRSFD-IN-opt were previously constructed in the Prather lab [59,63]. The initial set of 31 MIPS genes used in this work were obtained from the Joint Genome Institute (JGI) through the Community Science Program and were codon-optimized for *S. cerevisiae*. These gene sequences are listed in Table B.2.

Plasmids containing these MIPS genes were constructed by circular polymerase extension cloning (CPEC; Quan & Tian, 2009), using primers LG123 and LG124 to amplify the pRSFDuet-1 backbone. The primers used to amplify the MIPS genes are as follows: LG125 and LG126 (*T. maritima*), LG127 and LG128 (*A. fulgidus*), LG129 and LG130 (*M. tuberculosis*), LG131 and LG132 (*A. thaliana*), LG149 and LG150 (*A. clavatus*), LG133 and LG134 (*B. thetaiotaomicron*), LG151 and LG152 (*C. glabrata*), LG153 and LG154 (*C. orthopsilosis*), LG135 and LG136 (*C. halotolerans*), LG155 and LG156 (*D. squalens*), LG137 and LG138 (*D. melanogaster*), LG139 and LG140 (*G. vaginalis*), LG141 and LG142 (*H. sapiens*), LG157 and LG158 (*M. australicum*), LG159 and LG160 (*M. psychrophilus*), LG161 and LG162 (*M. paludis*), LG163 and LG164 (*N. nova*), LG165 and LG166 (*P. ramorum*), LG143 and LG144 (*P. buccae*), LG145 and LG146 (*S. indicum*), LG167 and LG168 (*S. thermophilus*), LG169 and LG170 (*S. cattleya*), LG171 and LG172 (*T. eurythermalis*), LG147 and LG148 (*V. radiata)*, LG173 and LG174 (*Z. bailii*), LG175 and LG176 (*N. maritimus*), LG177 and LG178 (*M. thermautrophicus*), LG179 and LG180 (*T. albus*), LG181 and LG182 (*B. mycoides*), LG183 and LG184 (*Bradyrhizobium sp.*), and LG185 and LG186 (*H. contortus*). The resulting plasmids were named using an abbreviation for the organism from which MIPS originated and a number referring to the order of the MIPS genes in the shipment (ex. pRSFD-Tm1-MIPS).

The original pRSFD-IN and pRSFD-IN-opt plasmids had the INO1 gene out of frame with the polyhistidine (His) tag on the pRSFDuet-1 backbone, so equivalent in-frame versions were created to allow protein purification. pRSFD-His-IN and pRSFD-His-IN-opt were created in the Prather lab by removing the start codon of INO1 to put the protein back in frame with the His tag.

Analogous plasmids containing the MIPS genes in frame with the pRSFDuet-1 His tag were constructed by CPEC. Primers LG123 and LG124 were used to amplify the pRSFD-His-IN backbone. MIPS genes were amplified from the pRSFDuet-derived plasmids described above. The primers used to amplify the MIPS genes are as follows: LG220 and LG126 (*T. maritima*), LG221 and LG128 (*A. fulgidus*), LG222 and LG130 (*M. tuberculosis*), LG223 and LG132 (*A. thaliana*), LG224 and LG150 (*A. clavatus*), LG225 and LG134 (*B. thetaiotaomicron*), LG226 and LG152 (*C. glabrata*), LG227 and LG154 (*C. orthopsilosis*), LG228 and LG136 (*C. halotolerans*), LG229 and LG156 (*D. squalens*), LG230 and LG138 (*D. melanogaster*), LG231 and LG140 (*G. vaginalis*), LG232 and LG142 (*H. sapiens*), LG233 and LG158 (*M. australicum*), LG234 and LG160 (*M. psychrophilus*), LG235 and LG162 (*M. paludis*), LG236 and LG164 (*N. nova*), LG237 and LG166 (*P. ramorum*), LG238 and LG144 (*P. buccae*), LG239 and LG146 (*S. indicum*), LG240 and LG168 (*S. thermophilus*), LG241 and LG170 (*S. cattleya*), LG242 and LG172 (*T. eurythermalis*), LG243 and LG148 (*V. radiata)*, LG244 and LG174 (*Z. bailii*), LG245 and LG176 (*N. maritimus*), LG246 and LG178 (*M. thermautrophicus*), LG247 and LG180 (*T. albus*), LG248 and LG182 (*B. mycoides*), LG249 and LG184 (*Bradyrhizobium sp.*), and LG250 and LG186 (*H. contortus*). These resulting plasmids were named to indicate the presence of the in-frame N-terminal His tag (ex. pRSFD-His-Tm1-MIPS).

N-terminal small ubiquitin-related modifier (SUMO) protein fusions were created for a subset of the MIPS genes by CPEC. Primers LG123 and LG251 were used to amplify the pRSFD-SUMO-MIOX backbone. MIPS genes were amplified from the His-tagged plasmids described above. The primers used to amplify the MIPS genes are as follows: LG253 and LG252 (*INO1*), LG254 and LG252 (*T. maritima*), LG255 and LG252 (*A. thaliana*), LG256 and LG252 (*H. sapiens*), LG257 and LG252 (*S. indicum*), LG258 and LG252 (*Z. bailii*), and LG259 and LG252 (*H. contortus*). The resulting plasmids were named to indicate the presence of the N-terminal SUMO tag (ex. pRSFD-SUMO-Tm1-MIPS).

A subset of the MIPS genes were codon-optimized for *E. coli* using Thermo Fisher's GeneArt GeneOptimizer tool.[152] The genes were designed to include in-frame N-terminal His tags and to avoid the EcoRI and HindIII restriction sites intended for construction. The *A. thaliana*, *B. thetaiotaomicron*, *C. glabrata*, *M. psychrophilus*, *S. indicum*, and *H. contortus* MIPS sequences used are included in Table B.3. The pRSFDuet-1 plasmid and the optimized MIPS gene inserts were each digested with EcoRI and HindIII and ligated. The resulting plasmids were named to indicate the presence of the N-terminal His tag as well as the *E. coli* codon optimization (ex. pRSFD-His-At4-MIPS-opt).

Finally, selected single amino acid mutations were introduced to pRSFD-His-IN, pRSFD-His-At4-MIPS-opt, and pRSFD-His-Pr18-MIPS by amplifying the appropriate plasmid using primers designed using Agilent's QuikChange primer design web tool.[108] The following primers were used with pRSFD-His-IN: LG271 and LG272 to create pRSFD-His-IN(V82M), LG273 and LG274 to create pRSFD-His-IN(A83G), LG279 and LG280 to create pRSFD-His-IN(N141D), LG275 and LG276 to create pRSFD-His-IN(Y250F), and LG277 and LG278 to create pRSFD-His-IN(V413R). With pRSFD-His-At4-MIPS-opt, primers LG281 and LG282 were used to create pRSFD-His-At4-MIPS(A79G)-opt, and primers LG283 and LG284 were used to create pRSFD-His-At4-MIPS(D146N)-opt. Primers LG285 and LG286 were used with pRSFD-His-Pr18-MIPS to create pRSFD-His-Pr18-MIPS(F234Y).

Verification of all Duet vector constructs was performed using primers LG73 and LG74. These were supplemented with LG266 and LG280 for verification of some of the pRSFD-His-IN mutations and with LG206 for verification of the pRSFD-His-Pr18-MIPS mutation.

### 3.2.3. Culture Conditions

Strains were grown in 1 mL of medium in 48-well flower plates (m2p-labs, Baesweiler, Germany) at 30°C or 37°C and 1200 rpm. Strains were grown in Luria-Bertani (LB) media supplemented with glucose (Sigma-Aldrich). Working cultures were inoculated from overnight cultures at a dilution of 1:20, induced with 100 μM isopropyl β-D-1-thiogalactopyranoside (IPTG), and supplemented with kanamycin (50 μg/mL) as required.

### 3.2.4. Measurement of MIPS Activity and Expression Level

For MIPS activity assays, cell pellets were taken from 750 μL of culture media at 48 hr after inoculation, washed twice in Tris-acetate buffer (0.5 M, pH 7.2), and resuspended in 200 μL B-PER (supplied in Tris buffer; Thermo Fisher Scientific, Waltham, MA) supplemented with an EDTA-free protease inhibitor cocktail (Roche Applied Science). Lysates were prepared by shaking at room temperature for 15 min followed by centrifugation, and total soluble protein was measured using a BCA assay kit (Thermo Fisher Scientific). MIPS activity was measured in crude lysates as previously described,[59] by the conversion of glucose-6-phosphate to *myo*-inositol-1-phosphate (MI-1-P) followed by release of inorganic phosphate ($P_i$) using sodium periodate. Activity was corrected using a no substrate control. A no periodate control was also tested, but we found little difference between the periodate and no periodate samples. We subsequently verified that periodate releases $P_i$ from added *myo*-inositol-1-phosphate (Sigma-Aldrich), whereas lysate released little $P_i$ from *myo*-inositol-1-phosphate. This might suggest that there was little MIPS activity, but our lysates produced significant $P_i$ from G6P as compared to controls without lysate. We therefore show relative measured activity for samples with periodate. Activity was normalized by the total protein concentration.

For analysis of MIPS expression levels, 15 μg of total protein for each lysate was separated by SDS-PAGE using a 10% polyacrylamide gel (Bio-Rad Laboratories, Hercules, CA) and transferred onto a nitrocellulose membrane via wet electroblotting. Membranes were blocked overnight in 5% milk at 4°C then incubated at room temperature for 2 hours in a 1:250 dilution of anti-His antibody conjugated to HRP (Santa Cruz Biotechnology, Santa Cruz, CA). Immunodetection was performed using Western Blotting Luminol Reagent (Santa Cruz Biotechnology) according to the manufacturer's instructions. The sum of pixel intensities (volume) for each band was measured and normalized to lane total protein using Bio-Rad's Image Lab software.[153]

### 3.2.5. Measurement of Extracellular Metabolites

Glucose, MI, and acetate concentrations in culture supernatant samples were quantified by high performance liquid chromatography (HPLC) on an Agilent 1200 series instrument (Santa Clara, CA) with an Aminex HPX-87H anion exchange column (300 mm by 7.8 mm; Bio-Rad Laboratories) using 5 mM sulfuric acid at a flow rate 0.6 mL/min as the mobile phase. The column and refractive index detector temperatures were held at 45°C and 35°C, respectively. Compounds were quantified from 10 μL injections using the refractive index signal.

### 3.2.6. MIPS Sequence Analysis

Multiple alignments of MIPS sequences were obtained using PROMALS3D,[154] using PDB structures 3QVS, 1GR0, 1P1I, 3CIN, and 1VKO. A phylogenetic tree was constructed using FastTree[155] and visualized in Archaeopteryx.[156] Python scripts for determining the number of differences in a given set of amino acid residue positions indexed with respect to a particular MIPS sequence and the amino acid identities of those differences from a given multiple alignment are included in Appendix B.2.1 and B.2.2, respectively. The amino acid positions used are listed in Table B.4. These include conserved amino acid positions from literature[144,145] as well as residues near the INO1 active site (PDB 1RM0),[139] as determined using PyMol.[157] For analysis of sequence differences, residues of similar size, hydropathy index, and chemistry were grouped according to their IMGT classes.[158]

Table 3.1. *E. coli* strains and plasmids used in this chapter

| Name | Genotype[a,b] | Source |
|---|---|---|
| **Strains** | | |
| LG1460 | BL21(DE3) ΔuxaC ΔgudD | This study |
| **Plasmids** | | |
| pRSFDuet-1 | RSF1030 *ori*, lacI, Kan$^R$ | Novagen (Darmstadt, Germany) |
| pRSFD-IN | pRSFDuet-1 with *S. cerevisiae INO1* inserted into the EcoRI and HindIII sites | Moon et al, 2009 |
| pRSFD-SUMO-MIOX | pRSFDuet-1 with SUMO-MIOX inserted into the NcoI and HindIII sites | Shiue & Prather, 2014 |
| pRSFD-IN-opt | pRSFDuet-1 with *E. coli* codon-optimized *INO1*[b] inserted into the EcoRI and HindIII sites | Prather Lab |
| pRSFD-His-IN | pRSFDuet-1 with *S. cerevisiae INO1* in frame with His tag | Prather Lab |
| pRSFD-His-IN-opt | pRSFDuet-1 with *E. coli* codon-optimized *INO1* in frame with His tag | Prather Lab |
| pRSFD-Tm1-MIPS | pRSFDuet-1 with *T. maritima MIPS*[a] | This study |
| pRSFD-Af2-MIPS | pRSFDuet-1 with *A. fulgidus MIPS*[a] | This study |
| pRSFD-Mtu3-MIPS | pRSFDuet-1 with *M. tuberculosis MIPS*[a] | This study |
| pRSFD-At4-MIPS | pRSFDuet-1 with *A. thaliana MIPS*[a] | This study |
| pRSFD-Ac5-MIPS | pRSFDuet-1 with *A. clavatus MIPS*[a] | This study |
| pRSFD-Bt6-MIPS | pRSFDuet-1 with *B. thetaiotaomicron MIPS*[a] | This study |
| pRSFD-Cg7-MIPS | pRSFDuet-1 with *C. glabrata MIPS*[a] | This study |
| pRSFD-Co8-MIPS | pRSFDuet-1 with *C. orthopsilosis MIPS*[a] | This study |
| pRSFD-Ch9-MIPS | pRSFDuet-1 with *C. halotolerans MIPS*[a] | This study |
| pRSFD-Ds10-MIPS | pRSFDuet-1 with *D. squalens MIPS*[a] | This study |
| pRSFD-Dm11-MIPS | pRSFDuet-1 with *D. melanogaster MIPS*[a] | This study |
| pRSFD-Gv12-MIPS | pRSFDuet-1 with *G. vaginalis MIPS*[a] | This study |

Table 3.1. *E. coli* strains and plasmids used in this chapter (cont.)

| Name | Genotype[a,b] | Source |
|---|---|---|
| **Plasmids** | | |
| pRSFD-Hs13-MIPS | pRSFDuet-1 with *H. sapiens MIPS*[a] | This study |
| pRSFD-Ma14-MIPS | pRSFDuet-1 with *M. australicum MIPS*[a] | This study |
| pRSFD-Mps15-MIPS | pRSFDuet-1 with *M. psychrophilus MIPS*[a] | This study |
| pRSFD-Mpa16-MIPS | pRSFDuet-1 with *M. paludis MIPS*[a] | This study |
| pRSFD-Nn17-MIPS | pRSFDuet-1 with *N. nova MIPS*[a] | This study |
| pRSFD-Pr18-MIPS | pRSFDuet-1 with *P. ramorum MIPS*[a] | This study |
| pRSFD-Pb19-MIPS | pRSFDuet-1 with *P. buccae MIPS*[a] | This study |
| pRSFD-Si20-MIPS | pRSFDuet-1 with *S. indicum MIPS*[a] | This study |
| pRSFD-St21-MIPS | pRSFDuet-1 with *S. thermophilus MIPS*[a] | This study |
| pRSFD-Sc22-MIPS | pRSFDuet-1 with *S. cattleya MIPS*[a] | This study |
| pRSFD-Te23-MIPS | pRSFDuet-1 with *T. eurythermalis MIPS*[a] | This study |
| pRSFD-Vr24-MIPS | pRSFDuet-1 with *V. radiata MIPS*[a] | This study |
| pRSFD-Zb25-MIPS | pRSFDuet-1 with *Z. bailii MIPS*[a] | This study |
| pRSFD-Nm26-MIPS | pRSFDuet-1 with *N. maritimus MIPS*[a] | This study |
| pRSFD-Mth27-MIPS | pRSFDuet-1 with *M. thermautrophicus MIPS*[a] | This study |
| pRSFD-Ta28-MIPS | pRSFDuet-1 with *T. albus MIPS*[a] | This study |
| pRSFD-Bm29-MIPS | pRSFDuet-1 with *B. mycoides MIPS*[a] | This study |
| pRSFD-B30-MIPS | pRSFDuet-1 with *Bradyrhizobium sp. MIPS*[a] | This study |
| pRSFD-Hc31-MIPS | pRSFDuet-1 with *H. contortus MIPS*[a] | This study |
| pRSFD-His-Tm1-MIPS | pRSFDuet-1 with *T. maritima MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Af2-MIPS | pRSFDuet-1 with *A. fulgidus MIPS*[a] in frame with His tag | This study |

Table 3.1. *E. coli* strains and plasmids used in this chapter (cont.)

| Name | Genotype[a,b] | Source |
|---|---|---|
| **Plasmids** | | |
| pRSFD-His-Mtu3-MIPS | pRSFDuet-1 with *M. tuberculosis MIPS*[a] in frame with His tag | This study |
| pRSFD-His-At4-MIPS | pRSFDuet-1 with *A. thaliana MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Ac5-MIPS | pRSFDuet-1 with *A. clavatus MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Bt6-MIPS | pRSFDuet-1 with *B. thetaiotaomicron MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Cg7-MIPS | pRSFDuet-1 with *C. glabrata MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Co8-MIPS | pRSFDuet-1 with *C. orthopsilosis MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Ch9-MIPS | pRSFDuet-1 with *C. halotolerans MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Ds10-MIPS | pRSFDuet-1 with *D. squalens MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Dm11-MIPS | pRSFDuet-1 with *D. melanogaster MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Gv12-MIPS | pRSFDuet-1 with *G. vaginalis MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Hs13-MIPS | pRSFDuet-1 with *H. sapiens MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Ma14-MIPS | pRSFDuet-1 with *M. australicum MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Mps15-MIPS | pRSFDuet-1 with *M. psychrophilus MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Mpa16-MIPS | pRSFDuet-1 with *M. paludis MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Nn17-MIPS | pRSFDuet-1 with *N. nova MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Pr18-MIPS | pRSFDuet-1 with *P. ramorum MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Pb19-MIPS | pRSFDuet-1 with *P. buccae MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Si20-MIPS | pRSFDuet-1 with *S. indicum MIPS*[a] in frame with His tag | This study |
| pRSFD-His-St21-MIPS | pRSFDuet-1 with *S. thermophilus MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Sc22-MIPS | pRSFDuet-1 with *S. cattleya MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Te23-MIPS | pRSFDuet-1 with *T. eurythermalis MIPS*[a] in frame with His tag | This study |

Table 3.1. *E. coli* strains and plasmids used in this chapter (cont.)

| Name | Genotype[a,b] | Source |
|---|---|---|
| **Plasmids** | | |
| pRSFD-His-Vr24-MIPS | pRSFDuet-1 with *V. radiata MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Zb25-MIPS | pRSFDuet-1 with *Z. bailii MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Nm26-MIPS | pRSFDuet-1 with *N. maritimus MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Mth27-MIPS | pRSFDuet-1 with *M. thermautrophicus MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Ta28-MIPS | pRSFDuet-1 with *T. albus MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Ba29-MIPS | pRSFDuet-1 with *B. mycoides MIPS*[a] in frame with His tag | This study |
| pRSFD-His-B30-MIPS | pRSFDuet-1 with *Bradyrhizobium sp. MIPS*[a] in frame with His tag | This study |
| pRSFD-His-Hc31-MIPS | pRSFDuet-1 with *H. contortus MIPS*[a] in frame with His tag | This study |
| pRSFD-SUMO-IN | pRSFDuet-1 with SUMO-INO1 | This study |
| pRSFD-SUMO-Tm1-MIPS | pRSFDuet-1 with SUMO-*T. maritima*-MIPS[a] | This study |
| pRSFD-SUMO-At4-MIPS | pRSFDuet-1 with SUMO-*A. thaliana*-MIPS[a] | This study |
| pRSFD-SUMO-Hs13-MIPS | pRSFDuet-1 with SUMO-*H. sapiens*-MIPS[a] | This study |
| pRSFD-SUMO-Si20-MIPS | pRSFDuet-1 with SUMO-*S. indicum*-MIPS[a] | This study |
| pRSFD-SUMO-Zb25-MIPS | pRSFDuet-1 with SUMO-*Z. bailii*-MIPS[a] | This study |
| pRSFD-SUMO-Hc31-MIPS | pRSFDuet-1 with SUMO-*H. contortus*-MIPS[a] | This study |
| pRSFD-His-At4-MIPS-opt | pRSFDuet-1 with codon-optimized *A. thaliana MIPS*[b] in frame with His tag | This study |
| pRSFD-His-Bt6-MIPS-opt | pRSFDuet-1 with codon-optimized *B. thetaiotaomicron MIPS*[b] in frame with His tag | This study |
| pRSFD-His-Cg7-MIPS-opt | pRSFDuet-1 with codon-optimized *C. glabrata MIPS*[b] in frame with His tag | This study |
| pRSFD-His-Mps15-MIPS-opt | pRSFDuet-1 with codon-optimized *M. psychrophilus MIPS*[b] in frame with His tag | This study |
| pRSFD-His-Si20-MIPS-opt | pRSFDuet-1 with codon-optimizedS. indicum *MIPS*[b] in frame with His tag | This study |
| pRSFD-His-Hc31-MIPS-opt | pRSFDuet-1 with codon-optimized *H. contortus MIPS*[b] in frame with His tag | This study |

Table 3.1. *E. coli* strains and plasmids used in this chapter (cont.)

| Name | Genotype[a,b] | Source |
|---|---|---|
| **Plasmids** | | |
| pRSFD-His-IN(V82M) | pRSFD-His-IN with Val-82 mutated to Met | This study |
| pRSFD-His-IN(A83G) | pRSFD-His-IN with Ala-83 mutated to Gly | This study |
| pRSFD-His-IN(N141D) | pRSFD-His-IN with Asn-141 mutated to Asp | This study |
| pRSFD-His-IN(Y250F) | pRSFD-His-IN with Tyr-250 mutated to Phe | This study |
| pRSFD-His-IN(V413R) | pRSFD-His-IN with Val-413 mutated to Arg | This study |
| pRSFD-His-At4-MIPS(A79G)-opt | pRSFD-His-At4-MIPS-opt with Ala-79 mutated to Gly | This study |
| pRSFD-His-At4-MIPS(D146N)-opt | pRSFD-His-At4-MIPS-opt with Asp-79 mutated to Asn | This study |
| pRSFD-His-Pr18-MIPS(F234Y) | pRSFD-His-Pr18-MIPS with Phe-234 mutated to Tyr | This study |

[a] Genes have been codon-optimized for *S. cerevisiae*

[b] Genes have been codon-optimized for *E. coli*

## 3.3. Results

### 3.3.1. MIPS Sequence Similarity Network and Representative Selection

An initial sequence similarity network with alignment score 170 was generated for Pfam PF01658, and a network view is shown in Figure 3.2A. The *S. cerevisiae* INO1 sequence currently used in the pathway is indicated in yellow. As suggested by the literature, eukaryotic sequences show strong sequence similarity, while bacterial and archaeal sequences are widely varied (Figure 3.2A ).[138,140] Optimum organism growth temperature was also mapped onto the network (Figure 3.2B),[159–163] further underscoring the diversity within the bacterial and archaeal clusters.

After generating the network, we selected representative sequences for further study. In general, these sequences were selected to span the available sequence diversity but were also biased towards proteins more likely to perform the MIPS reaction. To increase our confidence in the sequence clusters and better survey the sequence diversity present in the network, a Markov cluster algorithm was implemented using the *clusterMaker* Cytoscape plugin,[164] in addition to the organic clustering shown in Figure 3.2. The resulting alternative clustering allowed us to distinguish potentially different subclusters of sequences, which was particularly useful within the large eukaryotic cluster. The selection of cluster representatives was also guided by applying measures of node centrality (based on pairwise alignment scores as well as %ID) to individual clusters using *clusterMaker*.[164] We overweighted eukaryotic sequences in this initial sample, since most experimentally validated MIPS genes are eukaryotic.[32] Where possible, we selected previously validated sequences as well as those that contained key conserved active site residues identified in the literature (Figure 3.2C),[144] as determined using a multiple alignment. The thirty-one selected sequence representatives are indicated by the large orange nodes in Figure 3.2A and listed in Table 3.2.

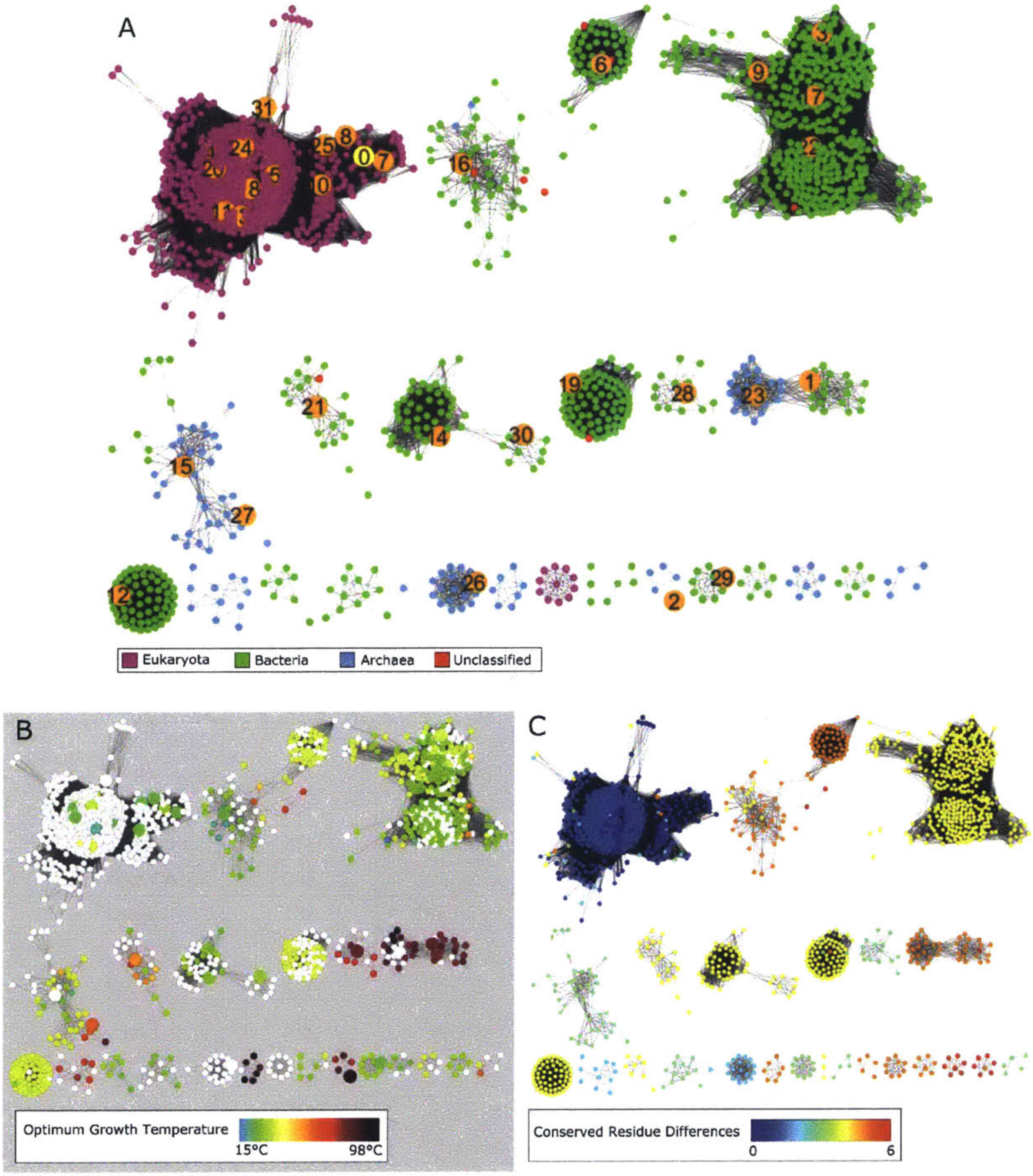

Figure 3.2. MIPS sequence similarity network for Pfam PF01658. This SSN was created using the University of Illinois Enzyme Function Initiative's Enzyme Similarity Tool and visualized in Cytoscape with an E-value cutoff of 170. A) Selected sequences and domains of life. The number labels indicate the variant number. The large yellow node (variant 0) is the *S. cerevisiae* INO1 sequence currently used in the glucaric acid pathway. The large orange nodes indicate sequences selected for functional validation. The remaining nodes are colored by domain of life. B) SSN colored by organism optimum growth temperature. Sequences lacking optimum organism temperature data are shown in white. Large nodes indicate selected sequences. C) SSN colored by number of differences in conserved amino acid residues reported in the literature relative to INO1.

Table 3.2. Selected MIPS Sequences for Experimental Verification.

| # | UniProt ID | Organism | Domain | Length (aa) | UniProt Status |
|---|---|---|---|---|---|
| 1 | Q9X1D6 | *Thermotoga maritima* | Bacteria | 533 | Unreviewed* |
| 2 | A0A075WEG3 | *Archaeoglobus fulgidus* | Archaea | 382 | Unreviewed* |
| 3 | P9WKI1 | *Mycobacterium tuberculosis* | Bacteria | 392 | Reviewed* |
| 4 | Q38862 | *Arabidopsis thaliana* | Eukaryota | 367 | Reviewed |
| 5 | A1CFT5 | *Aspergillus clavatus* | Eukaryota | 510 | Unreviewed |
| 6 | D7IFW4 | *Bacteroides thetaiotaomicron* | Bacteria | 534 | Unreviewed |
| 7 | Q6FQI1 | *Candida glabrata* | Eukaryota | 429 | Reviewed |
| 8 | H8X4H9 | *Candida orthopsilosis* | Eukaryota | 538 | Unreviewed |
| 9 | M1P1K8 | *Corynebacterium halotolerans* | Bacteria | 520 | Unreviewed |
| 10 | R7SX42 | *Dichomitus squalens* | Eukaryota | 363 | Unreviewed |
| 11 | O97477 | *Drosophila melanogaster* | Eukaryota | 549 | Reviewed |
| 12 | E3D8F4 | *Gardnerella vaginalis* | Bacteria | 565 | Unreviewed |
| 13 | Q9NPH2 | *Homo sapiens* | Eukaryota | 380 | Reviewed |
| 14 | L0KRR8 | *Mesorhizobium australicum* | Bacteria | 558 | Unreviewed |
| 15 | K4ME48 | *Methanolobus psychrophilus* | Archaea | 367 | Unreviewed |
| 16 | H1Y1B6 | *Mucilaginibacter paludis* | Bacteria | 376 | Unreviewed |
| 17 | W5TTL7 | *Nocardia nova* | Bacteria | 441 | Unreviewed |
| 18 | H3G8E9 | *Phytophthora ramorum* | Eukaryota | 363 | Unreviewed |
| 19 | D3HVK9 | *Prevotella buccae* | Bacteria | 517 | Unreviewed |
| 20 | Q9FYV1 | *Sesamum indicum* | Eukaryota | 435 | Reviewed |
| 21 | D1C4I3 | *Sphaerobacter thermophilus* | Bacteria | 510 | Unreviewed |
| 22 | F8JTE4 | *Streptomyces cattleya* | Bacteria | 375 | Unreviewed |
| 23 | A0A097QQW8 | *Thermococcus eurythermalis* | Archaea | 360 | Unreviewed |
| 24 | A8WEL5 | *Vigna radiata* | Eukaryota | 382 | Unreviewed |
| 25 | S6EIK9 | *Zygosaccharomyces bailii* | Eukaryota | 510 | Unreviewed |
| 26 | A9A3B6 | *Nitrosopumilus maritimus* | Archaea | 529 | Unreviewed |
| 27 | T2GII1 | *Methanothermobacter thermautotrophicus* | Archaea | 364 | Unreviewed |
| 28 | D3SMX0 | *Thermocrinis albus* | Bacteria | 365 | Unreviewed |
| 29 | A0A076W5U7 | *Bacillus mycoides* | Bacteria | 369 | Unreviewed |
| 30 | I2QG71 | *Bradyrhizobium sp. WSM1253* | Bacteria | 394 | Unreviewed |
| 31 | U6NKU3 | *Haemonchus contortus* | Eukaryota | 366 | Unreviewed |

*Has PDB Structure

### 3.3.2. Initial Evaluation of MIPS Genes

The purview of our JGI Community Science Program project also included synthesis of MIOX homologs, and all MIPS and MIOX variants were codon optimized for *S. cerevisiae* because we intended initial characterization to occur in yeast. However, difficulty with cloning and integration in yeast led us to pursue MIPS evaluation in *E. coli*.

Enzyme expression of many of the homologs initially hampered evaluation. When unmodified sequences were expressed at 30°C, only four achieved measurable MI production from glucose (blue bars in Figure 3.3). These were Cg7 (*C. glabrata*), Vr24 (*V. radiata*), Zb25 (*Z. bailii*), and Hc31 (*H. contortus*). Of these, only Hc31 produced comparable MI to INO1.

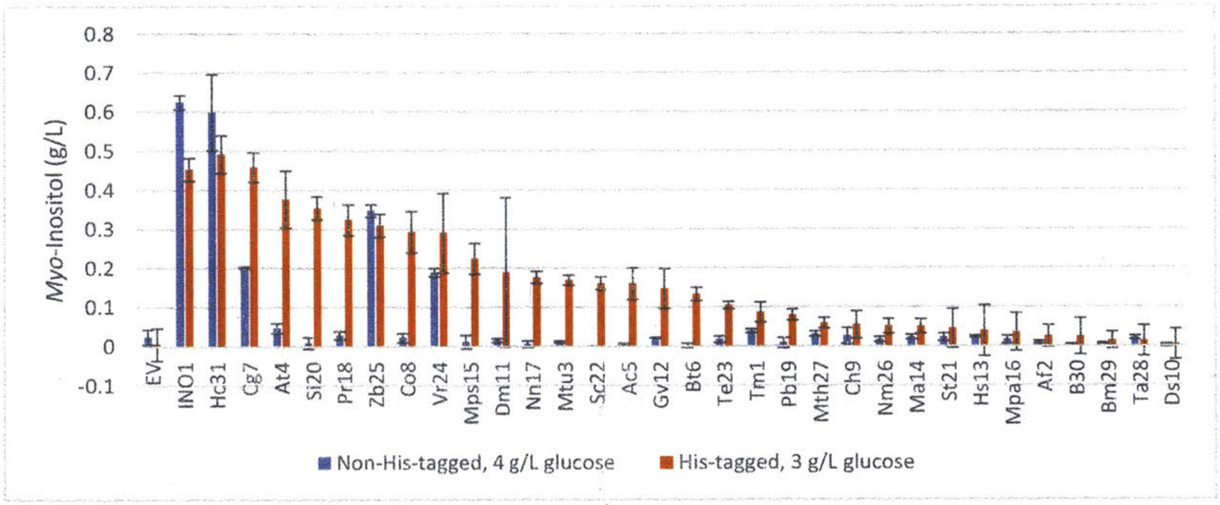

Figure 3.3. MI titers produced by selected MIPS variants. The MIPS variants were expressed from pRSFDuet vectors in LG1460. N-terminal His tags were added where indicated. Cells were grown at 30°C in LB supplemented with glucose as indicated and induced with 0.1 mM IPTG at inoculation. EV refers to the empty vector control, and INO1 refers to the S. *cerevisiae* MIPS. The other MIPS variants are indicated by organism abbreviation and number from the JGI synthesis order. MI concentration was measured at 48 hours after inoculation by HPLC. Error bars correspond to the standard error from three biological replicates.

In order to separate issues of catalytic activity from those of protein expression, the homologs were then His-tagged to allow for detection by Western blot. Interestingly, the addition of the N-terminal His tag dramatically improved MI production at 30°C for many homologs (orange bars in Figure 3.3), increasing the number of functional variants from 4 to 19. In addition, His-tagged MIPS homologs At4 (*A. thaliana*), Cg7, and Hc31 produced MI titers comparable to that of His-tagged INO1. Analysis by Western blot revealed a wide range of expression levels (Figure 3.4 and Table B.5). The variants with the lowest MI titers also showed undetectable or low expression. Even among the top MI producers, some variants had much higher expression than others. In addition, MIPS variants Ac5 (*A. clavatus*), Bt6 (*B. thetaiotaomicron*), Gv12 (*G. vaginalis*), and Mps15 (*M. psychrophilus*) had both moderate titers and low expression.

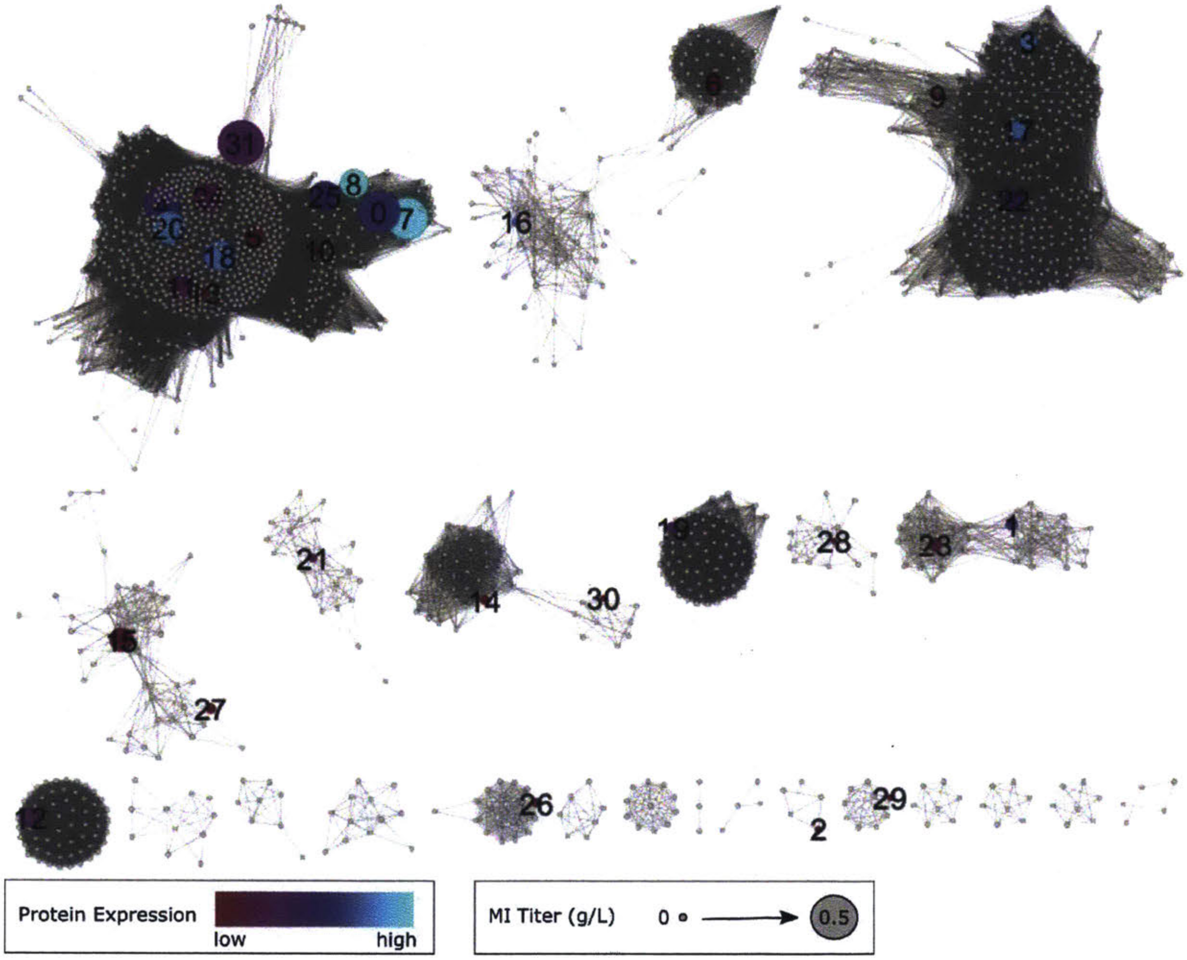

Figure 3.4. MIPS protein expression and MI titer data overlaid on SSN. Nodes are colored by the relative protein expression level, as measured by densitometry from Western blot images. The node size reflects the MI titer achieved from 3 g/L of glucose. The number labels indicate the variant number.

We also measured MIPS activity for the variants that produced detectible MI (Figure 3.5). While it was difficult to distinguish relatively low activity for INO1 and some low-activity MIPS variants from the empty vector control, several variants showed substantially higher activity. The highest measured activity was from variant Gv12, followed by Ac5. These were two of the enzymes that had moderate MI titers but low expression. In addition, variants At4, Bt6, Cg7, Dm11 (*D. melanogaster*), Nn17 (*N. nova*), Pb19 (*P. buccae*), Vr24, Zb25, and Hc31 also showed significant MIPS activity. Apart from Cg7, these variants showed moderate to low expression.

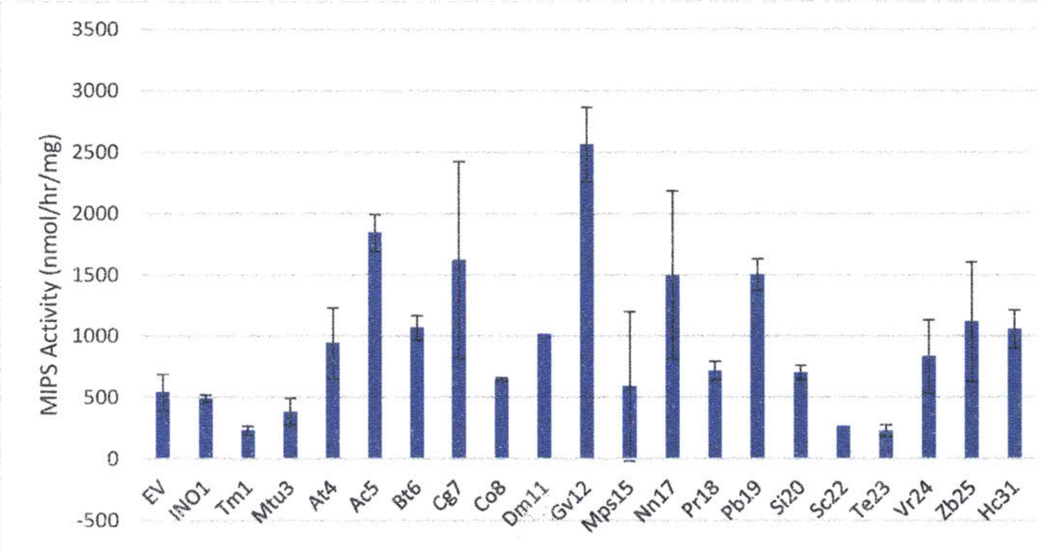

Figure 3.5. MIPS activity of selected MIPS variants. MIPS variants with N-terminal His tags were expressed from pRSFDuet vectors in LG1460. Cells were grown at 30°C in LB supplemented with 3 g/L glucose and induced with 0.1 mM IPTG at inoculation. EV refers to the empty vector control, and INO1 refers to the S. *cerevisiae* MIPS. The other MIPS variants are indicated by organism abbreviation and number from the JGI synthesis order. MIPS activity was measured from crude cell lysates taken at 48 hours. Error bars correspond to the standard error from three biological replicates.

Because the poor activity of INO1 above 30°C currently restricts the temperature at which the glucaric acid pathway is functional,[59] the His-tagged MIPS enzymes were also tested at 37°C. The MI titers are shown by the gray bars in Figure 3.6. While MI titers are generally lower at 37°C than at 30°C, the relative decrease varies widely, with some variants producing no MI at the higher temperature and others maintaining nearly the same titers at both temperatures. His-tagged MIPS variants Cg7, Co8 (*C. orthopsilosis*), Si20 (*S. indicum*), and Hc31 produced MI titers comparable to or better than that of His-tagged INO1. At4, Pr18 (*P. ramorum*), and Zb25 MIPS all experienced a precipitous drop in MI production at 37°C relative to 30°C, and Western blots showed no detectible enzyme at 37°C, suggesting these variants are not stably expressed at the higher temperature.

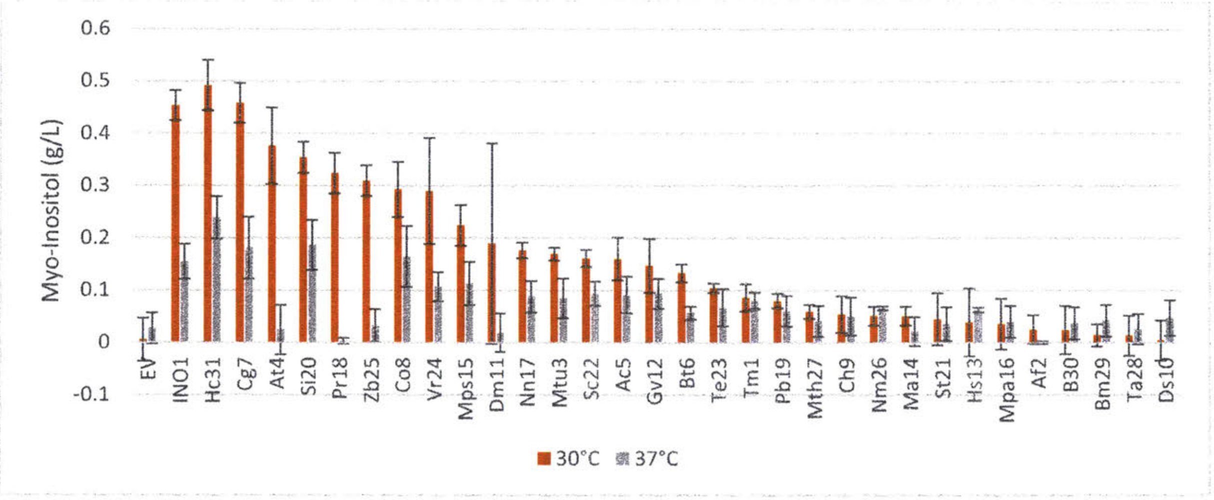

Figure 3.6. MI titers produced by His-tagged MIPS variants at 30°C and 37°C. The MIPS variants were expressed from pRSFDuet vectors in LG1460. Cells were grown at the indicated temperature in LB supplemented with 3 g/L glucose as indicated and induced with 0.1 mM IPTG at inoculation. EV refers to the empty vector control, and INO1 refers to the S. *cerevisiae* MIPS. The other MIPS variants are indicated by organism abbreviation and number from the JGI synthesis order. MI concentration was measured at 48 hours after inoculation by HPLC. Error bars correspond to the standard error from three biological replicates.

### 3.3.3. Improvement of Enzyme Expression

Several of the MIPS variants tested suffered from poor enzyme expression. Because many homologs tolerated and benefited from the addition of N-terminal His tags, N-terminal SUMO tags were added to a partial set of MIPS enzymes (INO1, Tm1 (*T. maritima*), At4, Hs13 (*H. sapiens*), Si20, Zb25, and Hc31). MI production for the His-tagged and SUMO-tagged variants were compared. As shown in Figure 3.7, At4 and Zb25, two variants with poor thermostability, showed an improvement in MI titers at 30°C with the SUMO tag. However, INO1 and Si20 actually performed worse with the SUMO tag. No variants showed improvement with the SUMO fusion at 37°C (data not shown).

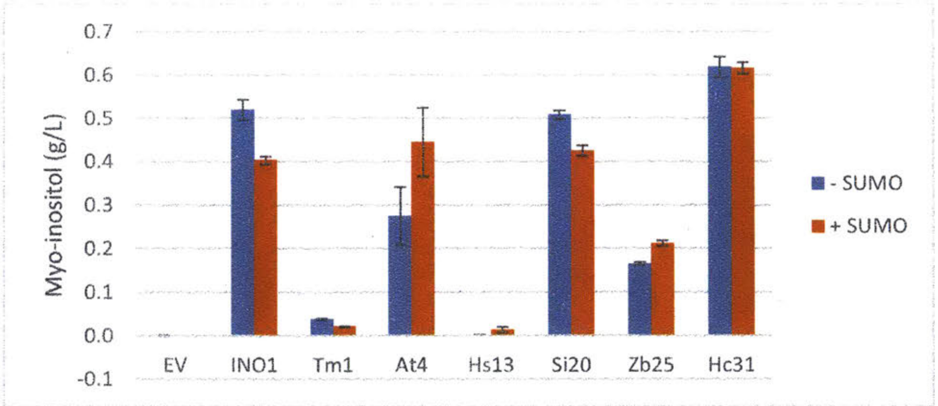

Figure 3.7. MI titers produced by selected SUMO-tagged MIPS variants at 30°C. MIPS variants were expressed from pRSFDuet vectors in LG1460. Cells were grown at 30°C in LB supplemented with 3 g/L glucose and induced with 0.1 mM IPTG at inoculation. EV refers to the empty vector control, and INO1 refers to the S. *cerevisiae* MIPS. The other MIPS variants are indicated by organism abbreviation and number from the JGI synthesis order. MI concentration was measured at 48 hours after inoculation by HPLC. Error bars correspond to the standard error from three biological replicates.

The *MIPS* genes we used in the above experiments were codon optimized for expression in *S. cerevisiae*, which may also contribute to expression limitations. We therefore codon optimized At4, Bt6, Cg7, Mps15, Si20, and Hc31 MIPS for expression in *E. coli*. The MI titers produced by the newly codon-optimized variants at 30°C and 37°C are shown in Figure 3.8. At 30°C, At4 benefits significantly from codon optimization, producing similar titers to non-optimized INO1. Bt6 and Cg7 also show a small benefit. The other MIPS variants perform similarly with and without codon optimization. However, *E. coli* codon-optimized INO1 performs significantly worse than its non-optimized analogue. Interestingly, the optimized INO1 performs well at 37°C. Hc31 also benefits from codon optimization at the higher temperature, but optimized Mps15 and Si20 show a clear decrease in production.

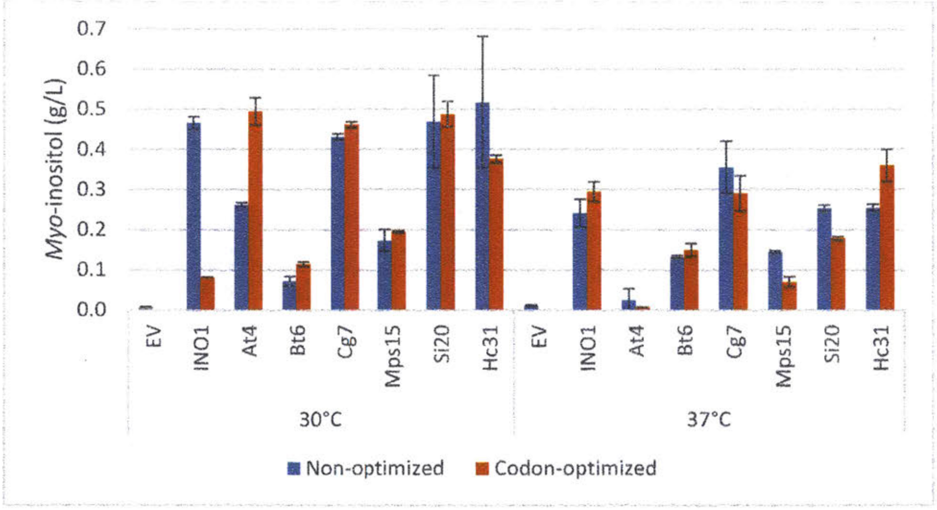

Figure 3.8. MI titers produced by selected *E. coli* codon-optimized MIPS variants. MIPS variants were expressed from pRSFDuet vectors in LG1460. Cells were grown in LB at the indicated temperature, supplemented with 3 g/L glucose, and induced with 0.1 mM IPTG at inoculation. EV refers to the empty vector control, and INO1 refers to the S. *cerevisiae* MIPS. The other MIPS variants are indicated by organism abbreviation and number from the JGI synthesis order. MI concentration was measured at 48 hours after inoculation by HPLC. Error bars correspond to the standard error from three biological replicates.

### 3.3.4. Sequence Analysis

Following initial evaluation, an updated SSN was prepared in September 2018. The original MIPS network was created in 2015 and contained 1,895 nodes representing a total of 4,104 sequences. Sequence databases have continued to grow and change since that time. The updated network is shown in Figure 3.9 at an alignment score cutoff of 180. It contains information for 9,902 sequences, including sequences recently added to the Pfam and UniProt databases as well as additional putative MIPS sequences retrieved from JGI's Phytozome, MycoCosm, and IMG databases[165] by BLAST or Pfam searches. These JGI sequences are shown in gray. A eukaryotic sub-network was also created and is shown in

Figure 3.10 at an alignment score cutoff of 210. Plant sequences cluster most closely together, while animal sequences are present in two distinct groups. Fungal and protist MIPS sequences are the most variable.

A new multiple alignment was also generated using the full set of sequences. This multiple alignment was used to construct a phylogenetic tree (not shown due to its large size) to complement the network, as well as compute residue differences between sequences. As expected, the phylogenetic tree showed many of the same features as the network. In the network, we noticed that the eukaryotic subcluster containing MIPS genes from nematodes, including the top producer Hc31, split off from the rest of the eukaryotic sequences at a relatively low alignment score. This was confirmed by the phylogenetic tree, which suggested the nematode sequences are part of a separate branch from other eukaryotic sequences.

The new multiple alignment was also used to search for sequence information given the performance data we obtained for the 31 MIPS representatives. Amino acid differences relative to INO1 and At4 MIPS were tabulated using the multiple alignment. These differences are summarized for the eukaryotic sequences in Table B.6. Prokaryotic and archaeal sequences encompassed too much sequence variation to explore with our small data set. Statistical power was also limited with only the eukaryotic sequences, but correlation coefficients were calculated for each amino acid difference to evaluate possible contributions of the sequence differences to observed MI titers and protein expression. These coefficients are also shown in Table B.6. Only amino acid differences that were observed in two or more sequences were considered potentially meaningful. Of these, five differences were selected for further study. Five single mutations were introduced into INO1, two were introduced into At4-MIPS-opt, and one was introduced into Pr18-MIPS. These mutations are listed in both Table 3.3 and Table B.6.

Table 3.3. Selected mutations for evaluation.

| **Selected Mutations** | | |
|---|---|---|
| INO1 | At4-opt | Pr18 |
| V82M | | |
| A83G | A79G | |
| Y250F | | F233Y |
| V413R | | |
| N151D | D146N | |

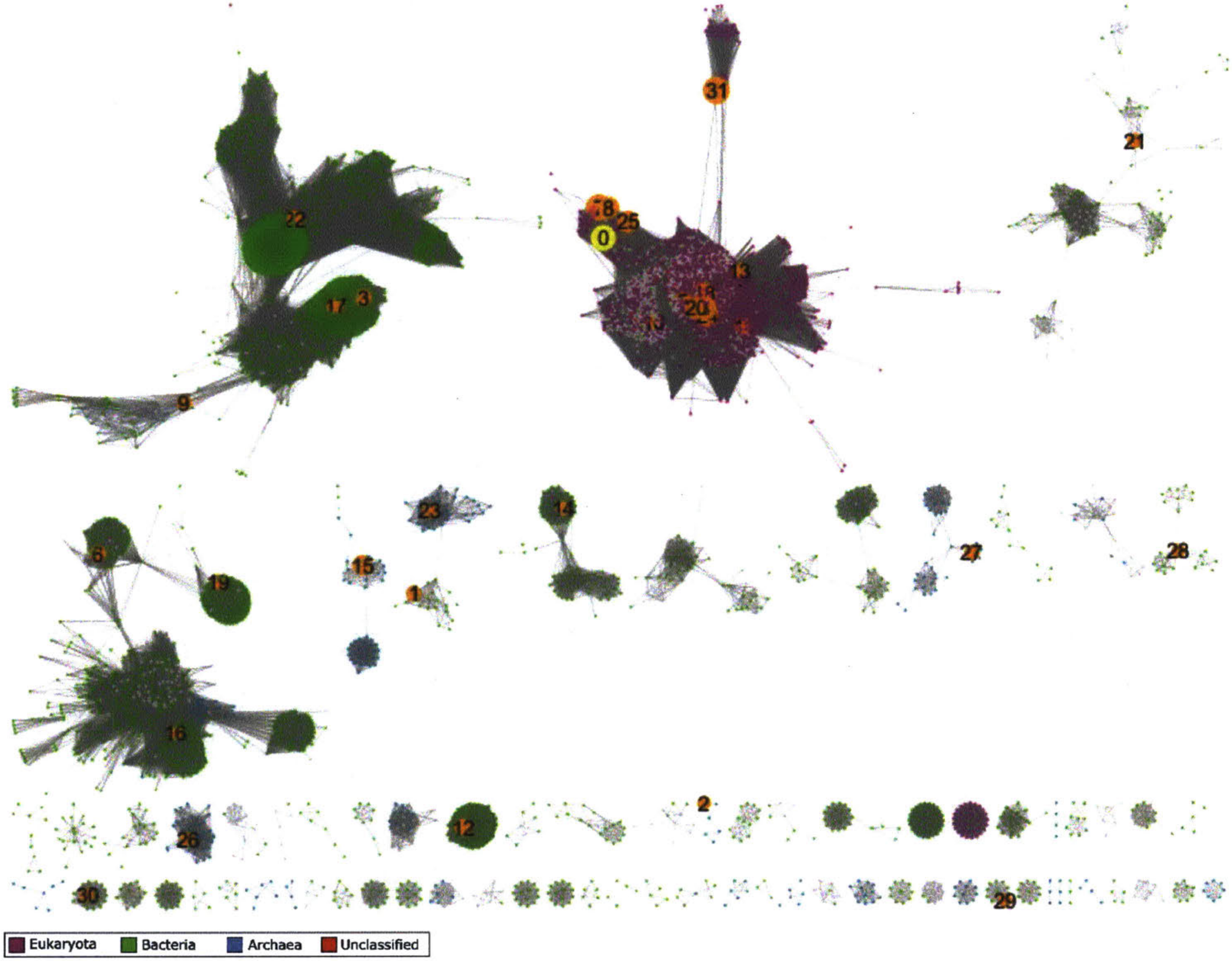

Figure 3.9 Updated MIPS SSN for full network with alignment score cutoff of 180. The yellow node is S. cerevisiae INO1. Orange nodes represent the 31 MIPS homologs (numbered 1-31 as in Table 3.2). The remaining nodes in the network are colored according to domains of life.

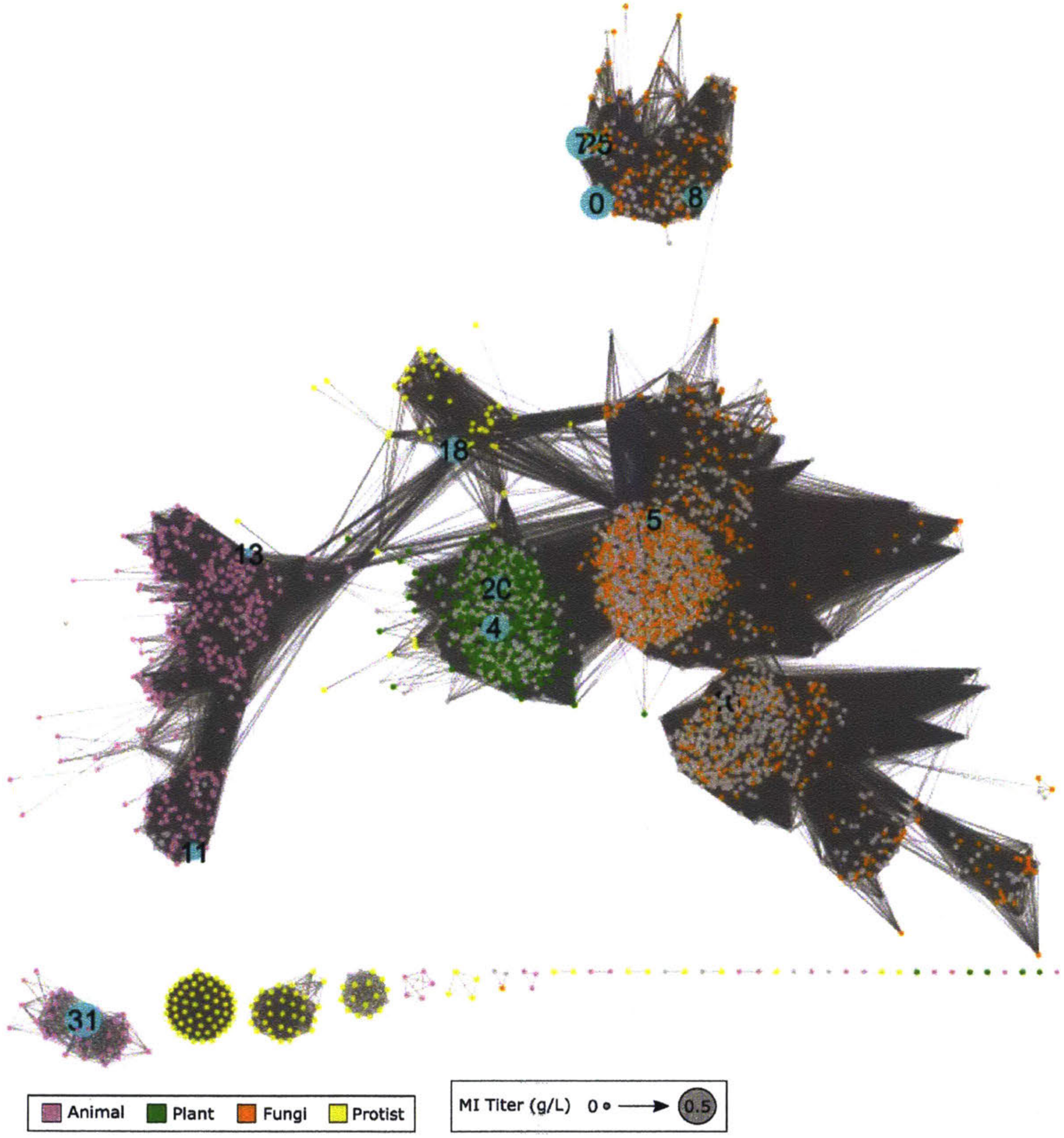

Figure 3.10 Updated MIPS SSN for eukaryotic sub-network with alignment score cutoff of 210. The blue nodes represent the initial set of 31 MIPS homologs, and their relative sizes correspond to MI titers produced by His-tagged variants at 30°C. The remaining nodes in the network are colored according to kingdoms of life.

The MI titers produced by these mutations are shown in Figure 3.11. Most mutations do not have significant effects. However, INO1(A83G) at 37°C and At4-opt(A79G) at 30°C both show reduced titers relative to their unmodified counterparts. These mutations correspond to the same position in the multiple alignment, which is located in the Rossmann fold domain of the enzyme.[140]

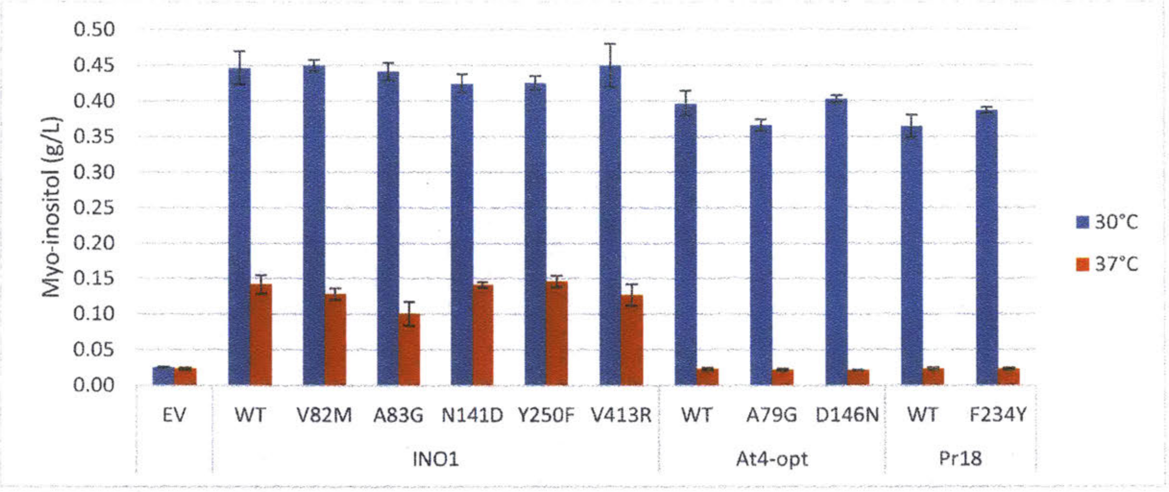

Figure 3.11. Effect of selected MIPS mutations on MI titers. MIPS variants and their mutant versions were expressed from pRSFDuet vectors in LG1460. Cells were grown in LB at the indicated temperature and supplemented with 3 g/L glucose. EV refers to the empty vector control, INO1 refers to the S. *cerevisiae* MIPS, At4-opt refers to the *A. thaliana* codon-optimized MIPS, and Pr18 refers to *P. ramorum* MIPS. MI concentration was measured at 48 hours after inoculation by HPLC. Error bars correspond to the standard error from three biological replicates.

### 3.4. Discussion

While many putative MIPS enzymes have been identified, most have not been experimentally validated. Here, we tested a diverse set of 31 MIPS enzymes in *E. coli* and confirmed activity for many previously uncharacterized enzymes. Active MIPS enzymes were discovered from all domains of life. For enzymes that had been previously studied in the literature, we were able to detect measurable MIPS activity for all but Hs13, which did not express in our system.

Interestingly, we found that His-tagging the MIPS enzymes often led to significant increases in performance. Of the set of MIPS enzymes that were not tagged, only 4 showed discernible MI production. However, 19 of the His-tagged enzymes produced MI, and many of them produced substantially more MI than their non-tagged counterparts. This is a somewhat unusual finding, as His tags have been generally shown to have a neutral or deleterious effect on enzyme activity and function.[166] The disparity in performance between tagged and untagged enzymes may be related to differences in stability in *E. coli* due to the N-end rule.[167]

We examined enzyme performance at both 30°C and 37°C. INO1 currently limits the glucaric acid pathway to operation at 30°C because it has substantially reduced activity at higher temperatures. However, MIOX is more active at 37°C, so finding a MIPS enzyme tolerant of higher temperatures could improve overall pathway flux. While MI titers generally were lower at 37°C than at 30°C for the His-tagged enzymes, the size of the difference varied substantially among the variants. MIPS variants At4, Pr18, and Zb25 retain essentially no MI production at the higher temperature, and no protein was detected by Western blot. In addition, At4 and Zb25 responded well to an N-terminal SUMO fusion tag, and At4 was improved by codon optimization, further suggesting that expression limits performance of these enzymes. At4, Pr18, and Zb25 are all eukaryotic, but they are otherwise quite different. For instance, plant MIPS Si20 and Vr24 (closely related to At4), as well as fungal MIPS INO1, Cg7, and Co8 (closely related to Zb25), did not show this behavior. To date, relatively few studies of MIPS expression or stability at different temperatures have been conducted.[141,168] Further work in this area may yield more information about sequence, structure, and function relationships.

The current analysis is limited by the relatively small number of sequences tested within the full sequence space. While there are fewer amino acid differences between the eukaryotic sequences tested versus the prokaryotic and archaeal sequences, it is still difficult to achieve sufficient statistical power to distinguish beneficial and deleterious amino acid changes. This is due to the large number of amino acid differences relative to sequences as well as the low frequency of most differences within the selected sequences. As a result, most of the differences we selected for mutational analysis did not have significant effects when moved into a new sequence context. However, A83G in INO1 and the analogous mutation A79G in At4-opt both slightly reduced MI production. This residue is located far from the active site within a conserved eukaryotic block identified by Basak and coworkers[145] that forms part of the Rossmann fold domain.[140] In addition to the challenges in identifying contributions of individual amino acids to function, pairwise or higher order relationships between residues and function are also difficult to access. This limitation is significant because evolutionary trajectories are often rugged and epistatic effects often confound analysis of individual amino acid changes.[169]

Our initial selection of variants from the SSN was intended to span the sequence space and identify which parts of the network merit further study. Hc31, which produced the most MI of all variants tested, is part of a distinct nematode subcluster in the SSN and in the phylogenetic tree. Gv12, Ac5, Cg7, Nn17, and Pb19 vary widely in sequence but have significantly higher measured MIPS activity than INO1. Consideration of additional sequences located near these productive variants in the network will help increase statistical power by reducing the number of differences between sequences, likely yielding more structure-function information. The same approach could be taken to better understand sequence features related to stable expression.

### 3.5. Conclusions

Many diverse MIPS enzymes were shown to be functional in *E. coli*, and this work provides a basis for additional exploration of the sequence similarity network to obtain structure-function information. Of the 31 sequences tested, 19 produced detectible MI production when expressed with an N-terminal polyhistidine tag. *H. contortus* (Hc31) MIPS performed as well as or better than INO1 under most experimental conditions. In addition, several homologs had significantly higher MIPS activity than INO1. However, stable protein expression appears to be a challenge for some variants, and the sequence features that affect expression are not yet clear. While statistical power was limited in this study due to the small number of relatively diverse sequences, a mutation in the Rossmann fold domain and far from the active site of INO1 and *A. thaliana* (At4) MIPS has a small negative effect on MI production. Further study of the regions in the network near high-performing variants may help discern additional sequence features that are important for activity and expression.

# 4. Development of Screening Methods for Glucuronic and Glucaric Acid

## Abstract

The improvement of glucaric acid pathway enzymes has been hampered by the lack of an effective screen for protein engineering. Both MIPS and MIOX have relatively low activity in the pathway and have been the focus of previous engineering work. To this end, two potential screens for detection of glucuronic acid or glucaric acid produced from glucose via the glucaric acid pathway were evaluated. The first, a growth screen, appears to be limited by pathway flux, as growth was possible from MI but not from glucose. The second, a previously developed biosensor based on the CdaR activator, was shown to respond to a downstream catabolic product of glucaric acid, likely glycerate, but not to glucaric acid itself. In addition, our desired application of the sensor to production of glucaric acid from glucose was hindered by catabolite repression of the fluorescent reporter or glucaric acid catabolism in the presence of glucose, regulation that was not previously confirmed. Further work to understand this regulation could point to strain engineering strategies to improve these approaches or to alternative screening schemes. While neither screen is currently ideal for use with the glucaric acid pathway, this work clarified native catabolite repression and CdaR regulation in *E. coli.*

## 4.1. Introduction

Efforts to improve the performance of the glucaric acid pathway have been hampered by a lack of effective screening and selection methods. In particular, substantial protein engineering efforts are impractical without high-throughput detection methods. This is significant because protein engineering has been critical for most commercial processes to reach economic viability.[1,3,170] Most substantial gains realized in the glucaric acid pathway to date have instead addressed limitations using small search spaces and low throughput HPLC quantification.

Protein engineering is a powerful tool for improving enzyme activity, specificity, and stability. As discussed in Section 1.3.3, directed evolution involves the generation of random mutations in the protein of interest and then screening of the resulting variation for improvements in the desired characteristic. Because the vast majority of mutations are detrimental, directed evolution requires a high-throughput screen to distinguish improved performance and thus identify the beneficial mutations.[171] In addition, screens can produce context-dependent results, as summarized by the maxim "you get what you screen for."[50]

In the glucaric acid pathway, MIPS and MIOX may both benefit from protein engineering. Both enzymes have low activity relative to Udh,[59] and each appears to limit flux through the pathway under some conditions. Increasing the flux through MIPS relative to glycolysis via dynamic knockdown of *pfk* expression improved titers and yield.[14,74] In addition, improving MIOX stability by adding an N-terminal SUMO fusion also boosted titers.[63] These previous findings suggest that protein engineering has the potential to increase activity and stability and further improve performance.

A previous growth screen was developed with the goal of evolving MIOX. *E. coli* can grow on glucuronic acid but not on MI, and the screen relied on the conversion of MI to glucuronic acid by MIOX to support growth in minimal media. However, instead of producing a MIOX variant with improved activity, directed evolution instead led to the discovery of a mechanism to increase MI transport into the cell.[63] While awareness of this MI transport limitation is valuable, improving MI transport is not beneficial in the full pathway from glucose. A screen design that begins from glucose may be more successful.

An alternative biosensor approach was offered by the Church lab at Harvard. *E. coli* CdaR was previously shown to respond to glycerate, galactarate, and glucarate, and the regulator activates genes involved in the catabolism of the three sugar derivatives.[172,173] The Church lab repurposed the *cdaR* gene to control the expression of GFP from a CdaR-responsive promoter.[54]

Here, we evaluate two different screening strategies for the glucaric acid pathway, one based on growth from glucose and one based on the fluorescent CdaR biosensor developed by the Church lab.

## 4.2. Materials and Methods

### 4.2.1. Strains and Plasmids

The *E. coli* strains and plasmids used in this study are listed in Table 4.1. Primers used for construction are listed in Table 4.2. *E. coli* strain DH5α was used for molecular cloning and plasmid preparation. The *E. coli* strains used for screening were derived from MG1655, MG1655 (DE3), and BL21Star (DE3). M4, MKTS3, and GALG20 were constructed previously in our lab. LG1458 and LG1460 were constructed as described in Chapter 2. Knockouts of *pgi* and *zwf* were performed in BL21Star (DE3) by sequential P1 transduction using Keio collection donor strains JW3985-1 and JW1841-1, respectively.[103] FLP recombinase expressed from plasmid pCP20 was used to cure the kanamycin resistance cassette after each transduction.[104] Transduction and curing were verified by PCR amplification and sequencing using primer pairs LG13 and LG14 for *pgi* and LG15 and LG16 for *zwf*. The resulting single and double knockout strains are LG2212 (Δ*pgi*) and LG2214 (Δ*pgi* Δ*zwf*).

Plasmids containing glucaric acid pathway genes were constructed previously.[59,63] pJKR-H-cdaR was obtained from Addgene.[54] Genes *gudD*, *garL*, and *cdaR* involved in glucaric acid catabolism were amplified from *E. coli* strain MG1655 genomic DNA with primer pairs LG105 and LG106, LG107 and LG108, and LG109 and LG110, respectively. pET-gudD was created from pETDuet-1 by inserting *gudD* into the NcoI and PstI sites. pET-gudD-garL was created from pET-gudD by inserting *garL* into the MfeI and AvrII sites. pACYC-cdaR was created from pACYCDuet-1 by inserting *cdaR* into the NcoI and PstI sites. pACYC-gudD was created by circular polymerase extension cloning (CPEC; Quan & Tian, 2009), using primers LG111 and LG112 to amplify gudD from pET-gudD and primers LG113 and LG114 to amplify the pACYCDuet-1 backbone. Verification of the Duet vector constructs was performed using primers LG73 and LG74 for the first multiple cloning site and LG75 and LG76 for the second site.

### 4.2.2. Culture Conditions

Strains were grown in 2-3 mL of medium in culture tubes at 30°C and 250 rpm. For the growth screen, strains were transformed and recovered in SOC medium, then transferred to liquid M9 medium. For the fluorescence screen, strains were grown in Luria-Bertani (LB) medium. For both, the medium was supplemented as described with *myo*-inositol (MI; Sigma-Aldrich, St. Louis, MA), glucuronic acid (Sigma-Aldrich), glucaric acid (Sigma-Aldrich), glucose (Sigma-Aldrich), and glycerate (Sigma-Aldrich). Working cultures were inoculated from overnight cultures at a dilution of 1:100 and were induced with 100 μM isopropyl β-D-1-thiogalactopyranoside (IPTG) and supplemented with kanamycin (50 μg/mL), carbenicillin (100 μg/mL), and chloramphenicol (34 μg/mL) as required.

### 4.2.3. GFP Measurements

For the fluorescence screen, culture samples were taken, washed in 0.1 M sodium phosphate buffer (pH 7), and diluted 1:2 or 1:4 in sodium phosphate buffer in a 96 well plate. Fluorescence and absorbance measurements were taken in a Tecan Infinite F200Pro plate reader (Männedorf, Switzerland). GFP fluorescence was read at an excitation wavelength of 485 nm and an emission wavelength of 535 nm. Cell density was measured by absorbance at 600 nm. Reported fluorescence values are normalized by cell density.

### 4.2.4. Measurement of Extracellular Metabolites

Where needed, MI, glucuronic acid, and glucaric acid concentrations in culture supernatant samples were quantified by high performance liquid chromatography (HPLC) on an Agilent 1200 series instrument (Santa Clara, CA) with an Aminex HPX-87H anion exchange column (300 mm by 7.8 mm; Bio-Rad Laboratories) using 5 mM sulfuric acid at a flow rate 0.6 mL/min as the mobile phase. The column and refractive index detector temperatures were held at 45°C and 35°C, respectively. Compounds were quantified from 10 μL injections using the refractive index signal.

Table 4.1. *E. coli* strains and plasmids used in this chapter

| Name | Genotype | Source |
|---|---|---|
| **Strains** | | |
| BL21Star(DE3) | F-, ompT, hsdSB (rB- mB-), gal, dcm, rne131, (DE3) | Thermo Fisher (Waltham, MA) |
| MG1655(DE3) | F-, λ-, ilvG-, frb-50, rph-1, (DE3) | Tseng, Martin, Nielsen, & Prather, 2009 |
| JW3985-1 | F-, Δ(araD-araB), ΔlacZ4787(::rrnB-3), lambda-, Δpgi-721::kan, rph-1, Δ(rhaD-rhaB)568, hsdR514 | CGSC #10867 |
| JW1841-1 | F-, Δ(araD-araB), ΔlacZ4787(::rrnB-3), lambda-, Δzwf-777::kan, rph-1, Δ(rhaD-rhaB)568, hsdR514 | CGSC #9537 |
| LG2212 | BL21(DE3) Δpgi | This study |
| LG2214 | BL21(DE3) Δpgi Δzwf | This study |
| M4 | MG1655(DE3) ΔendA ΔrecA Δpgi Δzwf | Shiue, Brockman, & Prather, 2015 |
| LG1458 | MG1655(DE3) ΔuxaC ΔgudD | This study |
| LG1460 | BL21(DE3) ΔuxaC ΔgudD | This study |
| MKTS3 | MG1655(DE3) PlacI-galP ΔptsHIcrr | Prather Lab |
| GALG20 | MG1655 Δpgi ΔendA ΔrecA | Gonçalves, Prazeres, Monteiro, & Prather, 2013 |
| **Plasmids** | | |
| pTrc99A | pBR322 ori, AmpR | Amann, Ochs, & Abel, 1988 |
| pETDuet-1 | ColE1(pBR322) ori, lacI, AmpR | Novagen (Darmstadt, Germany) |
| pACYCDuet-1 | p15A ori, lacI, CmR | Novagen (Darmstadt, Germany) |
| pCP20 | Repa, AmpR, CmR, FLP recombinase expressed by λ pr under control of λ cI857 | CGSC #7629 |
| pTrc-MIOX | pTrc99A with E. coli codon-optimized M. musculus MIOX inserted into the EcoRI and HindIII sites | Moon et al., 2009 |

Table 4.1. *E. coli* strains and plasmids used in this chapter (cont.)

| Name | Genotype | Source |
|---|---|---|
| **Plasmids** | | |
| pRSFD-IN | pRSFDuet-1 with S. cerevisiae INO1 inserted into the EcoRI and HindIII sites | Moon et al., 2009 |
| pRSFD-IN-MI | pRSFD-IN with MIOX inserted into the MfeI and XhoI sites | Moon et al., 2009 |
| pRSFD-MI | pRSFDuet-1 with MIOX inserted into the EcoRI and HindIII sites | Shiue & Prather, 2014 |
| pRSFD-MI-Udh | pRSFD-MI with P. syringae udh inserted into the MfeI and XhoI sites | Prather Lab |
| pJKR-H-cdaR | pUC ori, AmpR, PcdaR-cdaR, PgudP-sfGFP | Rogers et al., 2015 |
| pACYC-cdaR | pACYCDuet-1 with E. coli cdaR inserted into the NcoI and PstI sites | This study |
| pET-gudD | pETDuet-1 with E. coli gudD inserted into the NcoI and PstI sites | This study |
| pET-gudD-garL | pET-gudD with E. coli garL inserted into the MfeI and AvrII sites | This study |
| pACYC-gudD | pACYCDuet-1 with E. coli gudD inserted into the NcoI and PstI sites | This study |

Table 4.2. Oligonucleotides used in this chapter

| Name | Sequence[a] |
|---|---|
| LG13 | gctcctccaacaccgttacttg |
| LG14 | ggattaacctcacggtatgatttccg |
| LG15 | gatattacgcctgtgtgccgtg |
| LG16 | tctcgcgcgaacgttcaatg |
| LG105 | tgcttaCCATGGatgagttctcaatttacgacgc |
| LG106 | tccattCTGCAGttaacgcaccatgcacg |
| LG107 | tgcttaCAATTGatgaataacgatgttttcccgaa |
| LG108 | tgcattCCTAGGttatttttaaaggtatcagccagtttc |
| LG109 | tcgttaCCATGGatggctggctggcatc |
| LG110 | tcaataCTGCAGctaccgctcttcatccagttg |
| LG111 | ccctgtagaaataattttgtttaac |
| LG112 | gcgttcaaatttcgcag |
| LG113 | ctgcgaaatttgaacgc |
| LG114 | gttaaacaaaattatttctacaggg |
| LG73 | ggcgctatcatgccataccg |
| LG74 | gattatgcggccgtgtacaatacg |
| LG75 | cgtattgtacacggccgcataatc |
| LG76 | gctagttattgctcagcggtgg |

[a] Restriction sites used for cloning are capitalized and underlined.

## 4.3. Results

### 4.3.1. Growth Screen from Glucose

It was hypothesized that a growth screen from glucose could be created by extending the previous screen from MI. *E. coli* cannot grow from MI as a sole carbon source, but can grow on glucuronic acid and glucaric acid, the products of MIOX and Udh, respectively. To extend the screen to glucose, it was necessary to create a strain that could only grow on glucose if the glucose were converted to glucuronic acid or glucaric acid. This was done by knocking out *pgi* and *zwf*, which direct glucose-6-phosphate into glycolysis and the pentose phosphate pathway, respectively. The resulting strain LG2214 did not grow on glucose as the sole carbon source. The desired growth pathways with glucuronic acid and glucaric acid intermediates as mapped from the KEGG database[6] are summarized in Figure 4.1.

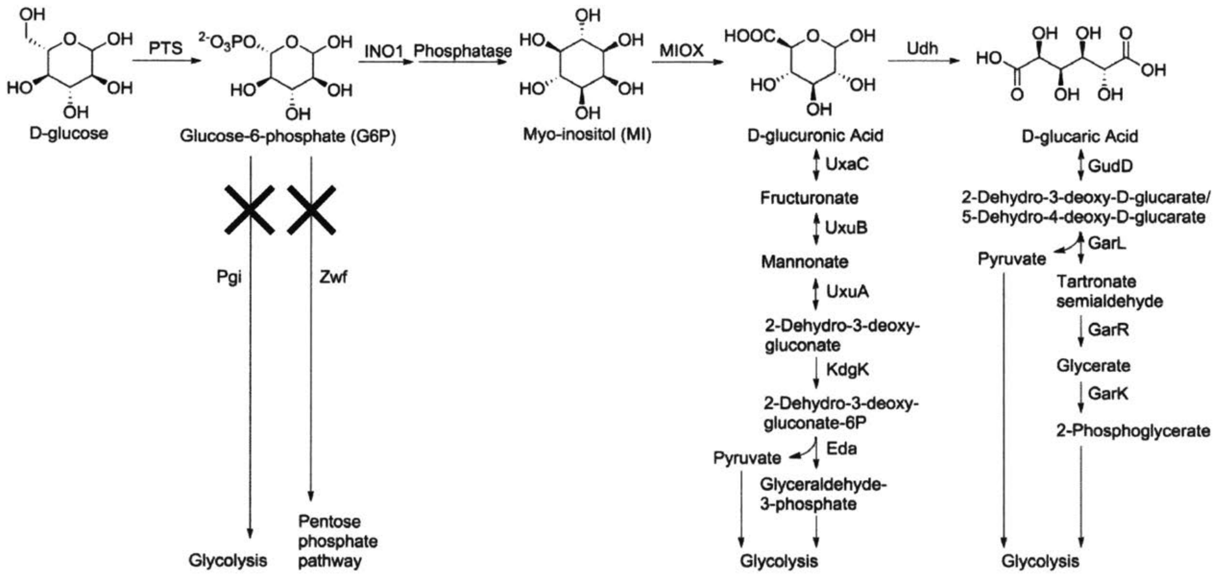

Figure 4.1. Growth pathways from glucose in the engineered strain LG2214. The strain cannot grow from glucose without expression of heterologous glucaric acid pathway genes due to knockouts of *pgi* and *zwf*, whose gene products direct G6P into glycolysis and the pentose phosphate pathway, respectively. Growth can be achieved from glucuronic acid catabolism via UxaC, UxuB, UxuA, KdgK, and Eda. Growth can also be achieved from glucaric acid catabolism via GudD, GarL, GarR, and GarK.

While growth was achieved in the knockout strain from glucuronic acid, glucaric acid, and MI, growth from glucose proved elusive. LG2214 could grow from glucuronic acid or glucaric acid, and LG2214 harboring pTrc-MIOX allowed growth from MI. Growth was relatively slow in all cases, and pretreatment with glucuronic or glucaric acid allowed for faster subsequent growth on MI. LG2214 harboring pTrc-MIOX and pRSFD-IN grew on glucuronic acid, glucaric acid, and MI, but did not grow on glucose, even with pretreatment.

We were concerned that the expression levels of the necessary catabolic enzymes may not be high enough when multiple glucaric acid pathway genes are highly expressed, due to metabolic burden effects or other regulation. Glucuronic acid catabolism requires at least five genes for growth, but glucaric acid catabolism appears to require just two, *gudD* and *garL*. We tested the effect of overexpression of these two genes from pET-gudD-garL. We also considered overexpression of the transcriptional activator *cdaR* from pACYC-cdaR. However, these strains did not show a growth benefit upon glucaric acid addition in LB media, suggesting that endogenous expression may not be the limiting factor.

### 4.3.2. Initial Evaluation of CdaR Biosensor

The function of the biosensor plasmid pJKR-H-cdaR is summarized in Figure 4.2. Briefly, CdaR is a native *E. coli* activator for glucaric acid catabolism genes, and it is autoregulated. pJKR-H-cdaR contains *cdaR* under the control of its native promoter, as well as superfolder GFP under the control of the native *gudP* promoter, which is subject to CdaR activation.[54,172] *gudP* encodes a putative glucarate transporter.[176]

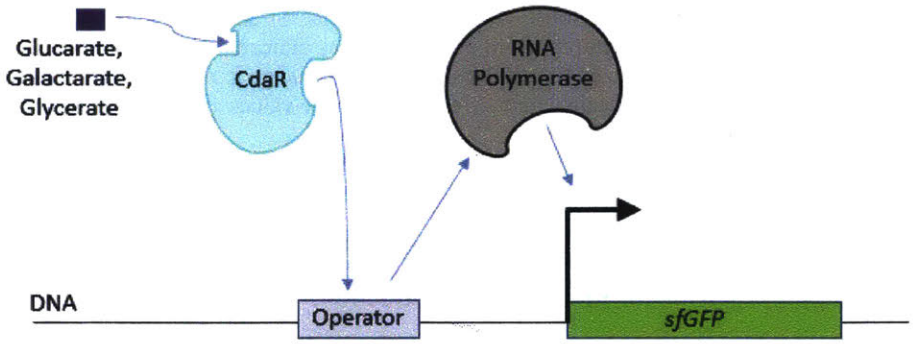

Figure 4.2. CdaR biosensor diagram. CdaR is believed to act as an activator when bound to glucarate, galactarate, or glycerate. CdaR then binds to its operator sequence in the *gudP* promoter region, recruits RNA polymerase, and promotes expression of superfolder GFP.

While the sensor had been previously characterized and applied to the production of glucaric acid,[54,177] we also characterized the sensor's behavior in our system. BL21Star (DE3) harboring pJKR-H-cdaR was grown with exogenous glucaric acid added to the culture medium, and the fluorescence response is shown in Figure 4.3. As expected, the sensor responds to glucaric acid, though it takes some time for the signal to fully develop for higher concentrations of glucaric acid. The sensor responded to glucaric acid at the lowest concentration tested, 0.1 mM (0.2 g/L). It also appears to have a large dynamic range, as the signal does not saturate even at 100 mM (21 g/L), the highest concentration tested. This general behavior is similar to that reported previously, and the dynamic range is well-suited for improving the glucaric acid production beyond its 1 g/L baseline level.

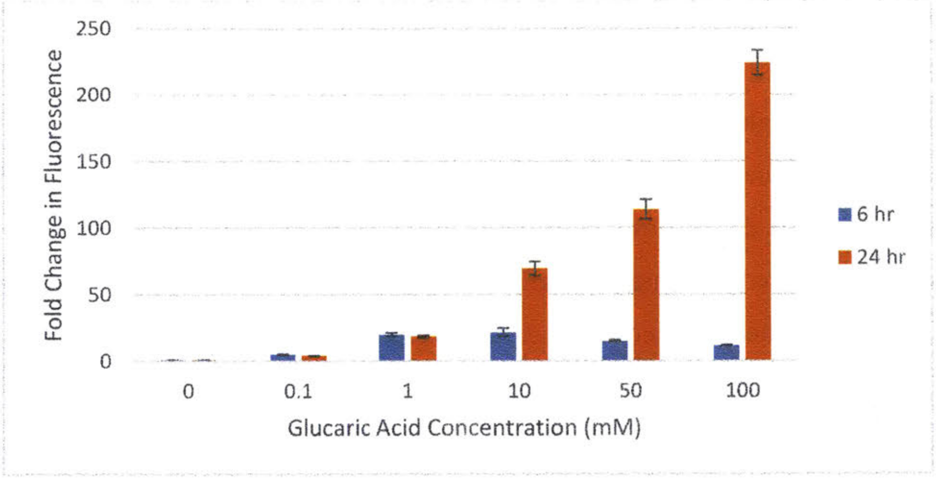

Figure 4.3. Response of CdaR sensor to exogenously added glucaric acid. Strain BL21Star (DE3) harboring pJKR-H-cdaR was grown in LB with various concentrations of glucaric acid as indicated. GFP fluorescence was measured at the indicated times and normalized to cell density, and fold change in normalized fluorescence relative to the 0 mM glucaric acid samples was then calculated. Mean fold change values ± SD for triplicate samples are shown.

The previous experiment tested the effect of exogenously added glucaric acid, but the ultimate goal is to apply a sensor for intracellular production. To test whether the sensor would work in this situation, we applied the sensor to detection of glucaric acid produced from MI via MIOX and Udh. As shown in Figure 4.4, both added glucaric acid and MI led to a substantial increase in fluorescence for MKTS3, with MI showing the stronger response. In contrast, only glucaric acid elicited a response in GALG20, which lacks the λDE3 lysogen that includes the gene for T7 polymerase, which is necessary for *MIOX* and *udh* expression from pRSFD-MI-Udh. Glucaric acid production was confirmed by HPLC.

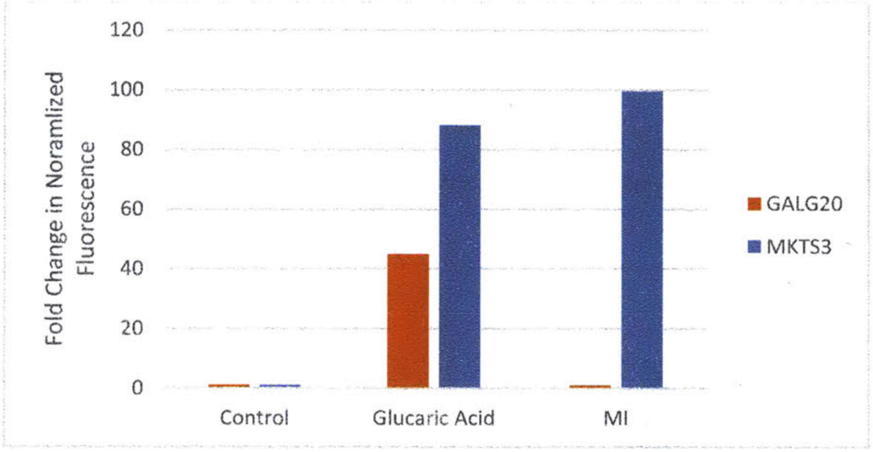

Figure 4.4. Response of CdaR sensor to glucaric acid produced from MI. Strains GALG20 and MKTS3 harboring pRSFD-MI-Udh and pJKR-H-cdaR were grown in LB with 10 mM glucaric acid or 30 mM MI as indicated. GFP fluorescence was measured at 24 hr and normalized to cell density, and fold change in normalized fluorescence relative to the control samples without added glucaric acid or MI was then calculated.

### 4.3.3. Catabolite Repression

In order to avoid the transport limitation found previously with production from MI, we sought a new screen that would allow detection of glucuronic acid or glucaric acid produced from glucose. However, some genes involved in transport and catabolism of glucuronic and glucaric acid have been shown or suggested to be subject to carbon catabolite repression of transcription in the presence of glucose. A common mechanism of catabolite repression is gene activation by CRP (cAMP receptor protein) or Cra (catabolite repressor and activator) in the absence of glucose [28]. As reported in RegulonDB, many genes in glucuronic acid catabolism require activation by CRP or Cra, including *uxaC*, *uxuA*, *uxuB*, *eda*, as well as the transporter *exuT* and regulators *exuR*, *uxuR*, and *kdgR* [178]. In addition, there is a CRP operator site upstream of the glucaric acid transporter *garP* that may control the operon *garPLRK*, and a Cra operator site upstream of a second glucaric acid transporter *gudP* that may control the operon *gudPXD,* though this regulation does not appear in RegulonDB [28,178,179].

With respect to the growth screen described in Section 4.3.1, the deletion of *pgi* and *zwf* has been shown to alleviate catabolite repression for the two sugars xylose and arabinose [72]. The strain from that work, M4, was evaluated for growth from glucuronic acid, glucaric acid, MI, and glucose, and we found it behaved similarly to LG2214. We also observed growth for LG2214 from glucuronic acid or glucaric acid in the presence of glucose.

With respect to the CdaR sensor described in Section 4.3.2, we reevaluated the sensor response to glucaric acid in the presence of glucose to evaluate the impact of catabolite repression. As shown in Figure 4.5, the fluorescence response of the sensor to glucaric acid is dramatically reduced even at low glucose concentrations. At 1.0 g/L of glucose, only a 1.5-fold change is evident, and the response is completely eliminated in the presence of 5.0 g/L glucose.

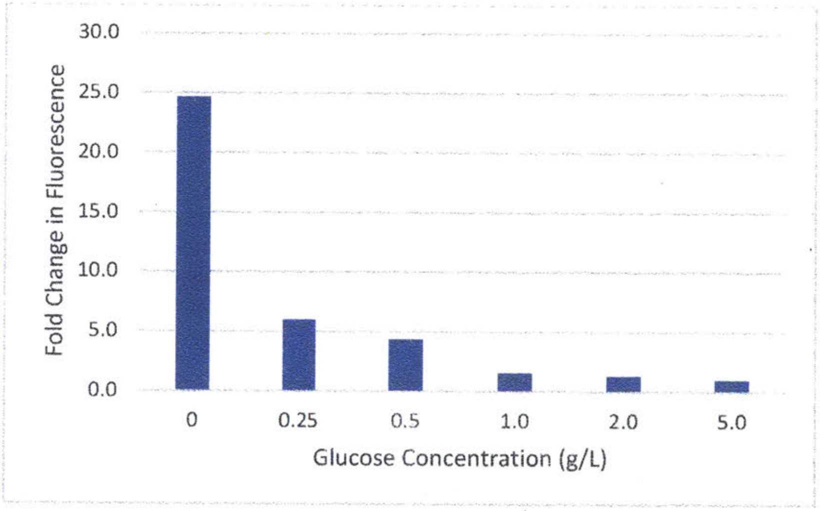

Figure 4.5. Effect of glucose on CdaR sensor response to glucaric acid. Strain BL21Star (DE3) harboring pJKR-H-cdaR was grown in LB with 10 mM glucaric acid and various concentrations of glucose as indicated. GFP fluorescence was measured at 22 hours and normalized to cell density, and fold change in normalized fluorescence was calculated relative to a control grown without glucose or glucaric acid.

This substantial reduction in the response to glucaric acid in the presence of glucose is problematic for a screen of production from glucose. To this end, we evaluated the effectiveness of strain engineering strategies for alleviating this catabolite repression. Previous work suggested that knocking out parts of the phosphotransferase (PTS) system and compensating with upregulation of galactose permease (GalP) could partially alleviate the effect of catabolite repression [30]. Strain MKTS3 is an MG1655 derivative that contains Δ*ptsHIcrr* and *galP* under the control of a constitutive promoter. As shown in Figure 4.6, this strain was tested with the CdaR sensor, and it substantially improved the signal's response at low levels of glucose (up to 1 g/L), maintaining the response near that observed with no added glucose. However, at 2 g/L of glucose, the signal fell dramatically to less than 25% of the response observed with no glucose.

Previous reports also suggested that knocking out *pgi* may reduce catabolite repression via decreased glucose consumption [29]. Strain GALG20, a MG1655 derivative that contains

Δ*pgi*, was also tested with the CdaR sensor, as shown in Figure 4.6. While the fluorescence signal still dropped with added glucose, this strain showed considerable improvement in sensor signal at all glucose concentrations tested (up to 5.0 g/L).

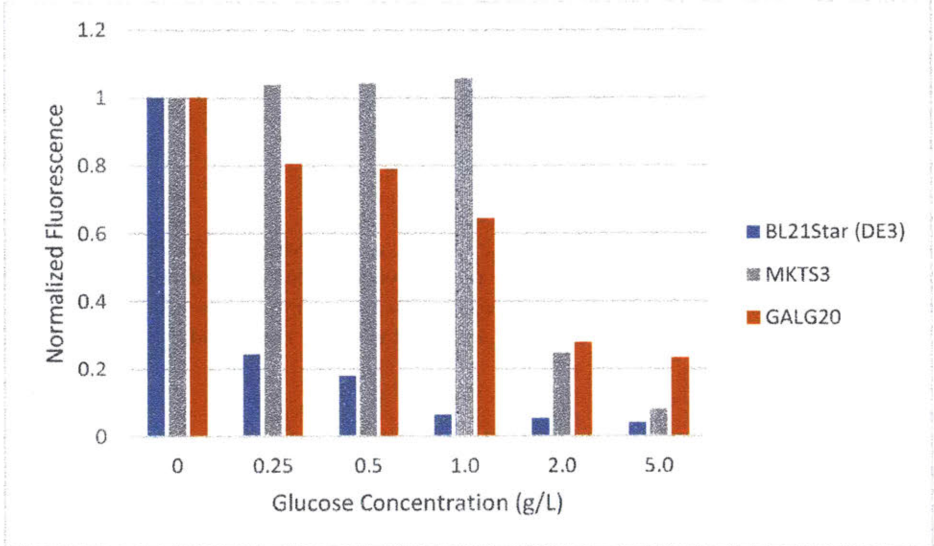

Figure 4.6. Effect of catabolite repression strain engineering strategies on CdaR sensor response to glucaric acid in the presence of glucose. Strains BL21Star (DE3), GALG20, and MKTS3, each harboring pJKR-H-cdaR, were grown in LB with 10 mM glucaric acid and various concentrations of glucose as indicated. GFP fluorescence was measured at 22, 27, and 25 hours for the three strains, respectively, and normalized to cell density. To enable comparisons between strains, the fold change in normalized fluorescence was calculated for each strain relative to its signal at 10 mM glucaric acid and 0 g/L glucose.

Because the CdaR biosensor was originally used to detect intracellularly produced glucaric acid, we also tested whether detection of intracellularly produced glucaric acid was subject to catabolite repression. Strains MG1655 (DE3) and MKTS3 were used test the CdaR sensor response to intracellular production of glucaric acid from MI (via MIOX and Udh) in the presence and absence of glucose. The results are shown in Figure 4.7. As before, glucaric acid alone activated the sensor, as did MI without added glucose. However, when glucose was added, the signal plummeted. The signal for MKTS3 in the presence of glucose is somewhat higher than for MG1655 (DE3), but MKTS3 also had significantly less residual glucose after 24 hr.

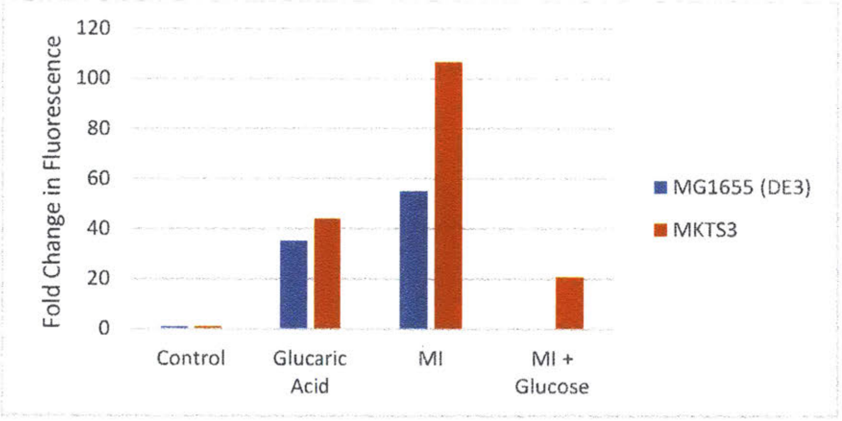

Figure 4.7. Response of CdaR sensor to glucaric acid produced from MI. Strains MG1655 (DE3) and MKTS3 harboring pRSFD-MI-Udh and pJKR-H-cdaR were grown in LB with 10 mM glucaric acid, 30 mM MI, and/or 5 g/L glucose as indicated. GFP fluorescence was measured at 24 hr and normalized to cell density, and fold change in normalized fluorescence relative to the control samples without added glucaric acid, MI, or glucose was then calculated.

### 4.3.4. Clarification of CdaR Sensor Function

In our initial evaluation of the CdaR sensor, we noticed that there was no response to glucaric acid in strains with *gudD* knocked out. GudD is the first enzyme in the catabolism of glucaric acid. Glucarate, galactarate, and glycerate were all shown to activate CdaR in the original study,[172] and glycerate is produced by GarR as a downstream intermediate in the breakdown of both glucarate and galactarate.[6] In addition, the Church lab noticed in a previous study using a slightly different sensor configuration that a Δ*garK* strain improved both the fluorescence signal and glucaric acid production.[180] GarK is the enzyme immediately downstream of glycerate in glucaric acid catabolism. We therefore hypothesized that the sensor may respond to glycerate rather than glucarate.

To evaluate this hypothesis, we first confirmed that GudD was essential for the sensor response. We expressed *gudD* from pACYC-gudD in LG1458, a strain with a Δ*gudD* genotype. The results are shown in Figure 4.8. Overexpression of *gudD* in LG1458 enables the sensor to respond to glucaric acid, whereas there is no response when expression is not induced with IPTG.

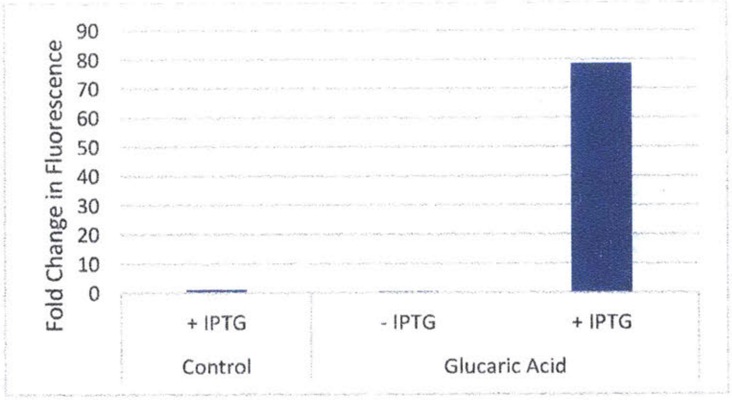

Figure 4.8. Effect of *gudD* overexpression on CdaR sensor response to glucaric acid. Strain LG1458 harboring pJKR-H-cdaR and pACYC-gudD was grown in LB with 10 mM glucaric acid and induced with IPTG as indicated. GFP fluorescence was measured at 24 hours and normalized to cell density, and fold change in normalized fluorescence was calculated relative to the control grown without glucaric acid.

We also compared the exogenous addition of glycerate with the addition of glucaric acid in BL21 and MG1655 strains with and without genomic *gudD* expression. As shown in Figure 4.9, glucaric acid produces a response in strains with *gudD* intact, while glycerate produces a response in all strains.

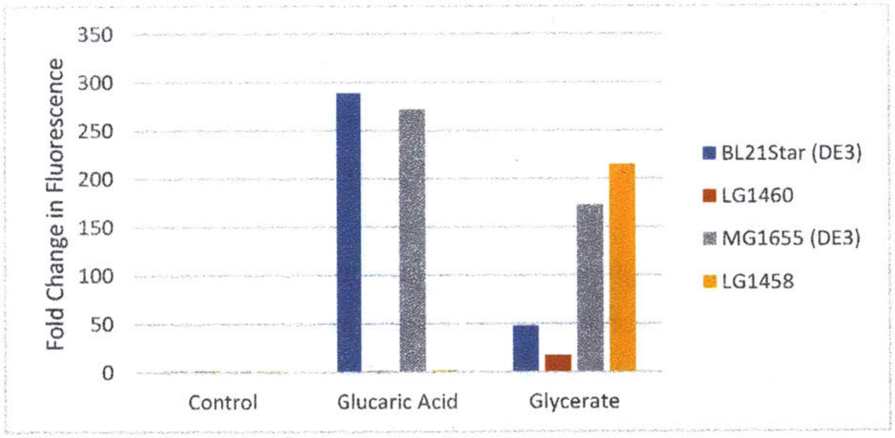

Figure 4.9. Response of CdaR sensor to exogenously added glucaric acid and glycerate. Strains BL21Star (DE3), LG1460, MG1655 (DE3), and LG1458, each harboring pJKR-H-cdaR, were grown in LB with 10 mM glucaric acid or glycerate as indicated. GFP fluorescence was measured at 24 hours and normalized to cell density, and fold change in normalized fluorescence was calculated for each strain relative to the control grown without glucaric acid or glycerate.

### 4.4. Discussion

In the context of the growth screen, we were able to achieve growth as previously described from MI but not from glucose. The lack of growth on glucose is likely due to low pathway flux, as native catabolite repression in the presence of glucose does not appear to be active in LG2214. While growth was possible from exogenously added glucuronic acid and glucaric acid within one or two days in minimal medium, growth from MI often took significantly longer, even after pretreatment with glucuronic acid or glucaric acid. Expression of the additional genes *INO1* and *udh* likely reduces flux compared to the previous MI screen that only required *MIOX*. Moreover, typical production of glucuronic or glucaric acid from MI in production strains with intact *pgi* and *zwf* in LB is 5-7 g/L,[63] but typical production from glucose is much lower, around 1 g/L.[59] While some of this reduction is likely due to competition between INO1, Pgi, and Zwf for glucose-6-phosphate, the dramatic decrease in production is consistent with the lack of growth we observed from glucose in our growth screen strain.

The CdaR biosensor responded to glucaric acid in strains with intact glucaric acid catabolism genes. In particular, strains without *gudD* did not respond to the sensor, and plasmid-based expression restored the response. In addition, addition of glycerate, an intermediate in glucaric acid catabolism and another known activator of CdaR, produced a response in all strains tested. These results show that CdaR does not respond directly to glucaric acid and instead suggest that the true activator is glycerate. While we did not directly interrogate the response from galactaric acid, its catabolism also produces glycerate, so it may also activate CdaR via glycerate.

While a sensor for a downstream catabolic product may theoretically support a screen, in this case the catabolic pathway is branched. As already mentioned, galactaric acid shares much of the same catabolic pathway as glucaric acid. In addition, glycerate is produced by GarR from tartronate semialdehyde, which can itself be produced from other metabolites that connect to central carbon metabolism, including glyoxylate and hydroxypyruvate.[6] Further work is necessary to ensure the sensor response is directly tied to glucaric acid catabolism. If other pathways contribute significantly, it may be possible to eliminate them via strain engineering.

In addition, the CdaR sensor suffers from catabolite repression. The precipitous decline we observed in the sensor's response to fed or produced glucaric acid in the presence of glucose suggests that *gfp* on the sensor plasmid or glucaric acid catabolic genes are affected. The CdaR sensor plasmid uses the *gudP* promoter to drive expression of *gfp*. Since the relevant operons are *gudPXD* and *garPLRK*, both the fluorescent signal and catabolism are likely affected. We were able to partially alleviate the repression using strains MKTS3 and GALG20, and it may be possible to optimize the starting concentration of glucose to allow sufficient production of glucaric acid but minimize repression of the fluorescence signal, similar to the response we saw from MKTS3 in Figure 4.7. However, for the purpose of screening for protein and strain engineering, variation in glucose consumption rates is likely to affect the response. Further work to clarify the genes affected by catabolite repression may allow for targeted overexpression to help alleviate it. In addition, a more complete understanding of native regulation may point to alternative screening strategies that may be more effective for the glucaric acid pathway.

### 4.5. Conclusions

At this stage, neither the growth screen nor the CdaR fluorescent biosensor is well-suited for screening in the context of the glucaric acid pathway in *E. coli*, but we did uncover new pathway regulatory information. Genes involved in glucaric acid transport and catabolism appear to be subject to catabolite repression, which was suggested by computational motif searches but was not previously confirmed by experimental evidence. In addition, CdaR is not directly activated by glucaric acid but instead by a downstream product of glucaric acid catabolism, likely glycerate. Further work to clarify which genes are subject to catabolite repression and to eliminate other pathways that produce glycerate may improve these screening approaches or point to alternative screening opportunities.

# 5. Conclusions and Future Directions

## 5.1. Summary of Goals and Conclusions

The overall goal of this thesis was to further improve the productivity of the glucaric acid pathway. We did so by alleviating oxidative stress, leveraging natural homology, and evaluating screening strategies to improve the reactions catalyzed by MIPS and MIOX.

### 5.1.1. Alleviation of oxidative stress for MI production

MIOX is sensitive to hydrogen peroxide, and MIOX turnover may also produce ROS. However, it was unclear whether either of these phenomena was significant in the context of the glucaric acid pathway. We first verified that MIOX activity in crude cell lysates was sensitive to hydrogen peroxide. Then we took a systematic approach to reduce the prevalence of major ROS species hydrogen peroxide, superoxide, and hydroxyl radicals. We did this by overexpressing native catalase and superoxide dismutases. Overexpression of *katE* substantially increased overall glucuronic acid titers as well as soluble MIOX levels and activity. Overexpression of superoxide dismutases *sodA* or *sodB* in combination with *katE* led to a small additional increase in titers, suggesting that endogenous hydrogen peroxide and superoxide scavenging are insufficient in this system.

Interestingly, overexpression of catalytically inactive versions of iron-binding enzymes *katE* and *sodB* also improved glucuronic acid production. Labile iron has been linked to the production of hydroxyl radicals, so we hypothesized that the inactive enzymes may function as iron chelators. We confirmed that chemical iron chelators were able to produce the same effect.

The strategies used here to alleviate oxidative stress significantly improved performance of the glucaric acid pathway. Moreover, they are general and may be applied in other biological systems.

### 5.1.2. Exploration of natural diversity in MIPS enzymes

The MIPS enzyme appears to limit glucaric acid pathway flux due to its competition with central carbon metabolism for its substrate, glucose-6-phosphate. Many putative MIPS enzymes exist in sequence databases, and we aimed to leverage this natural diversity to help identify improved homologs. Thirty-one MIPS enzymes were selected from a sequence similarity network for Pfam family PF01658. Of these 31 sequences, 19 produced detectible MI production when expressed with an N-terminal polyhistidine tag. One homolog, *H. contortus*

(Hc31) MIPS, performed as well as or better than INO1 under most experimental conditions. Several eukaryotic and prokaryotic enzymes also appear to have significantly higher activity than INO1.

However, stable enzyme expression and thermostability seems to be a significant challenge for some variants. MIPS stability has received relatively little attention in the literature. The strong positive effect of N-terminal His tags on many enzyme variants led us to also test N-terminal SUMO tags and codon optimization. While these methods appeared to help stabilize some variants at 30°C, the effect was not maintained at 37°C.

The small number of relatively diverse sequences tested so far limits statistical power to uncover sequence features that contribute to stability and activity. Mutations at five locations in the multiple sequence alignment were tested based on the limited information we did obtain, and one appears to slightly reduce performance in both the INO1 and At4 MIPS sequences. Despite this challenge, our initial survey of the MIPS sequence network provides guidance for further exploration.

### 5.1.3. Evaluation of glucuronic and glucaric acid screening methods

Improvement of glucaric acid pathway enzymes by protein engineering has been hampered by the lack of an effective screen. Both MIPS and MIOX have low activity in the pathway and may benefit from such engineering efforts. To this end, two potential screens for detection of glucuronic acid or glucaric acid produced from glucose were evaluated.

The first was a growth screen from glucose. A previously-developed growth screen from MI showed that MI import into the cell, rather than MIOX activity, was limiting. In our attempt to extend the screen to glucose, we developed an *E. coli* strain that could not grow from glucose without the expression of glucaric acid pathway genes. This engineered strain was able to grow from MI, but no growth was detected from glucose. Because catabolite repression in the presence of glucose does not appear to prevent consumption of glucuronic or glucaric acid in our strain, the problem is likely insufficient pathway flux.

The second was a fluorescence screen based on the previously-developed CdaR biosensor. While glucaric acid has been reported as an effector of CdaR, we found that the sensor did not respond to glucaric acid itself. Only when glucaric acid was allowed to be catabolized was a response observed. Further work to understand the sensor mechanism suggested that the actual effector molecule may be glycerate, a downstream catabolic product of glucaric acid. In addition, the biosensor suffers from catabolite repression in the presence of glucose, which was not previously recognized. Partial alleviation of this repression can be achieved using strain engineering to reduce glucose import via the PTS system as well as glycolytic flux.

While neither screen is currently ideal for use with the glucaric acid pathway, this work served to clarify native catabolite repression and CdaR regulation in *E. coli.*

## 5.2. Future Directions

This thesis work led to significant improvements in the glucaric acid pathway. In addition, we have gained an increased understanding of pathway enzymes and native regulation in *E. coli*. These findings can be applied for further improvement of glucaric acid production and to other pathways with similar limitations.

### 5.2.1. Oxidative stress

The unexpected finding that a reduction in labile iron levels improves MIOX performance suggests that further work to investigate and improve iron regulation may be worthwhile. Overexpression of genes for iron sequestration proteins, such as *E. coli* ferritin-like *dps* that is part of the OxyR regulon,[75] may produce positive results. This and other systems sensitive to ROS may also benefit from increased attention to iron content in media formulation.

As previously mentioned, the strategies we used to alleviate oxidative stress are quite general and can be used for other pathways and likely also other organisms. We selected catalase and superoxide dismutase because they are efficient enzymes that do not require reducing power. However, several other methods have been used in the literature. In order to evaluate which methods are the most effective, it would be useful to compare them side by side in a variety of systems known to be affected by ROS.

Finally, it is possible that overexpression of ROS scavenging enzymes may be beneficial for other systems under stress. This work showed that the native antioxidant network is not able to reduce ROS to sufficiently low levels for optimal MIOX performance. The regulatory responses to oxidative stress, heat shock, and osmotic stress overlap.[181,182] The likelihood of overwhelming native capacity would be tied to the extent that ROS scavenging contributes to these other stress responses.

### 5.2.2. Protein engineering of MIPS

Further exploration of the MIPS network, coupled with directed evolution, is likely to produce an improved enzyme. Our initial survey of the MIPS sequence network showed a large number of active variants. Expression and stability appear to be significant problems for several enzymes with otherwise good performance, and these are problems that are likely amenable to further bioinformatic analysis. In addition, we found significant differences in stability between enzymes that are very similar in sequence, so the study of additional nearby sequences may illuminate sequence features associated with stability.

Since this work began, a biosensor for MI was developed in our lab. This allows for directed evolution of MIPS enzymes for enhanced MI production from glucose. Gene shuffling using a variety of active homologs may be able to produce an improved MIPS enzyme while offering further information about sequence and function.

### 5.2.3. Screen development for directed evolution in the glucaric acid pathway

The difficulties we encountered in our screen development work underscore that our understanding of native regulation, even in a comparatively well-characterized model organism like *E. coli*, is still incomplete. This finding motivates careful study and confirmation of regulatory mechanisms prior to deployment of biosensors.

It may be possible to modify the CdaR sensor for glucaric acid detection. First, the biosensor should be optimized for detection of glycerate, the likely true effector of the signal response. The consumption of glycerate should be prevented by knockout of *garK*, and other reactions that produce glycerate should be eliminated. If characterization of the glycerate sensor shows that it responds as expected, then relief of catabolite repression can be attempted. Based on our work, the *gudP* promoter used to drive *gfp* appears to be subject to catabolite repression, containing both a CdaR binding site and a CRP-cAMP binding site. The other CdaR-responsive promoters are likely to behave similarly.[172] Because both regulators function as activators and there could be interactions between them, the relief of catabolite repression may be challenging using the native promoter sequences. However, it may be possible to identify the operator sequence and repurpose CdaR as a repressor in a new biosensor.[61,183] Constitutive expression of catabolic genes may also be necessary.

However, while it may be possible to overcome the regulatory limitations of the CdaR sensor, the development of a glucuronic acid biosensor may be more straightforward. UxuR is a repressor that responds to fructuronic acid and regulates glucuronic acid catabolism.[184–187] Fructuronic acid is reversibly produced from glucuronic acid via UxaC. Glucuronic acid catabolism genes are also subject to catabolite repression in the presence of glucose, but only UxaC is likely needed for sensor function.

# References

1. Burk, M. J. & Van Dien, S. Biotechnology for chemical production: Challenges and opportunities. *Trends Biotechnol.* **34,** 187–190 (2016).
2. Burk, M. J. Sustainable production of industrial chemicals from sugars. *Int. Sugar J.* **112,** 30–35 (2010).
3. Erickson, B., Nelson & Winters, P. Perspective on opportunities in industrial biotechnology in renewable chemicals. *Biotechnol. J.* **7,** 176–185 (2012).
4. Beauprez, J. J., De Mey, M. & Soetaert, W. K. Microbial succinic acid production: Natural versus metabolic engineered producers. *Process Biochem.* **45,** 1103–1114 (2010).
5. Paddon, C. J. & Keasling, J. D. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol.* **12,** 355–367 (2014).
6. Kanehisa, M. & Goto, S. KEGG : Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acid Res.* **28,** 27–30 (2000).
7. King, Z. A., Lloyd, C. J., Feist, A. M. & Palsson, B. O. Next-generation genome-scale models for metabolic engineering. *Curr. Opin. Biotechnol.* **35,** 23–29 (2015).
8. Burgard, A. P., Pharkya, P. & Maranas, C. D. OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84,** 647–657 (2003).
9. Oberhardt, M. A. *et al.* Systems-wide prediction of enzyme promiscuity reveals a new underground alternative route for pyridoxal 5'-phosphate production in *E. coli. PLoS Comput. Biol.* **12,** 1–19 (2016).
10. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183 (2013).
11. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8,** 2180–2196 (2013).
12. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. *Nat. Biotechnol.* **33,** 1061–1072 (2015).
13. Cameron, D. E. & Collins, J. J. Tunable protein degradation in bacteria. *Nat. Biotechnol.* **32,** 1276–1281 (2014).
14. Gupta, A., Reizman, I. M. B., Reisch, C. R. & Prather, K. L. J. Dynamic regulation of metabolic flux in engineered bacteria using a pathway-independent quorum-sensing circuit. *Nat. Biotechnol.* **35,** 273–279 (2017).
15. Brockman, I. M. & Prather, K. L. J. Dynamic metabolic engineering: New strategies for developing responsive cell factories. *Biotechnol. J.* **10,** 1360–1369 (2015).
16. Tan, S. Z., Manchester, S. & Prather, K. L. J. Controlling central carbon metabolism for improved pathway yields in *Saccharomyces cerevisiae. ACS Synth. Biol.* **5,** 116–124 (2016).
17. Dai, Z. & Nielsen, J. Advancing metabolic engineering through systems biology of industrial microorganisms. *Curr. Opin. Biotechnol.* **36,** 8–15 (2015).
18. Boyarskiy, S. & Tullman-Ercek, D. Getting pumped: membrane efflux transporters for enhanced biomolecule production. *Curr. Opin. Chem. Biol.* **28,** 15–19 (2015).
19. Lv, Y., Cheng, X., Du, G., Zhou, J. & Chen, J. Engineering of an $H_2O_2$ auto-scavenging in vivo cascade for pinoresinol production. *Biotechnol. Bioeng.* **114,** 2066–2074 (2017).
20. Nielsen, J. & Keasling, J. D. Engineering cellular metabolism. *Cell* **164,** 1185–1197

(2016).

21. Pisithkul, T., Patel, N. M. & Amador-Noguez, D. Post-translational modifications as key regulators of bacterial metabolic fluxes. *Curr. Opin. Microbiol.* **24,** 29–37 (2015).
22. Yang, J.-S., Seo, S. W., Jang, S., Jung, G. Y. & Kim, S. Rational engineering of enzyme allosteric regulation through sequence evolution analysis. *PLoS Comput. Biol.* **8,** e1002612 (2012).
23. Yao, R. & Shimizu, K. Recent progress in metabolic engineering for the production of biofuels and biochemicals from renewable sources with particular emphasis on catabolite regulation and its modulation. *Process Biochem.* **48,** 1409–1417 (2013).
24. Görke, B. & Stülke, J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat. Rev. Microbiol.* **6,** 613–624 (2008).
25. Basak, S. & Jiang, R. Enhancing E. coli tolerance towards oxidative stress via engineering its global regulator cAMP receptor protein (CRP). *PLoS One* **7,** e51179 (2012).
26. Västermark, A. & Saier, M. H. The involvement of transport proteins in transcriptional and metabolic regulation. *Curr. Opin. Microbiol.* **18C,** 8–15 (2014).
27. Chubukov, V., Gerosa, L., Kochanowski, K. & Sauer, U. Coordination of microbial metabolism. *Nat. Rev. Microbiol.* **12,** 327–340 (2014).
28. Kim, D. *et al.* Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP. *Nucleic Acids Res.* **46,** 2901–2917 (2018).
29. Yao, R. *et al.* Catabolic regulation analysis of *Escherichia coli* and its *crp*, *mlc*, *mgsA*, *pgi* and *ptsG* mutants. *Microb. Cell Fact.* **10,** 67 (2011).
30. Solomon, K. V., Sanders, T. M. & Prather, K. L. J. A dynamic metabolite valve for the control of central carbon metabolism. *Metab. Eng.* **14,** 661–671 (2012).
31. Bar-Even, A. *et al.* The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50,** 4402–4410 (2011).
32. Wasmuth, E. V & Lima, C. D. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).
33. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37,** D211–D215 (2009).
34. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2015).
35. Damborsky, J. & Brezovsky, J. Computational tools for designing and engineering enzymes. *Curr. Opin. Chem. Biol.* **19,** 8–16 (2014).
36. Suplatov, D., Voevodin, V. & Švedas, V. Robust enzyme design: Bioinformatic tools for improved protein stability. *Biotechnol. J.* **10,** 344–355 (2015).
37. Uberto, R. & Moomaw, E. W. Protein similarity networks reveal relationships among sequence, structure, and function within the cupin superfamily. *PLoS One* **8,** e74477 (2013).
38. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci.* **103,** 5869–5874 (2006).
39. Crameri, A., Raillard, S.-A., Bermudez, E. & Stemmer, W. P. C. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391,** 288–291 (1998).
40. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16,** 379–394 (2015).
41. Martínez, R. & Schwaneberg, U. A roadmap to directed enzyme evolution and screening

systems for biotechnological applications. *Biol. Res.* **46,** 395–405 (2013).

42. Wilding, M., Scott, C. & Warden, A. C. Computer-guided surface engineering for enzyme improvement. *Sci. Rep.* **8,** 11998 (2018).
43. Chen, Z. & Zhao, H. Rapid creation of a novel protein function by *in vitro* coevolution. *J. Mol. Biol.* **348,** 1273–1282 (2005).
44. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10,** 866–876 (2009).
45. Tracewell, C. A. & Arnold, F. H. Directed enzyme evolution: Climbing fitness peaks one amino acid at a time. *Curr. Opin. Chem. Biol.* **13,** 3–9 (2009).
46. Cobb, R. E., Chao, R. & Zhao, H. Directed evolution: Past, present, and future. *AIChE J.* **59,** 1432–1440 (2013).
47. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31,** 24–33 (2015).
48. Wong, T. S. Sequence saturation mutagenesis (SeSaM): A novel method for directed evolution. *Nucleic Acids Res.* **32,** e26 (2004).
49. Wong, T., Zhurina, D. & Schwaneberg, U. The diversity challenge in directed protein evolution. *Comb. Chem. High Throughput Screen.* **9,** 271–288 (2006).
50. Schmidt-Dannert, C. & Arnold, F. H. Directed evolution of industrial enzymes. *Trends Biotechnol.* **17,** 135–136 (1999).
51. Fernandez-López, R., Ruiz, R., de la Cruz, F. & Moncalián, G. Transcription factor-based biosensors enlightened by the analyte. *Front. Microbiol.* **6,** 648 (2015).
52. Zhang, J., Jensen, M. K. & Keasling, J. D. Development of biosensors and their application in metabolic engineering. *Curr. Opin. Chem. Biol.* **28,** 1–8 (2015).
53. Tamura, T. & Hamachi, I. Recent progress in design of protein-based fluorescent biosensors and their cellular applications. *ACS Chem. Biol.* **9,** 2708–2717 (2014).
54. Rogers, J. K. *et al.* Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Res.* **43,** 7648–7660 (2015).
55. Williams, T. C., Pretorius, I. S. & Paulsen, I. T. Synthetic evolution of metabolic productivity using biosensors. *Trends Biotechnol.* **34,** 371–381 (2016).
56. Eggeling, L., Bott, M. & Marienhagen, J. Novel screening methods — biosensors. *Curr. Opin. Biotechnol.* **35,** 30–36 (2015).
57. Werpy, T. & Petersen, G. Top value added chemicals from biomass. *U.S. Dep. Energy Energy Effic. Renew. Energy* (2004).
58. Smith, T. N. *et al.* Modifications in the nitric acid oxidation of D-glucose. *Carbohydr. Res.* **350,** 6–13 (2012).
59. Moon, T. S., Yoon, S.-H., Lanza, A. M., Roy-Mayhew, J. D. & Prather, K. L. J. Production of glucaric acid from a synthetic pathway in recombinant *Escherichia coli. Appl. Environ. Microbiol.* **75,** 589–595 (2009).
60. Perez, J. L., Jayaprakasha, G. K., Yoo, K. S. & Patil, B. S. Development of a method for the quantification of d-glucaric acid in different varieties of grapefruits by high-performance liquid chromatography and mass spectra. *J. Chromatogr. A* **1190,** 394–397 (2008).
61. Doong, S. J., Gupta, A. & Prather, K. L. J. Layered dynamic regulation for improving metabolic pathway productivity in *Escherichia coli. Proc. Natl. Acad. Sci.* **115,** 2964–2969 (2018).
62. Bollinger, J. M., Diao, Y., Matthews, M. L., Xing, G. & Krebs, C. *myo*-Inositol

oxygenase: a radical new pathway for $O_2$ and C–H activation at a nonheme diiron cluster. *Dalt. Trans.* 905–914 (2009). doi:10.1039/B811885J

63. Shiue, E. & Prather, K. L. J. Improving D-glucaric acid production from *myo*-inositol in *E. coli* by increasing MIOX stability and *myo*-inositol transport. *Metab. Eng.* **22,** 22–31 (2014).
64. Snyder, R. A. *et al.* Circular dichroism, magnetic circular dichroism, and variable temperature variable field magnetic circular dichroism studies of biferrous and mixed-valent myo-inositol oxygenase: insights into substrate activation of O2 reactivity. *J. Am. Chem. Soc.* **135,** 15851–15863 (2013).
65. Worsdorfer, B. *et al.* Organophosphonate-degrading PhnZ reveals an emerging family of HD domain mixed-valent diiron oxygenases. *Proc. Natl. Acad. Sci.* **110,** 18874–18879 (2013).
66. Reddy, C. C., Pierzchala, P. A. & Hamilton, G. A. *myo*-Inositol oxygenase from hog kidney. II. Catalytic properties of the homogeneous enzyme. *J. Biol. Chem.* **256,** 8519–8524 (1981).
67. Naber, N. I., Swan, J. S. & Hamilton, G. A. L-*myo*-Inosose-1 as a probable intermediate in the reaction catalyzed by *myo*-inositol oxygenase. *Biochemistry* **25,** 7201–7207 (1986).
68. Dutta, R. K. *et al.* Beneficial effects of *myo*-inositol oxygenase deficiency in cisplatin-induced AKI. *J. Am. Soc. Nephrol.* **28,** 1421–1436 (2017).
69. Sun, L., Dutta, R. K., Xie, P. & Kanwar, Y. S. *myo*-Inositol oxygenase overexpression accentuates generation of reactive oxygen species and exacerbates cellular injury following high glucose ambience: A new mechanism relevant to the pathogenesis of diabetic nephropathy. *J. Biol. Chem.* **291,** 5688–5707 (2016).
70. Duan, J. *et al. OsMIOX,* a *myo*-inositol oxygenase gene, improves drought tolerance through scavenging of reactive oxygen species in rice (*Oryza sativa* L.). *Plant Sci.* **196,** 143–151 (2012).
71. Valluru, R. & Van den Ende, W. *Myo*-inositol and beyond--emerging networks under stress. *Plant Sci.* **181,** 387–400 (2011).
72. Shiue, E., Brockman, I. M. & Prather, K. L. J. Improving product yields on D-glucose in *Escherichia coli* via knockout of *pgi* and *zwf* and feeding of supplemental carbon sources. *Biotechnol. Bioeng.* **112,** 579–587 (2015).
73. Moon, T. S., Dueber, J. E., Shiue, E. & Prather, K. L. J. Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered *E. coli. Metab. Eng.* **12,** 298–305 (2010).
74. Reizman, I. M. B. *et al.* Improvement of glucaric acid production in *E. coli* via dynamic control of metabolic fluxes. *Metab. Eng. Commun.* **2,** 109–116 (2015).
75. Imlay, J. A. Transcription factors that defend bacteria against reactive oxygen species. *Annu. Rev. Microbiol.* **69,** 93–108 (2015).
76. Mishra, S. & Imlay, J. Why do bacteria use so many enzymes to scavenge hydrogen peroxide? *Arch. Biochem. Biophys.* **525,** 145–160 (2012).
77. Adolfsen, K. J. & Brynildsen, M. P. A kinetic platform to determine the fate of hydrogen peroxide in *Escherichia coli. PLOS Comput. Biol.* **11,** e1004562 (2015).
78. Jang, S. & Imlay, J. A. Micromolar intracellular hydrogen peroxide disrupts metabolism by damaging iron-sulfur enzymes. *J. Biol. Chem.* **282,** 929–937 (2007).
79. Winterbourn, C. C. Reconciling the chemistry and biology of reactive oxygen species. *Nat. Chem. Biol.* **4,** 278–286 (2008).

80. Farrugia, G. & Balzan, R. Oxidative stress and programmed cell death in yeast. *Front. Oncol.* **2,** 64 (2012).
81. Imlay, J. & Linn, S. DNA damage and oxygen radical toxicity. *Science* **240,** 1302–1309 (1988).
82. Kakhlon, O. & Cabantchik, Z. I. The labile iron pool: Characterization, measurement, and participation in cellular processes. *Free Radic. Biol. Med.* **33,** 1037–1046 (2002).
83. Winterbourn, C. C. Superoxide as an intracellular radical sink. *Free Radic. Biol. Med.* **14,** 85–90 (1993).
84. Kehrer, J. P. The Haber – Weiss reaction and mechanisms of toxicity. *Toxicology* **149,** 43–50 (2000).
85. Liochev, S. I. & Fridovich, I. The Haber-Weiss cycle—70 years later: An alternative view. *Redox Rep.* **7,** 55–57 (2002).
86. Imlay, J. A. The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. *Nat. Rev. Microbiol.* **11,** 443–454 (2013).
87. Miller, A.-F. Superoxide dismutases: Active sites that save, but a protein that kills. *Curr. Opin. Chem. Biol.* **8,** 162–168 (2004).
88. Winterbourn, C. C. & Metodiewa, D. The reaction of superoxide with reduced glutathione. *Arch. Biochem. Biophys.* **314,** 284–290 (1994).
89. Fontecave, M. & Pierre, J. L. Iron: Metabolism, toxicity and therapy. *Biochimie* **75,** 767–773 (1993).
90. Andre, C., Kim, S. W., Yu, X.-H. & Shanklin, J. Fusing catalase to an alkane-producing enzyme maintains enzymatic activity by converting the inhibitory byproduct $H_2O_2$ to the cosubstrate $O_2$. *Proc. Natl. Acad. Sci.* **110,** 3191–3196 (2013).
91. Xu, P., Qiao, K. & Stephanopoulos, G. Engineering oxidative stress defense pathways to build a robust lipid production platform in *Yarrowia lipolytica. Biotechnol. Bioeng.* **114,** 1521–1530 (2017).
92. Zhang, S. *et al.* Alleviation of reactive oxygen species enhances PUFA accumulation in *Schizochytrium* sp. through regulating genes involved in lipid metabolism. *Metab. Eng. Commun.* **6,** 39–48 (2018).
93. Abbott, D. A. *et al.* Catalase overexpression reduces lactic acid-induced oxidative stress in *Saccharomyces cerevisiae. Appl. Environ. Microbiol.* **75,** 2320–2325 (2009).
94. Chin, W.-C., Lin, K.-H., Liu, C.-C., Tsuge, K. & Huang, C.-C. Improved n-butanol production via co-expression of membrane-targeted tilapia metallothionein and the clostridial metabolic pathway in *Escherichia coli. BMC Biotechnol.* **17,** 36 (2017).
95. Chen, Z., Sun, X., Li, Y., Yan, Y. & Yuan, Q. Metabolic engineering of *Escherichia coli* for microbial synthesis of monolignols. *Metab. Eng.* **39,** 102–109 (2017).
96. Wang, J., Shen, X., Yuan, Q. & Yan, Y. Microbial synthesis of pyrogallol using genetically engineered *Escherichia coli. Metab. Eng.* **45,** 134–141 (2018).
97. Lewis, D. F. V. Oxidative stress: The role of cytochromes P450 in oxygen activation. *J. Chem. Technol. Biotechnol.* **77,** 1095–1100 (2002).
98. Bruno-Barcena, J. M., Andrea Azcarate-Peril, M. & Hassan, H. M. Role of antioxidant enzymes in bacterial resistance to organic acids. *Appl. Environ. Microbiol.* **76,** 2747–2753 (2010).
99. Fu, R. Y. *et al.* Introducing glutathione biosynthetic capability into *Lactococcus lactis* subsp. *cremoris* NZ9000 improves the oxidative-stress resistance of the host. *Metab. Eng.* **8,** 662–671 (2006).

100. Gómez-Pastor, R., Pérez-Torrado, R., Cabiscol, E., Ros, J. & Matallana, E. Engineered Trx2p industrial yeast strain protects glycolysis and fermentation proteins from oxidative carbonylation during biomass propagation. *Microb. Cell Fact.* **11,** 4 (2012).

101. Gómez-Pastor, R., Pérez-Torrado, R., Cabiscol, E., Ros, J. & Matallana, E. Reduction of oxidative cellular damage by overexpression of the thioredoxin TRX2 gene improves yield and quality of wine yeast dry active biomass. *Microb. Cell Fact.* **9,** 9 (2010).

102. Halliwell, B. Cell culture, oxidative stress, and antioxidants: Avoiding pitfalls. *Biomed. J.* **37,** 99–105 (2014).

103. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 2006.0008 (2006).

104. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci.* **97,** 6640–6645 (2000).

105. St-Pierre, F. *et al.* One-step cloning and chromosomal integration of DNA. *ACS Synth. Biol.* **2,** 537–541 (2013).

106. Yoon, S.-H., Moon, T. S., Iranpour, P., Lanza, A. M. & Prather, K. J. Cloning and characterization of uronate dehydrogenases from two pseudomonads and *Agrobacterium tumefaciens* strain C58. *J. Bacteriol.* **191,** 1565–1573 (2009).

107. International Genetically Engineered Machine Foundation. Registry of Standard Biological Parts. (2018). Available at: http://parts.igem.org/.

108. Agilent. QuikChange primer design. (2018). Available at: http://www.agilent.com/genomics/qcpd.

109. Moon, T. S., Yoon, S.-H., Tsang Mui Ching, M.-J., Lanza, A. M. & Prather, K. L. J. Enzymatic assay of D-glucuronate using uronate dehydrogenase. *Anal. Biochem.* **392,** 183–185 (2009).

110. Tseng, H.-C., Martin, C. H., Nielsen, D. R. & Prather, K. L. J. Metabolic engineering of *Escherichia coli* for enhanced production of (*R*)- and (*S*)-3-hydroxybutyrate. *Appl. Environ. Microbiol.* **75,** 3137–3145 (2009).

111. Obinger, C., Maj, M., Nicholls, P. & Loewen, P. Activity, peroxide compound formation, and heme d synthesis in *Escherichia coli* HPII catalase. *Arch. Biochem. Biophys.* **342,** 58–67 (1997).

112. Sevcenco, A.-M. *et al.* Exploring the microbial metalloproteome using MIRAGE. *Metallomics* **3,** 1324–1330 (2011).

113. Miller, A.-F. & Wang, T. A single outer-sphere mutation stabilizes apo-Mn superoxide dismutase by 35 °C and disfavors Mn binding. *Biochemistry* **56,** 3787–3799 (2017).

114. Yikilmaz, E., Rodgers, D. W. & Miller, A. F. The crucial importance of chemistry in the structure-function link: Manipulating hydrogen bonding in iron-containing superoxide dismutase. *Biochemistry* **45,** 1151–1161 (2006).

115. Jacques, J.-F. *et al.* RyhB small RNA modulates the free intracellular iron pool and is essential for normal growth during iron limitation in Escherichia coli. *Mol. Microbiol.* **62,** 1181–1190 (2006).

116. Liu, Z. D., Liu, D. Y. & Hider, R. C. Iron chelator chemistry. In *Iron Chelation Therapy* (ed. Hershko, C.) 141–166 (Springer, 2002).

117. Kicic, A., Chua, A. C. & Baker, E. Effect of iron chelators on proliferation and iron uptake in hepatoma cells. *Cancer* **92,** 3093–3110 (2001).

118. Beauchene, N. A. *et al.* $O_2$ availability impacts iron homeostasis in *Escherichia coli. Proc. Natl. Acad. Sci.* **114,** 12261–12266 (2017).

119. Imlay, J. A. Diagnosing oxidative stress in bacteria: Not as easy as you might think. *Curr. Opin. Microbiol.* **24,** 124–131 (2015).
120. Rhee, S. G., Chang, T.-S., Jeong, W. & Kang, D. Methods for detection and measurement of hydrogen peroxide inside and outside of cells. *Mol. Cells* **29,** 539–549 (2010).
121. Kalyanaraman, B. *et al.* Measuring reactive oxygen and nitrogen species with fluorescent probes: Challenges and limitations. *Free Radic. Biol. Med.* **52,** 1–6 (2012).
122. Seaver, L. C. & Imlay, J. A. Hydrogen peroxide fluxes and compartmentalization inside growing *Escherichia coli. J. Bacteriol.* **183,** 7182–7189 (2001).
123. Liu, Y. & Imlay, J. A. Cell death from antibiotics without the involvement of reactive oxygen species. *Science* **339,** 1210–1213 (2013).
124. Seaver, L. C. & Imlay, J. A. Are respiratory enzymes the primary sources of intracellular hydrogen peroxide? *J. Biol. Chem.* **279,** 48742–48750 (2004).
125. Imlay, J. A. Cellular defenses against superoxide and hydrogen peroxide. *Annu. Rev. Biochem.* **77,** 755–776 (2008).
126. Zamocky, M., Furtmüller, P. G. & Obinger, C. Evolution of catalases from bacteria to humans. *Antioxid. Redox Signal.* **10,** 1527–1548 (2008).
127. Perry, J. J. P., Shin, D. S., Getzoff, E. D. & Tainer, J. A. The structural biochemistry of the superoxide dismutases. *Biochim. Biophys. Acta - Proteins Proteomics* **1804,** 245–262 (2010).
128. Miller, A.-F. Superoxide dismutases: Ancient enzymes and new insights. *FEBS Lett.* **586,** 585–595 (2012).
129. Hopkin, K. A., Papazian, M. A. & Steinman, H. M. Functional differences between manganese and iron superoxide dismutases in Escherichia coli K-12. *J. Biol. Chem.* **267,** 24253–24258 (1992).
130. Niederhoffer, E. C., Naranjo, C. M., Bradley, K. L. & Fee, J. A. Control of *Escherichia coli* superoxide dismutase (*sodA* and *sodB*) genes by the ferric uptake regulation (*fur*) locus. *J. Bacteriol.* **172,** 1930–1938 (1990).
131. Lim, J. B., Barker, K. A., Huang, B. K. & Sikes, H. D. In-depth characterization of the fluorescent signal of HyPer, a probe for hydrogen peroxide, in bacteria exposed to external oxidative stress. *J. Microbiol. Methods* **106,** 33–39 (2014).
132. Lennen, R. M. & Herrgård, M. J. Combinatorial strategies for improving multiple-stress resistance in industrially relevant *Escherichia coli* strains. *Appl. Environ. Microbiol.* **80,** 6223–6242 (2014).
133. Van Der Heijden, J. *et al.* Exploring the redox balance inside gram-negative bacteria with redox-sensitive GFP. *Free Radic. Biol. Med.* **91,** 34–44 (2016).
134. Hantke, K. Iron and metal regulation in bacteria. *Curr. Opin. Microbiol.* **4,** 172–177 (2001).
135. Sheng, Y. *et al.* Superoxide dismutases and superoxide reductases. *Chem. Rev.* **114,** 3854–3918 (2014).
136. Gupta, A., Hicks, M. A., Manchester, S. P. & Prather, K. L. J. Porting the synthetic D-glucaric acid pathway from *Escherichia coli* to *Saccharomyces cerevisiae. Biotechnol. J.* **11,** 1201–1208 (2016).
137. Brockman, I. M. & Prather, K. L. J. Dynamic knockdown of *E. coli* central metabolism for redirecting fluxes of primary metabolites. *Metab. Eng.* **28,** 104–113 (2015).
138. Majumder, A. L., Chatterjee, A., Ghosh Dastidar, K. & Majee, M. Diversification and evolution of L-*myo*-inositol 1-phosphate synthase. *FEBS Lett.* **553,** 3–10 (2003).

139. Jin, X., Foley, K. M. & Geiger, J. H. The structure of the 1L-*myo*-inositol-1-phosphate synthase-NAD$^{+}$-2-deoxy-D-glucitol 6-(*E*)-vinylhomophosphonate complex demands a revision of the enzyme mechanism. *J. Biol. Chem.* **279,** 13889–13895 (2004).

140. Geiger, J. H. & Jin, X. The structure and mechanism of *myo*-inositol-1-phosphate synthase. In *Biology of Inositols and Phosphoinositides* (eds. Majumder, A. L. & Biswas, B. B.) **39,** 157–180 (Springer, 2006).

141. Stieglitz, K. A., Yang, H., Roberts, M. F. & Stec, B. Reaching for mechanistic consensus across life kingdoms: structure and insights into catalysis of the myo-inositol-1-phosphate synthase (mIPS) from Archaeoglobus fulgidus. *Biochemistry* **44,** 213–224 (2005).

142. Norman, R. a *et al.* Crystal structure of inositol 1-phosphate synthase from *Mycobacterium tuberculosis*, a key enzyme in phosphatidylinositol synthesis. *Structure* **10,** 393–402 (2002).

143. Joint Center for Structural Genomics. Crystal structure of a myo-inositol-1-phosphate synthase-related protein (TM_1419) from Thermotoga maritima MSB8 at 1.70 Å resolution. (2008). doi:10.2210/pdb3CIN/pdb

144. Dastidar, K. & Chatterjee, A. Evolutionary divergence of L-*myo*-inositol 1-phosphate synthase: significance of a "core catalytic structure." In *Biology of Inositols and Phosphoinositols* (eds. Majumder, A. L. & Biswas, B. B.) 313–338 (Springer, 2006).

145. Basak, P. *et al.* An evolutionary analysis identifies a conserved pentapeptide stretch containing the two essential lysine residues for rice L-*myo*-inositol 1-phosphate synthase catalytic activity. *PLoS One* **12,** e0185351 (2017).

146. Atkinson, H. J., Morris, J. H., Ferrin, T. E. & Babbitt, P. C. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **4,** e4345 (2009).

147. Barber, A. E. & Babbitt, P. C. Pythoscape: A framework for generation of large protein similarity networks. *Bioinformatics* **28,** 2845–2846 (2012).

148. Zhao, S. *et al.* Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife* **3,** e03275 (2014).

149. Gerlt, J. a. *et al.* Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta - Proteins Proteomics* **1854,** 1019–1037 (2015).

150. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–504 (2003).

151. Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS One* **4,** e6441 (2009).

152. Thermo Fisher Scientific. GeneOptimizer process for successful gene optimization. (2018). Available at: www.thermofisher.com/us/en/home/life-science/cloning/gene-synthesis/geneart-gene-synthesis/geneoptimizer.html.

153. Bio-Rad. Image Lab software. (2018). Available at: www.bio-rad.com/en-us/product/image-lab-software?ID=KRE6P5E8Z.

154. Pei, J., Tang, M. & Grishin, N. V. PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36,** 30–34 (2008).

155. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26,** 1641–1650 (2009).

156. Han, M. V. & Zmasek, C. M. PhyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10,** 356 (2009).

157. Schrödinger, LLC. The PyMOL molecular graphics system, Version 1.8. (2015).
158. Pommié, C., Levadoux, S., Sabatier, R., Lefranc, G. & Lefranc, M. P. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.* **17,** 17–32 (2004).
159. Reimer, L. C. *et al.* BacDive in 2019: Bacterial phenotypic data for high-throughput biodiversity analysis. *Nucleic Acids Res.* gky879 (2018). doi:10.1093/nar/gky879
160. Orbović, V. & Poff, K. L. Effect of temperature on growth and phototropism of *Arabidopsis thaliana* seedlings. *J. Plant Growth Regul.* **26,** 222–228 (2007).
161. Kwon-Chung, K. J. & Sugui, J. A. *Aspergillus fumigatus*—What makes the species a ubiquitous human fungal pathogen? *PLoS Pathog.* **9,** e1003743 (2013).
162. ATCC. Cells and microorganisms. (2018). Available at: www.atcc.org/en/Products/Cells_and_Microorganisms.aspx.
163. Oplinger, E. S. *et al.* Sesame. in *Alternative Field Crops Manual* (University of Wisconsin Cooperative Extension Service, University of Minnesota Extension Service, and Center for Alternative Plant and Animal Products, 1990).
164. Morris, J. H. *et al.* ClusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* **12,** 436 (2011).
165. Nordberg, H. *et al.* The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **42,** 26–31 (2014).
166. Booth, W. T. *et al.* Impact of an N-terminal polyhistidine tag on protein thermal stability. *ACS Omega* **3,** 760–768 (2018).
167. Dougan, D. A., Truscott, K. N. & Zeth, K. The bacterial N-end rule pathway: Expect the unexpected. *Mol. Microbiol.* **76,** 545–558 (2010).
168. Chhetri, D. R., Adhikari, J. & Mukherjee, A. K. NAD$^+$ mediated differential thermotolerance between chloroplastic and cytosolic L-*myo*-inositol-1-phosphate synthase from *Diplopterygium glaucum* (Thunb.) Nakai. *Prep. Biochem. Biotechnol.* **36,** 307–319 (2006).
169. Hartl, D. L. What can we learn from fitness landscapes? *Curr. Opin. Microbiol.* **21,** 51–57 (2014).
170. Bailey, J. E. Toward a science of metabolic engineering. *Science* **252,** 1668–1675 (1991).
171. Abatemarco, J., Hill, A. & Alper, H. S. Expanding the metabolic engineering toolbox with directed evolution. *Biotechnol. J.* **8,** 1397–410 (2013).
172. Monterrubio, R., Baldoma, L., Obradors, N., Aguilar, J. & Badia, J. A common regulator for the operons encoding the enzymes involved in D-galactarate , D-glucarate , and D-glycerate utilization in *Escherichia coli. J. Bacteriol.* **182,** 2672–2674 (2000).
173. Roberton, A. M., Sullivan, P. A., Jones-Mortimer, M. C. & Kornberg, H. L. Two genes affecting glucarate utilization in *Escherichia coli* K12. *J. Gen. Microbiol.* **117,** 377–382 (1980).
174. Gonçalves, G. A. L., Prazeres, D. M. F., Monteiro, G. A. & Prather, K. L. J. De novo creation of MG1655-derived *E. coli* strains specifically designed for plasmid DNA production. *Appl. Microbiol. Biotechnol.* **97,** 611–620 (2013).
175. Amann, E., Ochs, B. & Abel, K. Tightly regulated *tac* promoter vectors useful for the expression of unfused and fused proteins in *Escherichia coli. Gene* **69,** 301–315 (1988).
176. Sampaio, M.-M. *et al.* Phosphotransferase-mediated transport of the osmolyte 2-O-α-mannosyl-D-glycerate in *Escherichia coli* occurs by the product of the *mngA* (*hrsA*) gene and is regulated by the *mngR* (*farR*) gene product acting as repressor. *J. Biol. Chem.* **279,**

5537–5548 (2004).

177. Rogers, J. K. & Church, G. M. Genetically encoded sensors enable real-time observation of metabolite production. *Proc. Natl. Acad. Sci.* **113,** 2388–2393 (2016).

178. Gama-Castro, S. *et al.* RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* **44,** D133–D143 (2016).

179. Shimada, T., Fujita, N., Yamamoto, K. & Ishihama, A. Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. *PLoS One* **6,** e20081 (2011).

180. Raman, S., Rogers, J. K., Taylor, N. D. & Church, G. M. Evolution-guided optimization of biosynthetic pathways. *Proc. Natl. Acad. Sci.* **111,** 17803–17808 (2014).

181. Lin, Z., Zhang, Y. & Wang, J. Engineering of transcriptional regulators enhances microbial stress tolerance. *Biotechnol. Adv.* **31,** 986–991 (2013).

182. Lemire, J., Alhasawi, A., Appanna, V. P., Tharmalingam, S. & Appanna, V. D. Metabolic defence against oxidative stress: The road less travelled so far. *J. Appl. Microbiol.* **123,** 798–809 (2017).

183. Cox, R. S., Surette, M. G. & Elowitz, M. B. Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.* **3,** (2007).

184. Tutukina, M. N., Potapova, A. V., Cole, J. A. & Ozoline, O. N. Control of hexuronate metabolism in *Escherichia coli* by the two interdependent regulators, ExuR and UxuR: Derepression by heterodimer formation. *Microbiology* **162,** 1220–1231 (2016).

185. Robert-Baudouy, J., Portalier, R. & Stoeber, F. Regulation of hexuronate system genes in *Escherichica coli* K-12: Multiple regulation of the *uxu* operon by *exuR* and *uxuR* gene products. *J. Bacteriol.* **145,** 211–220 (1981).

186. Hugouvieux-Cotte-Pattat, N. & Robert-Baudouy, J. Regulation of expression of the uxu operon and of the uxuR regulatory gene in Escherichia coli K12. *J. Gen. Microbiol.* **129,** 3345–3353 (1983).

187. Suvorova, I. A. *et al.* Comparative genomic analysis of the hexuronate metabolism genes and their regulation in gammaproteobacteria. *J. Bacteriol.* **193,** 3956–3963 (2011).

# Appendix A. Supplemental Information for Chapter 2

## A.1. Genomic *udh* expression and verification

Genomic integration of *udh* was performed to allow expression of the full glucaric acid pathway (*INO1*, *MIOX*, and *udh*) from fewer plasmids, minimizing metabolic burden. *Pseudomonas syringae* Udh, used in most prior glucaric acid work, was previously shown to have activity two orders of magnitude greater than that of either INO1 or MIOX.[59] *Agrobacterium tumefaciens udh* was selected for integration because the rate constant for the *A. tumefaciens* enzyme is more than twice that of the *P. syringae* enzyme.[106]

After construction as described in Materials and Methods, genomic expression was validated to ensure it was sufficiently high to convert glucuronic acid to glucaric acid in the context of the glucaric acid pathway. This was done by growing LG2512 (genomic *udh*) and LG1460 harboring pTATudh2 in LB supplemented with 10 g/L of glucuronic acid (pH 7). Neither strain can consume glucuronic acid for growth. Supernatant samples were taken at 0 hours and 72 hours, and glucuronic acid concentrations were measured by NADH generation at 340 nm by purified Udh. LG2512 converted 7.4 ± 0.2 g/L of glucuronic acid to glucaric acid, which is just slightly less than the 8.0 ± 0.2 g/L converted by LG1458 with pTATudh2. These are equivalent to 8.0 g/L and 8.6 g/L of glucaric acid production, respectively. This rate of conversion is sufficient for the pathway because the maximum 72 hour glucaric acid titer we have observed from glucose is about 2 g/L [61] and from myo-inositol is about 5 g/L.[63]

# Appendix B. Supplemental Information for Chapter 3

## B.1. Tables

Table B.1. Oligonucleotides used in Chapter 3

| Name | Sequence[a] |
|---|---|
| LG123 | aagcttgcggccgc |
| LG124 | gaattcggatcctggctgtg |
| LG125 | cacagccaggatccgaattcATGGTCAAGGTTTTGATCTTGG |
| LG126 | gcggccgcaagcttTTATCACAACCACTTTGGTTTTAAAC |
| LG127 | cacagccaggatccgaattcATGAAGGTTTGGTTAGTCGGT |
| LG128 | gcggccgcaagcttTTATCACTTCAAGTTAGAGTACCATTCC |
| LG129 | cacagccaggatccgaattcATGTCCGAGCACCAATCTT |
| LG130 | gcggccgcaagcttTTATCAACCAATAATAAATTCTTCCAATTGAG |
| LG131 | cacagccaggatccgaattcATGTTCATCGAATCTTTCAAGGTT |
| LG132 | gcggccgcaagcttTTATCACTTATATTCCATGATCATGTTGT |
| LG133 | cacagccaggatccgaattcATGAAGCAAGAGATTAAGCCAG |
| LG134 | gcggccgcaagcttTTATCAGGCCAAGTAGTCTGG |
| LG135 | cacagccaggatccgaattcATGGGTAAGGTCAGAGTCG |
| LG136 | gcggccgcaagcttTTATCAATCTTCGGAACCGATAATG |
| LG137 | cacagccaggatccgaattcATGAAGCCAACTAATAACTCTACTTTG |
| LG138 | gcggccgcaagcttTTATCAGTGACCGTTACCATTAGT |
| LG139 | cacagccaggatccgaattcATGTCTATTAGAGTTGCTATTGCC |
| LG140 | gcggccgcaagcttTTATCAGGCTCTCCAAACGG |
| LG141 | cacagccaggatccgaattcATGGAGGCTGCTGCTC |
| LG142 | gcggccgcaagcttTTATCAAGTGGTTGGCATTGG |
| LG143 | cacagccaggatccgaattcATGGAAAGAACCAACGTTAAGC |
| LG144 | gcggccgcaagcttTTATCAATCGATTTCTTCATCTGGTTC |
| LG145 | cacagccaggatccgaattcATGTTCATCGAATCCTTCAAGG |
| LG146 | gcggccgcaagcttTTATCACTTGTATTCCAAAATCATGTTG |
| LG147 | cacagccaggatccgaattcATGTTCATCGAAAACTTTAAGGTTGA |
| LG148 | gcggccgcaagcttTTATCACTTGTATTCCAAAATCATGTTG |
| LG149 | cacagccaggatccgaattcATGGCCCCACATGCTT |
| LG150 | gcggccgcaagcttTTATCAGAACAACTTATGTTCCAAAGTCAT |
| LG151 | cacagccaggatccgaattcATGACTGTTAATAAGGGTATTTCCATC |
| LG152 | gcggccgcaagcttTTATCACTTCAATCTTTCTTCGAAAC |
| LG153 | cacagccaggatccgaattcATGTCTTCCATTGACTTCAAATCTT |
| LG154 | gcggccgcaagcttTTATCAAGTCAAACGTTCCTCG |
| LG155 | cacagccaggatccgaattcATGTCTTCCGGTGCTAACA |
| LG156 | gcggccgcaagcttTTATCACCAGATTCTAGTCTCCAAC |
| LG157 | cacagccaggatccgaattcATGGGTTCCAAGAAGGTTAGA |
| LG158 | gcggccgcaagcttTTATCACTCAGCAGCGTCC |
| LG159 | cacagccaggatccgaattcATGTATTACTTCGACAGAGGTAAC |
| LG160 | gcggccgcaagcttTTATCAGTTAGTTCTGTCACCGT |

Table B.1. Oligonucleotides used in Chapter 3 (cont.)

| Name | Sequence[a] |
|---|---|
| LG161 | cacagccaggatccgaattcATGAAGACTAACATTGAACCAGC |
| LG162 | gcggccgcaagcttTTATCACATGGATTCAACCAATTCT |
| LG163 | cacagccaggatccgaattcATGTCTGATGTTAACCCAGCT |
| LG164 | gcggccgcaagcttTTATCATTCAGCACCAATAATGAAAG |
| LG165 | cacagccaggatccgaattcATGGCTTCTTCTGATTTCTTTCAA |
| LG166 | gcggccgcaagcttTTATCATTGTCTAGCATCAATTTCGG |
| LG167 | cacagccaggatccgaattcATGTCTTCCAGAAAGATCAGAGT |
| LG168 | gcggccgcaagcttTTATCAACCTTGTTCAGCAGG |
| LG169 | cacagccaggatccgaattcATGGGTTCTGTTAGAGTCGC |
| LG170 | gcggccgcaagcttTTATCATCTCTCAACTTCACCTCT |
| LG171 | cacagccaggatccgaattcATGGTTAAGGTTGTCATTTTGGG |
| LG172 | gcggccgcaagcttTTATCACAACCATCTAGGTTTTAAACC |
| LG173 | cacagccaggatccgaattcATGACTACTGATTCTTACTTCACC |
| LG174 | gcggccgcaagcttTTATCACTTTAATCTTTCTTCAAAACGC |
| LG175 | cacagccaggatccgaattcATGACTGGTAGAATTAAGGTTGG |
| LG176 | gcggccgcaagcttTTATTCACCAGCAACGAACTT |
| LG177 | cacagccaggatccgaattcATGGATAAGATTAAGATTGCTATTGTTG |
| LG178 | gcggccgcaagcttTTATCTTTCTCTTTCACCAGCAATA |
| LG179 | cacagccaggatccgaattcATGGCTGACAGAAAAATTAGAGTT |
| LG180 | gcggccgcaagcttTTATCTTTCTCTTTCACCTCTAATGAAT |
| LG181 | cacagccaggatccgaattcATGACTTACCAAACTGGTGTTTTAT |
| LG182 | gcggccgcaagcttTTAGGCGTTGTATTCAAAGTGC |
| LG183 | cacagccaggatccgaattcATGCACTCCAGATTGCAAG |
| LG184 | gcggccgcaagcttTTAAGCGTAAGCCTTATCGTC |
| LG185 | cacagccaggatccgaattcATGAACGGTTACGCTAACG |
| LG186 | gcggccgcaagcttTTAGTTAGCTTTTGGTAATTGAGTGA |
| LG220 | cacagccaggatccgaattcGGTCAAGGTTTTGATCTTGGGTC |
| LG221 | cacagccaggatccgaattcGAAGGTTTGGTTAGTCGGTGC |
| LG222 | cacagccaggatccgaattcGTCCGAGCACCAATCTTTGC |
| LG223 | cacagccaggatccgaattcGTTCATCGAATCTTTCAAGGTTGAATCTC |
| LG224 | cacagccaggatccgaattcGGCCCCACATGCTTCTTC |
| LG225 | cacagccaggatccgaattcGAAGCAAGAGATTAAGCCAGCTAC |
| LG226 | cacagccaggatccgaattcGACTGTTAATAAGGGTATTTCCATCAGAGT |
| LG227 | cacagccaggatccgaattcGTCTTCCATTGACTTCAAATCTTCTAAGTC |
| LG228 | cacagccaggatccgaattcGGGTAAGGTCAGAGTCGCC |
| LG229 | cacagccaggatccgaattcGTCTTCCGGTGCTAACACTC |
| LG230 | cacagccaggatccgaattcGAAGCCAACTAATAACTCTACTTTGGAAG |
| LG231 | cacagccaggatccgaattcGTCTATTAGAGTTGCTATTGCCGG |
| LG232 | cacagccaggatccgaattcGGAGGCTGCTGCTCAATTC |

Table B.1. Oligonucleotides used in Chapter 3 (cont.)

| Name | Sequence[a] |
|---|---|
| LG233 | cacagccaggatccgaattcGGGTTCCAAGAAGGTTAGAGTCG |
| LG234 | cacagccaggatccgaattcGTATTACTTCGACAGAGGTAACGTCAT |
| LG235 | cacagccaggatccgaattcGAAGACTAACATTGAACCAGCTGAAG |
| LG236 | cacagccaggatccgaattcGTCTGATGTTAACCCAGCTGC |
| LG237 | cacagccaggatccgaattcGGCTTCTTCTGATTTCTTTCAAGAACC |
| LG238 | cacagccaggatccgaattcGGAAAGAACCAACGTTAAGCCAG |
| LG239 | cacagccaggatccgaattcGTTCATCGAATCCTTCAAGGTTGAATC |
| LG240 | cacagccaggatccgaattcGTCTTCCAGAAAGATCAGAGTCGC |
| LG241 | cacagccaggatccgaattcGGGTTCTGTTAGAGTCGCTATTGT |
| LG242 | cacagccaggatccgaattcGGTTAAGGTTGTCATTTTGGGTCAAG |
| LG243 | cacagccaggatccgaattcGTTCATCGAAAACTTTAAGGTTGAATGTCC |
| LG244 | cacagccaggatccgaattcGACTACTGATTCTTACTTCACCCCATC |
| LG245 | cacagccaggatccgaattcGACTGGTAGAATTAAGGTTGGTTTGG |
| LG246 | cacagccaggatccgaattcGGATAAGATTAAGATTGCTATTGTTGGTGTTG |
| LG247 | cacagccaggatccgaattcGGCTGACAGAAAAATTAGAGTTGCTATC |
| LG248 | cacagccaggatccgaattcGACTTACCAAACTGGTGTTTTATTCGTTG |
| LG249 | cacagccaggatccgaattcGCACTCCAGATTGCAAGATAGAAG |
| LG250 | cacagccaggatccgaattcGAACGGTTACGCTAACGGTAC |
| LG251 | acctccaatctgttcgcgg |
| LG252 | gcggccgcaagctt |
| LG253 | cgcgaacagattggaggtACAGAAGATAATATTGCTCCAATCACC |
| LG260 | cgcgaacagattggaggtGTCAAGGTTTTGATCTTGGGTCAA |
| LG261 | cgcgaacagattggaggtTTCATCGAATCTTTCAAGGTTGAATCTC |
| LG262 | cgcgaacagattggaggtGAGGCTGCTGCTCAATTCTT |
| LG263 | cgcgaacagattggaggtTTCATCGAATCCTTCAAGGTTGAATC |
| LG264 | cgcgaacagattggaggtACTACTGATTCTTACTTCACCCCATC |
| LG265 | cgcgaacagattggaggtAACGGTTACGCTAACGGTACT |
| LG271 | taccgaggccattaaagtggagccattgttgcc |
| LG272 | ggcaacaatggctccactttaatggcctcggta |
| LG273 | ttcgccaataccgagcccactaaagtggagc |
| LG274 | gctccactttagtgggctcggtattggcgaa |
| LG275 | gactgcaaatactgagaggttcgtagaagtatctcctg |
| LG276 | caggagatacttctacgaacctctcagtatttgcagtc |
| LG277 | ctcgtccattgcccttttttgagtccccgacggg |
| LG278 | cccgtcggggactcaaaaagggcaatggacgag |
| LG279 | tatagatctgcgtcattgatgtcccaaccagagacg |
| LG280 | cgtctctggttgggacatcaatgacgcagatctata |
| LG281 | gtagcaccctgaccggaggcgttattgcaaa |
| LG282 | tttgcaataacgcctccggtcagggtgctac |

Table B.1. Oligonucleotides used in Chapter 3 (cont.)

| Name | Sequence[a] |
|---|---|
| LG283 | ggtggttgggatattagcaatatgaatctggcagacg |
| LG284 | cgtctgccagattcatattgctaatatcccaaccacc |
| LG285 | gtggtctgccaacactgaacgttattccgacatcgttgaag |
| LG286 | cttcaacgatgtcggaataacgttcagtgttggcagaccac |
| LG73 | ggcgctatcatgccataccg |
| LG74 | gattatgcggccgtgtacaatacg |
| LG206 | ggcttcttctgatttctttcaagaacc |
| LG266 | cagccaggatccgaattcg |

[a] Homologous regions to MIPS genes are capitalized.

Table B.2. MIPS DNA sequences from JGI

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Thermotoga maritima* | Q9X1D6 | ATGGTCAAGGTTTTGATCTTGGGTCAAGGTTATGTCGCTTCCACTTTCGTCGCTGGTTTGGAAAAATTGCGTAAGGGTGAAATCGAACCATACGGTGTTCCATTAGCCCGTGAATTGCCAATTGGTTTCGAAGACATTAAAATTGTTGGTTCTTACGACGTTGATAGAGCTAAGATTGGTAAGAAATTGTCTGAAGTCGTTAAGCAATACTGGAACGATGTTGATTCTTTAACTTCCGACCCAGAAATTAGAAAGGGTGTTCACTTGGGTTCCGTCAGAAATTTGCCAATTGAAGCCGAAGGTTTAGAAGATTCTATGACCTTAAAGGAAGCTGTTGATACCTTGGTTAAAGAATGGACTGAATTGGACCCAGACGTTATCGTTAATACCTGTACCACTGAAGCCTTTGTTCCATTTGGTAACAAGGAAGACTTATTAAAGGCTATTGAAAATAATGACAAGGAAAGATTGACCGCTACCCAAGTCTACGCTTACGCCGCCGCTTTGTATGCTAACAAGAGAGGTGGTGCTGCTTTTGTTAACGTTATTCCAACTTTCATTGCTAACGACCCAGCTTTCGTCGAGTTGGCTAAAGAAAACAACTTAGTCGTCTTCGGTGACGATGGTGCTACTGGTGCTACTCCATTTACTGCTGATGTCTTATCCCATTTGGCCCAAAGAAACCGTTACGTTAAAGACGTCGCTCAATTTAACATTGGTGGTAATATGGACTTTTTGGCTTTAACTGACGATGGTAAGAACAAATCCAAGGAATTCACTAAGTCTTCTATTGTCAAGGACATTTTGGGTTACGACGCTCCACATTATATTAAGCCAACCGGTTACTTAGAACCATTGGGTGACAAAAAATTCATTGCTATTCATATCGAATACGTTTCTTTCAATGGTGCTACTGATGAATTGATGATTAACGGTAGAATTAATGACTCTCCAGCTTTGGGTGGTTTGTTAGTCGACTTGGTTAGATTGGGTAAGATTGCTTTGGATAGAAAGGAATTCGGTACTGTTTACCCAGTTAACGCTTTCTACATGAAGAACCCTGGTCCAGCTGAAGAAAAGAACATCCCACGTATTATCGCTTACGAAAAGATGAGAATTTGGGCCGGTTTAAAACCAAAGTGGTTGTGATAA |
| *Archaeoglobus fulgidus* | A0A075WEG3 | ATGAAGGTTTGGTTAGTCGGTGCCTACGGTATCGTTTCTACCACTGCCATGGTCGGTGCCCGTGCTATTGAAAGAGGTATTGCTCCAAAGATCGGTTTGGTTTCTGAATTGCCACACTTCGAAGGTATTGAAAAATATGCTCCATTCTCTTTCGAATTCGGTGGTCACGAAATTAGATTGTTATCTAACGCTTATGAGGCCGCTAAGGAACACTGGGAGTTGAACAGACACTTCGATAGAGAAATCTTGGAAGCCGTCAAGTCCGATTTGGAAGGTATCGTTGCCAGAAAGGGTACTGCCTTGAATTGTGGTTCCGGTATCAAAGAATTGGGTGATATCAAGACCTTGGAAGGTGAAGGTTTGTCCTTGGCCGAAATGGTCTCCAGAATTGAAGAAGATATTAAGTCCTTTGCCGATGACGAAACTGTTGTTATTAATGTTGCTTCTACCGAACCATTGCCAAACTACTCTGAAGAATACCACGGTTCTTTGGAGGGTTTCGAACGTATGATTGACGAAGACAGAAAGGAATACGCCTCCGCCTCCATGTTGTACGCTTACGCTGCTTTGAAGTTGGGTTTACCATACGCTAACTTTACCCCATCTCCTGGTTCCGCTATCCCAGCTTTGAAAGAATTGGCTGAAAAGAAGGGTGTTCCTCACGCCGGTAACGATGGTAAAACCGGTGAAACCTTGGTTAAGACTACCTTGGCTCCAATGTTTGCTTACAGAAACATGGAAGTTGTTGGTTGGATGTCTTACAACATTTTGGGTGATTACGATGGTAAAGTCTTGTCTGCTAGAGACAACAAGGAATCCAAGGTTTTGTCTAAGGACAAAGTCTTGGAAAAGATGTTAGGTTACTCTCCATACTCTATTACCGAAATCCAATATTTCCCATCCTTGGTTGATAACAAGACCGCCTTCGATTTTGTCCATTTCAAGGGTTTCTTAGGTAAGTTAATGAAGTTCTACTTCATTTGGGATGCTATCGACGCTATTGTCGCCGCTCCTTTGATTTTAGACATCGCCAGATTCTTGTTGTTTGCTAAGAAGAAAGGTGTTAAGGGTGTTGTTAAAGAAATGGCTTTCTTTTTCAAGTCTCCTATGGACACTAACGTCATCAACACTCACGAACAATTTGTTGTCTTAAAGGAATGGTACTCTAACTTGAAGTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Mycobacterium tuberculosis* | P9WKI1 | ATGTCCGAGCACCAATCTTTGCCAGCCCCAGAAGCTTCCACTGAAGTTAGAGTCGCCATCGTCGGTGTCGGTAACTGTGCTTCCTCTTTGGTTCAAGGTGTTGAGTACTATTATAATGCTGATGATACTTCTACCGTTCCAGGTTTGATGCATGTCAGATTTGGTCCTTACCACGTTAGAGACGTCAAATTCGTTGCCGCTTTTGACGTTGATGCCAAGAAGGTTGGTTTTGACTTGTCTGATGCTATCTTCGCCTCCGAAAACAATACTATTAAGATCGCTGATGTTGCTCCAACTAACGTCATTGTTCAAAGAGGTCCAACTTTGGATGGTATCGGTAAATACTACGCCGACACTATTGAATTGTCCGATGCTGAACCAGTCGATGTTGTTCAAGCTTTAAAGGAAGCTAAGGTTGACGTTTTGGTTTCCTACTTGCCAGTCGGTTCTGAAGAAGCCGACAAATTCTACGCTCAATGTGCTATCGATGCTGGTGTCGCCTTCGTTAACGCTTTGCCAGTTTTTATTGCTTCTGACCCAGTTTGGGCTAAAAAGTTCACTGATGCTAGAGTCCCTATCGTCGGTGACGACATCAAATCTCAAGTCGGTGCTACTATTACTCACAGAGTTTTGGCTAAATTGTTCGAAGACAGAGGTGTTCAATTAGATCGTACTATGCAATTGAACGTCGGTGGTAATATGGATTTCTTGAACATGTTGGAAAGAGAAAGATTGGAATCTAAGAAGATCTCTAAGACTCAAGCCGTTACTTCTAACTTGAAGAGAGAATTCAAGACCAAAGACGTTCACATCGGTCCATCTGACCACGTTGGTTGGTTGGATGATAGAAAGTGGGCTTACGTTAGATTGGAAGGTAGAGCTTTTGGTGATGTCCCATTGAATTTGGAATACAAGTTAGAGGTTTGGGATTCTCCAAACTCTGCCGGTGTTATCATCGATGCTGTTAGAGCCGCTAAGATTGCTAAAGATAGAGGTATTGGTGGTCCTGTTATTCCAGCTTCTGCCTACTTGATGAAGTCTCCACCAGAACAATTGCCAGACGACATCGCCAGAGCTCAATTGGAAGAATTTATTATTGGTTGATAA |
| *Arabidopsis thaliana* | Q38862 | ATGTTCATCGAATCTTTCAAGGTTGAATCTCCAAACGTTAAATACACTGAAAACGAAATTAACTCTGTCTACGATTACGAAACTACTGAAGTTGTCCACGAAAACCGTAATGGTACCTATCAATGGGTTGTCAAACCAAAGACTGTTAAGTACGACTTCAAGACTGACACCAGAGTCCCAAAGTTGGGTGTCATGTTGGTTGGTTGGGGTGGTAATAACGGTTCTACCTTAACTGCTGGTGTCATCGCCAACAAAGAAGGTATTTCTTGGGCTACCAAGGATAAGGTTCAACAAGCTAACTACTTCGGTTCTTTAACTCAAGCTTCTTCCATTAGAGTTGGTTCTTACAACGGTGAGGAAATCTACGCTCCTTTCAAGTCTTTATTGCCAATGGTTAACCCAGAAGATGTCGTCTTTGGTGGTTGGGATATCTCTGACATGAATTTGGCCGATGCTATGGCCAGAGCTAGAGTCTTAGACATCGACTTGCAAAAACAATTAAGACCTTACATGGAAAACATGATCCCATTGCCAGGTATTTACGACCCAGATTTCATTGCTGCTAATCAAGGTTCCAGAGCCAATTCTGTTATTAAGGGTACCAAGAAGGAACAAGTTGATCATATCATCAAGGATATGAGAGAATTCAAGGAAAAGAACAAGGTTGATAAATTGGTTGTCTTGTGGACTGCTAACACCGAAAGATACTCCAACGTTATTGTTGGTTTGAACGATACTACCGAAAACTTGTTAGCCTCCGTCGAAAAGGACGAATCTGAAATCTCCCCATCTACTTTGTATGCTATTGCTTGTGTTTTGGAAGGTATTCCATTCATCAACGGTTCTCCACAAAACACTTTCGTTCCAGGTTTAATTGAATTGGCCATCTCTAAGAACTGTTTAATCGGTGGTGATGATTTTAAGTCCGGTCAAACTAAGATGAAGTCCGTCTTAGTTGACTTCTTGGTCGGTGCCGGTATCAAACCAACTTCTATCGTTTCTTACAATCACTTGGGTAACAACGATGGTATGAACTTATCTGCTCCACAAACCTTTAGATCTAAGGAAATCTCTAAATCCAACGTTGTTGACGACATGGTTGCTTCTAATGGTATTTTATTCGAGCCAGGTGAACACCCAGACCATGTCGTTGTCATTAAGTACGTTCCATACGTCGCTGATTCCAAAAGAGCTATGGACGAATATACCTCTGAAATTTTCATGGGTGGTAGAAACACCATCGTTTTGCACAATACTTGTGAAGATTCTTTGTTGGCCGCTCCAATCATTTTAGATTTGGTTTTGTTGGCTGAATTATCTACTCGTATTCAATTCAAGGCTGAAGGTGAAGGTAAGTTTCACTCTTTTCACCCAGTTGCTACTATTTTATCCTACTTGACTAAGGCTCCATTGGTTCCACCAGGTACCCCAGTTGTCAACGCCTTGTCTAAGCAAAGAGCTATGTTGGAAAACATCTTGAGAGCTTGTGTTGGTTTGGCTCCAGAAAACAACATGATCATGGAATATAAGTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Aspergillus clavatus* | A1CFT5 | ATGGCCCCACATGCTTCTTCCGATGTTGCTGCCAACGGTGCCGTCAACGGTTCCGCTCGTGCTACCTCCGCCCCATTGTTCACTGTCGCTTCCCCAAATGTCGAATACACTGACAACGAAATTAAATCTCAATATGCCTACCACACTACTGAAATTACCAGAAACGCTGACGGTAAGTGGGTTGCTACTCCAAAAGTCACTAACTACCAATTCAAGGTTGACCGTAAGGTTGGTAAGGTTGGTATGATGTTAGTTGGTTGGGGTGGTAACAACGGTTCTACCGTCACCGCTGGTATCATTGCTAACAGAAGAAACTTGTCCTGGGAGACCAGAGAAGGTGAAAGAGCTTCTAACTATTACGGTTCTGTCGTTATGTCTTCTACCGTTAAATTAGGTACCGAGACCAAGACTGGTGAAGAGATCAACATCCCATTCCACGATTTATTGCCAATGGTCCACCCAAACGACTTGGTTATTGGTGGTTGGGACATCTCTTCCTTGAACTTGGCTGAATCTATGGATAGAGCTCAAGTCTTAGAACCAACCTTGAAGCAATTGGTTAGAAAGGAAATGGCTGAAATGAAACCATTGCCTTCTATTTATTACCCAGATTTCATTGCTGCTAATCAAGAAGACAGAGCTGACAATGTTATCGAAGGTGATAAGGCTTGTTGGGCTCATGTTGAAAGAATCCAAAAGGATATGCGTGATTTCAAAACCCAACATGGTTTGGATAAGGTTATTGTCATGTGGACTGCCAATACCGAACGTTACGCTGATATCTTGCCAGGTATTAACGACACTGCCGACAACTTGTTGAATGCTATCAAGAACGGTCACGAAGAAGTTTCTCCATCTACTGTTTTCGCTGTTGCTTGTATTTTGGACAACGTTCCATTTATCAATGGTTCCCCACAAAACACTTTCGTTCCAGGTGCTATCCAATTAGCCGAAAAGCATAACGCTTTCATCGGTGGTGACGATTTCAAGTCCGGTCAAACCAAGATGAAGTCCGCTTTGGTTGATTTTTTGATTAACGCTGGTATTAAATTGACTTCTATCGCTTCTTACAACCACTTGGGTAATAACGACGGTAAGAATTTGTCCTCCCAAAAGCAATTCCGTTCTAAGGAAATTTCTAAGTCTAACGTTGTCGATGACATGGTCGCCGCCAACAACATTTTGTACAAGGAAGGTGAACACCCTGATCACACCGTTGTTATCAAGTACATGCCAGCTGTTGGTGATAACAAAAGAGCTTTAGACGAGTACTACGCTGAAATTTTCATGGGTGGTCATCAAACTATCTCTTTGTTCAATATTTGTGAGGACTCTTTGTTAGCCTCCCCATTGATCATCGACTTGGTCGTCATCGCTGAAATGATGACCAGAATTTCTTGGAAGTCTGCTGAAGAGGCCGACTACAAAGGTTTCCACTCCGTCTTATCCATTTTATCCTATATGTTAAAAGCCCCATTGACCCCACCAGGTACCCCTGTTGTCAATGCTTTGGCTAAGCAAAGATCTGCCTTGACCAACATTTTCCGTGCTTGTGTTGGTTTGCAACCAGACTCTGAAATGACTTTGGAACATAAGTTGTTCTGATAA |
| *Bacteroides thetaiotaomicron* | D7IFW4 | ATGAAGCAAGAGATTAAGCCAGCTACTGGTAGATTGGGTGTCTTAGTCGTTGGTGTCGGTGGTGCTGTCGCTACTACCATGATCGTCGGTACTTTGGCTTCCCGTAAGGGTTTGGCCAAACCAATCGGTTCTATTACTCAATTGGCTACCATGAGAATGGAAAACAACGAGGAAAAGTTGATTAAGGATGTTGTTCCATTGACCGACTTGAACGATATTGTCTTCGGTGGTTGGGACATTTTCCCTGACAACGCTTATGAAGCTGCCATGTACGCTGAAGTCTTGAAGGAAAAGGACTTAAACGGTGTTAAAGATGAATTGGAAGCCATCAAACCAATGCCAGCTGCTTTCGATCACAATTGGGCCAAACGTTTAAACGGTACTCACATTAAGAAGGCTGCCACTAGATGGGAAATGGTCGAGCAATTAAGACAAGACATTCGTGATTTCAAGGCTGCCAACAATTGTGAAAGAGTTGTTGTTTTATGGGCTGCTTCCACCGAAATTTACATCCCATTATCTGATGAACATATGTCTTTGGCTGCTTTGGAAAAGGCTATGAAGGACAACAACACCGAAGTCATTTCTCCATCTATGTGTTACGCTTACGCTGCCATCGCCGAAGATGCTCCATTCGTTATGGGTGCTCCAAACTTATGTGTCGATACCCCTGCCATGTGGGAGTTCTCTAAGCAAAAAAACGTCCCTATCTCTGGTAAAGACTTCAAGTCTGGTCAAACCTTAATGAAAACTGTCTTAGCTCCAATGTTCAAGACTAGAATGTTGGGTGTTAACGGTTGGTTCTCCACCAACATCTTGGGTAACAGAGATGGTGAAGTTTTGGACGACCCAGATAACTTCAAGACTAAGGAAGTTTCTAAGTTGTCTGTCATTGACACTATTTTCGAACCAGAAAAGTACCCAGACTTATACGGTGACGTCTATCACAAGGTTAGAATTAATTACTATCCTCCAAGAAAGGATAACAAGGAAGCTTGGGACAATATTGATATCTTTGGTTGGATGGGTTACCCAATGGAGATTAAAGTTAACTTTTTGTGTAGAGACTCTATCTTGGCTGCTCCAATCGCCTTGGATTTGGTTTTATTCTCTGACTTGGCTATGAGAGCTGGTATGTGTGGTATTCAAACTTGGTTGTCCTTTTTCTGTAAGTCCCCAATGCACGATTTCGAACACCAACCAGAACACGACTTATTTACTCAATGGAGAATGGTTAAACAAACTTTGAGAAACATGATCGGTGAAAAGGAACCAGACTACTTGGCCTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Candida glabrata* | Q6FQI1 | ATGACTGTTAATAAGGGTATTTCCATCAGAGTCAACAACGTTGGTGATAAGGTTTCTTACAAGGAAAACGAATTGTTAACTAACTACACTTACCACACCAATGTTGTTCACACTAACTCTGACAAAACTCAATTTGAAGTCACTCCATTGGATAAGAACTATCAATTTAAGGTCGATTTAAACAAACCAGAAAGATTGGGTGTTATGTTGGTTGGTTTGGGTGGTAATAACGGTTCTACTATGATGGCCGCTGTCTTGGCTAACAAGCACAACGTTTGTTTCAGAACTCGTGACAAGGAAGGTTTAACCGAGCCTAACTACTATGGTTCTTTGACCCAATCTTCTACTATCAAGTTGGGTGTCGATTCTAAGGGTAAGGATGTTTACGTTCCATTTAACTCTTTGGTTCCTATGGTCAACCCAAATGATTTCGTTGTCTCTGGTTGGGATATCAACGGTGCTACCATGGATCAAGCTATGGAAAGAGCCTCTGTTTTGGAAGTCGACTTGAGAAACAAGTTGGCTCCAATGATGAAGGATCACAAACCATTAAAGTCTGTCTACTACCCAGACTTTATCGCCGCTAATCAAGATGAGAGAGCTGACAACTGTTTGAACGTTGACCCTCAAACTGGTAAGGTCACCACCACCGGTAAGTGGGAACATTTAAATCACATCCGTAATGACATCCGTACCTTCAAGCAACAAAACGACTTGGACAAGGTTATCATTTTATGGACTGCTAATACTGAACGTTATGTTGAGATCTTGCCAGGTGTTAACGATACTATGGAAAACTTGTTGGAAGCTATCAAGAACGACCACACTGAAATTGCTCCATCTACCATTTTTGCTGCCGCCTCCATCTTAGAACACTGTCCTTACATCAACGGTTCCCCTCAAAACACCTTTGTTCCAGGTTTGATCGAATTGGCTGAAAAGAACGACTCTTTGATCGCTGGTGACGATTTCAAGTCCGGTCAAACTAAAATGAAGTCTGTTTTGGCTCAATTTTTGGTCGACGCTGGTATCCGTCCTGTTTCCATTGCTTCTTATAACCATTTGGGTAACAACGACGGTTACAACTTGTCTTCTCCACAACAATTCAGATCTAAGGAAATTTCTAAGGCTTCCGTCGTCGACGACATCATTGAATCTAACCCAATCTTGTACAACGATAAGTTGGGTAACAAGATTGATCACTGTATCGTTATCAAGTACATGCACGCTGTTGGTGACTCTAAGGTCGCTATGGATGAATACTACTCCGAATTGATGTTGGGTGGTCATAATAGAATTTCTATTCATAACGTTTGTGAAGATTCTTTGTTGGCCACCCCATTGATTATTGACTTAATTGTTATGACCGAATTCTGTTCCAGAGTTACCTACAGAAATGTCGACGGTCAAGATGGTGCTGAAGCTAAGGGTGACTTCGAGAACTTCTACCCTGTTTTATCTTTCTTGTCTTACTGGTTGAAGGCCCCTTTGACTAAGCCAGGTTACCAACCAATTAACGGTTTGAACAAACAAAGAACTGCTTTAGAAAATTTTTTAAGATTGTTGATTGGTTTACCAGCTATTGATGAATTGCGTTTCGAAGAAAGATTGAAGTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Candida orthopsilosis* | H8X4H9 | ATGTCTTCCATTGACTTCAAATCTTCTAAGTCTGTCACCAAAGACGATCATTTGTATACTAAGTTCACTTACGAAAATTCTTTGGTCGAAAAGGATGCCCAAGGTAAGTTCCATGTTACTCCTACTTCTGTTGACTACGATTTCAAATTGGACTTGAAAGTTCCAAAAGTTGGTTTATTATTGGTCGGTTTAGGTGGTAACAACGGTACCACCTTAATGGCCTCCATCTTGGCCAACAAGCACAACATTTCCTTTGAAAACAAGGAAGGTGTCGTTAAGCCTAATTACTACGGTTCTGTTACCCAATCTTCTACTATTAAGATTGGTGTTGACGCTGAAGGTAACGATGTTTACGCTCCATTCAACTCTATCGTTCCTTTTGTTAATCCAAATGACTTGGTTGTTGACGGTTGGGATATTTCCGGTTTGGAATTGGATCAAGCTATGAAAAGAGCTAAGGTCTTGGACGTTACCTTGCAAAAGCAATTGGCTCCACACTTGCAAGGTAAGAAGCCAATGGAATCCATTTATTACCCAGACTTCATCGCTGCCAATCAAGGTGATAGAGCTGATAACGTCTTCAACAAGGTTAACGGTGAAATTAAGACCGACGACAAATGGAAGGACGTCGAAAAGATCAGAAAGGACATTAGAGATTTCAAACAAAAGAACGGTTTGGATAAGGTCATCGTTTTGTGGACCGCTAATACCGAACGTTACGCTGACGTTTTGCCAAAAGTTAACGATACTGCCGATAACTTGATCGCCTCTATTAAATCCAATCACGAAGAAATTGCTCCATCCACCATTTTCGCCGTTGCTTCTATCTTGGAGAACGTTCCATACATCAACGGTTCTCCACAAAACACTTTCGTCCCAGGTGTCATTGAATTAGCTGAAAAACACCATTCTTTCATCGGTGGTGACGATTTTAAATCTGGTCAAACTAAGATCAAGTCTGTCTTGGCTCAATTCTTAGTCGACGCTGGTATTAAGCCAATTTCTATTGCTTCCTACAATCACTTGGGTAACAACGATGGTTACAATTTGTCCGCTCCTAAGCAATTCCGTTCCAAGGAAATCTCCAAACAATCCGTTGTCGATGACATGATCGAATCTAACGAAATCTTGTACAACAAGGAGACCGGTGACAAGGTTGACCATTGCATTGTCATTAAGTACTTGCCAGCTGTTGGTGACTCTAAGGTTGCCATGGACGAATACTACTCCGAGTTAATGTTGGGTGGTCATAACAAAATTTCCATTCACAACGTTTGTGAAGATTCTTTGTTGGCTACTCCATTGATTATCGACTTAGTTGTCGTTACCGAATTCTTGCAACGTGTTCAATACAAAAAAATCCCAAGATTCTGAAGACAAGTACCACGACTTCTACGCTGTTTTAACTTTGTTGTCTTATTGGTTGAAAGCCCCTTTGTCTCGTCCTGGTTTCAAGACCATTAACGGTTTGAATAAGCAAAGACAAGCCTTGGAAAACTTGTTGAGATTATTGGTTGGTTTGCCTATCAACAATGAATTGAGATTCGAGGAACGTTTGACTTGATAA |
| *Corynebacterium halotolerans* | M1P1K8 | ATGGGTAAGGTCAGAGTCGCCATCGCTGGTGTCGGTAACTGTGCTGCCTCTTTGGTTCAAGGTGTCGAGTTCTACAGAGACACCCCAGTTGAGGAAAAGGTTCCAGGTTTGATGCACGTTGCCTTTGGTGAATACCACGTTTCTGATGTCGAATTTGTTGCTGCTTTTGATGTCGATGCTGAAAAGGTTGGTAGAGATTTGGCTGAAGCCTTGGATGCTTCTGAAAACTGTACTATTAAGATCGCCGACGTCCCTACCACCGGTGTTACCGTTCAACGTGGTCCTACTTTGGATGGTTTGGGTAGACACTACAGAGAAACTGTCACCGAATCTACTGCTGAACCAGTTGATGTTGCTCAAGCCTTGAGAGACGCTGAAGTTGACGTTTTGGTCTCCTACTTGCCAGTTGGTTCCGAACAAGCTGATAAGTTCTACGCCAGAGCTGCTTTGGATGCTGGTGTCGCTTTTGTCAACGCTTTGCCAGTCTTTATCGCTTCTGATCCAGAATGGGCCCAAAAGTTTGTTGACGCCGGTTTACCAATTGTCGGTGATGATATCAAGTCTCAAGTCGGTGCTACTATTACCCACAGAGTTATGGCTAAGTTGTTCGAGGATAGAGGTGTCAGATTGGAAAGAACTATGCAATTGAACGTTGGTGGTAACATGGACTTCAAGAACATGTTGGACAGAGACCGTTTAGAATCTAAAAAAATCTCCAAAACTCAAGCCGTCACCTCTAATTTGCACGAATCTCCATTGGCTGGTAAAGTCTCCGACAGAAATGTTCACATCGGTCCATCCGATTACGTTGAATGGTTGGACGACAGAAAGTGGGCTTACGTTAGATTGGAAGGTAGAGCTTTCGGTGAAGTTCCATTAAACTTGGAATATAAGTTAGAAGTCTGGGACTCTCCAAACTCTGCCGGTATCATCATCGACGCTGTCAGAGCTGCTAAGATCGCTTTGGACAGAGGTGTTGCCGGTCCAGTTTTGCCAGCTTCCGCTTACTTGATGAAGTCCCCACCAGTTCAATTGGGTGATGATGAAGCCAGAGCTCAATTGGAAGCCTTCATTATCGGTTCCGAAGATTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Dichomitus squalens* | R7SX42 | ATGTCTTCCGGTGCTAACACTCCAGAATCTCAATTGGAATCTGTTTTGCCAGTTCACCCAACCGCTGTTAGAAGAGCTTCTCCAATTGTTGTTCAATCCGAACACACCTCCTACACTAACGATCACATTATTTCCAAATTCACCAACAGAGGTGCTGACGTCACTATCGTTGAAGGTCAATACATCGTTACCCCAACTGCCAAGCCATACGAATTCCAAACCGCTAGAAAGGTTGCTAAGACTGGTTTGATGATGGTCGGTTGGGGTGGTAACAACGGTTCCACTTTGTCTGCCACCATTTTAGCTAATCGTCACAACATTGTCTGGAGAACTAAGTCCGGTGTCCAACAACCTAACTACATTGGTTCCTTATTAAGAGCCTCCACTGTTAGATTGGGTGCTGACCCATCTACCGGTAAGGATGTTTACGTTCCTATCTCCGATGTTTTGCCTATGGTTCATCCAAACGACTTAGTCTTAGGTGGTTGGGATATCTCTGGTGCTAGATTGGACGAAGCTATGAAGAGAGCTCAAGTTTTGGATTGGGATTTACAAAGACAAGTTATGCCACATATGGCCGCTTTGGGTTCCCCATTGCCATCTATTTATTACCCAGACTTCATCGCTGCCAATCAAGAAGCTAGAGCCGACAACGTTGTTCCAGGTACCGATAAACAAGCCCACTTGGAACACTTAAGAGCCGACATCAGAAAATTCAAAGAAACTCACGGTTTAGACAGAGTTGTTGTCTTTTGGACTGCCAATACCGAAAGATATTCCGACATCATCCCAGGTGTCAACGACACCGCTGATAACTTGTTGAACGCTATTAAAGCTTCTCATTCTGAAGTCTCTCCTTCCACTTTGTTTGCTGTTGCCGCCATTTTGGAAGGTGAACCATTCGTTAACGGTGCCCCACAAAACACTTTCGTTCCAGGTGTTATCGAATTAGCCGAAAGATTGCAATCCTTTATCGGTGGTGATGATTTGAAGTCCGGTCAAACTAAGTTGAAGTCTGTTTTCGCCGAATTTTTAGTCAACGCTGGTATTAAGCCATTGTCCATTGCTTCTTACAACCACTTGGGTAACAACGATGGTCATAACTTGTCCGCCGAACCACAATTCAAGTCCAAGGAAATTTCTAAGTCTTCTGTTGTTGATGACATGGTTTCCGCCAACGCTTTGTTATTCAAGCCATCTGCCGTTGGTGCTCCAGCTGGTTCTAAGGAAGCTAAGGGTGAACATCCAGATCACATCGTTGTCATTAAGTACGTTCCAGCTGTCGGTGATTCTAAGAGAGCTATTGACGAATATTACTCCGAAATTTTCTGTGGTGGTAGATCTACTATCAACATTTTTAACGAATGTGAAGACTCCTTGTTGGCTACTCCATTGATCTTGGACTTGACCATCTTGACTGAATTATTGACTCGTGTCAAGTACAGAGACGCTTCTGCCGGTAAGGACTTCAAACCTTTGTATCCAATTTTATCCTTGTTGTCTTACATGTTGAAGGCCCCATTGGTCAAGCCAGGTACCGATGTCGTCAACTCCTTGAATAGACAAAGAAATGCTTTGGAAACCTTTTTGAAGGCCTGTATCGGTTTGGAAGGTTCTTCCGACTTATTGTTGGAGACTAGAATCTGGTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Drosophila melanogaster* | O97477 | ATGAAGCCAACTAATAACTCTACTTTGGAAGTTATCTCCCCAAAGGTCCAAGTTGACGATGAATTCATTACCACTGACTACGATTACCAAACTTCCCACGTCAAGCGTACTGCTGACGGTCAATTGCAAGTTCACCCTCAAACTACCTCTTTAAAAATCAGAACCGGTCGTCATGTTCCAAAATTAGGTGTTATGTTAGTTGGTTGGGGTGGTAACAACGGTTCTACCTTGACTGCTGCCTTGGAAGCTAACAGAAGACAATTGAAATGGAGAAAGAGAACCGGTGTTCAAGAAGCTAATTGGTACGGTTCCATCACTCAAGCCTCTACCGTTTTCATCGGTTCCGATGAAGACGGTGGTGATGTCTACGTTCCAATGAAAGAATTGTTGCCTATGGTTGAACCTGATAACATTATCGTCGATGGTTGGGACATCTCCGGTTTGCATTTAGGTGACGCTATGAGAAGAGCCGAAGTTTTAGATGTTGCTTTGCAAGATCAAATCTACGATCAATTGGCTCAATTGAGACCAAGACCATCTATTTATGACCCAGACTTTATTGCTGCTAACCAATCTGACAGAGCTGACAACGTTATTAGAGGTACTAGATTGGAACAATACGAACAAATCAGAAAGGACATTAGAGACTTCCGTGAGAGATCTGGTGTTGATTCTGTCATCGTCTTGTGGACCGCTAACACCGAAAGATTCGCTGACGTCCAACCAGGTTTGAATACTACTTCCCAAGAATTAATTGCTTCTTTGGAAGCCAACCACTCTGAAGTTTCCCCATCTACCATCTTTGCCATGGCTTCTATCGCTGAAGGTTGTACCTACATTAATGGTTCTCCTCAAAATACTTTTGTCCCAGGTTTGATTCAATTGGCTGAAGAAAAGAACGTCTTCATTGCTGGTGATGATTTCAAGTCTGGTCAAACCAAGATTAAGTCTGTTTTGGTCGATTTCTTGGTCGGTGCCGGTATCAAACCAGTCTCTATTGCTTCCTACAACCACTTGGGTAACAACGATGGTAAGAACTTGTCTGCTCCTCAACAATTCAGATCTAAAGAAATCTCTAAATCTAACGTTGTTGATGACATGGTTGCCTCTAATCGTTTGTTGTACGGTCCAGACGAACACCCAGATCATGTCGTTGTTATCAAGTACGTTCCATACGTTGGTGACTCCAAGAGAGCTATGGACGAATATACCTCTGAAATTATGATGGGTGGTCACAACACCTTGGTTATCCACAACACTTGTGAAGATTCTTTGTTAGCTACCCCATTGATTTTAGATTTGGTTATTTTAGGTGAATTATCCACCAGAATTCAATTGAGAAATGCCGAAAAGGAATCTGCTCCATGGGTTCCATTCAAGCCAGTCTTATCCTTGTTATCTTATTTGTGTAAAGCTCCTTTGGTCCCACAAGGTTCTCAAGTCGTTAACTCTTTATTCAGACAAAGAGCTGCTATTGAAAACATTTTGCGTGGTTGTATTGGTTTGCCACCTATCTCTCACATGACTTTGGAACAAAGATTCGATTTCTCTACCATTACTAACGAACCACCATTGAAAAGAGTTAAAATTTTGGGTCAACCTTGCTCCGTTGAATCTGTTACTAACGGTAAAAAGTTACACGCTAACGGTCACTCCAACGGTTCTGCTAAGTTGGCCACTAATGGTAACGGTCACTGATAA |
| *Gardnerella vaginalis* | E3D8F4 | ATGTCTATTAGAGTTGCTATTGCCGGTGTTGGTAATTGTGCTTCTTCCTTGGTTCAAGGTGTCGAGTACTATAAGAACGCCAACGATGGTGATAAGATCCCTGGTTTGATGCATGCCGTTTTCGGTCAATACAGAGTTAGAGATATTGAGTTTGTTGCTGCTTTCGACGTTGACGCTTTGAAGGTTGGTCACGACTTGTCTGAAGCCATTTATGCTTCTCAAAACAACACCATTCGTTTCGCCGACGTTCCTAACTTGGGTGTCAAGGTTCAAAGAGGTCCAACCTACGACGGTTTGGGTGACTACTACAAGCAAATGATCGAAGAGTCTAAGGAAGAACCAGTTAACGTTGCTGCTGTCTTGAGAGATTTACATGTCGACGTTTTGGTCTCTTACTTGCCAGTTGGTTCTGAACAAGCTGACAAGGCTTACGCTACCGCTGCTATGGAAGCCGGTTGTGCCTTCGTTAACTGTTTACCAGTCTTCATTGCTTCTGACCCAGTCTGGGCTCAAAAGTTTAGAGATGCTGGTGTCCCAATTATCGGTGATGATATCAAGTCTCAAGTTGGTGCTACTATTACTCACAGAGTTATGGCTCGTTTGTTTGAAGATAGAGGTGTTCGTTTAGATAGAACCTACCAATTAAACGTCGGTGGTAATATGGACTTTATGAACATGTTGCAAAGATCCAGATTAGAATCCAAAAAAATTTCTAAGACCCGTGCTGTTACTTCCATTGTTCCTCACGATATGGATGACCATAACGTTCACATTGGTCCATCTGACTACGTTGCTTGGTTGGATGATCGTAAGTTCGCTTTCGTTAGATTGGAAGGTACTACTTTTGGTGATGTCCCATTATCTTTGGAATACAAGTTGGAAGTTTGGGATTCTCCTAACTCTGCTGGTATCGTCATTGACGCCGTTAGAGCTGCTAAAATTGCTTTGGATAGAAAATTGTCTGGTCCAATCTTAGCTCCATCTTCTTACTTCATGAAATCTCCAGCTGTCCAACACGAAGATTCTGAAGCCAGAGAATTGGTCGAAAGATATATCGCTGGTGACGTTGAAGCCGACGAATCCCAATTGAATGCCGATGTCGAGGCTGCTAAGGAACACGGTAAGTCCGTTTGGAGAGCCTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| **Organism** | **UniProt ID** | **MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*)** |
|---|---|---|
| *Homo sapiens* | Q9NPH2 | ATGGAGGCTGCTGCTCAATTCTTCGTTGAATCTCCAGACGTCGTCTACGGTCCTGAGGCTATCGAAGCTCAATACGAATATAGAACTACTAGAGTTTCTAGAGAAGGTGGTGTTTTGAAGGTCCACCCAACTTCCACTAGATTTACTTTCAGAACTGCCAGACAAGTTCCACGTTTGGGTGTCATGTTAGTTGGTTGGGGTGGTAACAACGGTTCTACTTTGACTGCCGCTGTTTTGGCCAACAGATTAAGATTGTCTTGGCCAACTAGATCCGGTAGAAAGGAAGCTAATTACTATGGTTCTTTAACTCAAGCCGGTACCGTTTCTTTGGGTTTAGACGCTGAAGGTCAAGAAGTCTTCGTTCCATTCTCCGCCGTTTTACCAATGGTTGCTCCAAACGATTTGGTTTTTGATGGTTGGGATATTTCCTCTTTAAACTTGGCTGAAGCTATGAGAAGAGCTAAGGTTTTGGACTGGGGTTTGCAAGAACAATTGTGGCCACATATGGAAGCTTTGAGACCAAGACCATCTGTCTACATTCCAGAATTTATTGCTGCTAACCAATCCGCTAGAGCTGACAATTTGATTCCAGGTTCCAGAGCTCAACAATTGGAACAAATTAGAAGAGATATTAGAGACTTCAGATCCTCTGCCGGTTTGGACAAAGTCATCGTCTTATGGACCGCCAACACCGAAAGATTCTGTGAAGTTATTCCAGGTTTAAACGACACTGCTGAAAATTTGTTGCGTACCATCGAATTGGGTTTGGAAGTTTCTCCATCTACCTTATTCGCCGTTGCTTCCATCTTGGAAGGTTGTGCTTTCTTGAACGGTTCTCCTCAAAACACCTTGGTCCCAGGTGCTTTGGAGTTAGCTTGGCAACATAGAGTCTTCGTCGGTGGTGATGACTTCAAGTCTGGTCAAACTAAGGTCAAATCCGTCTTGGTCGATTTCTTGATCGGTTCCGGTTTGAAGACCATGTCCATTGTTTCTTACAATCATTTGGGTAACAACGACGGTGAAAACTTGTCCGCTCCATTGCAATTCAGATCTAAAGAAGTTTCCAAGTCTAACGTCGTCGATGACATGGTTCAATCCAATCCAGTTTTATACACTCCAGGTGAAGAACCAGACCACTGCGTTGTTATTAAATACGTCCCATATGTCGGTGACTCTAAACGTGCTTTAGACGAATATACCTCCGAATTAATGTTGGGTGGTACTAACACCTTGGTTTTACATAACACTTGTGAAGACTCTTTGTTGGCTGCTCCAATTATGTTGGATTTGGCTTTATTGACTGAATTATGCCAAAGAGTCTCTTTCTGCACCGATATGGATCCAGAACCACAAACCTTCCATCCAGTTTTGTCCTTATTGTCTTTCTTGTTTAAGGCTCCTTTGGTTCCACCAGGTTCTCCAGTTGTTAACGCTTTGTTCAGACAAAGATCTTGTATCGAAAACATTTTGAGAGCCTGTGTTGGTTTGCCACCACAAAACCACATGTTGTTGGAACACAAGATGGAAAGACCAGGTCCTTCCTTGAAGAGAGTCGGTCCAGTTGCTGCTACTTACCCAATGTTAAATAAGAAGGGTCCAGTTCCAGCTGCTACCAACGGTTGCACTGGTGATGCTAACGGTCATTTGCAAGAAGAACCTCCAATGCCAACCACTTGATAA |
| *Mesorhizobium australicum* | L0KRR8 | ATGGGTTCCAAGAAGGTTAGAGTCGGTATTGTTGGTGTTGGTAACTGTGCCTCCTCCTTGGTTCAAGGTTTGTCTTATTACAGACACGCCAAGTCTAACGAACCAATTCCTGGTTTAGTTCATGCCGACTTGGGTGGTTACCATGTCGATGACATTGAAATTGTCTGTGCTTTCGATGTTGCTAAGTCTAAGGTCGGTCGTGACGTTGCTGACGCTATTTACGCTCCACCAAATAATACCTTCAGATTCGCCGATGCTCCAACTACCGGTGTTTTGGTTGAAAGAGGTCCAACTTTAGATGGTATTGGTAAGTATTTGAGAGATGAAATCGAAGAAGCCCCAGAACCAGTCGCTAACGTTTCCGAAATTTTGCGTGATTCCGGTGCTGATGTCTTGGTCTCTTATTTGCCAGTCGGTTCCGAAGAAGCCACTCATTTTTACGCTGAATGTGCTTTGGAAGCCGGTTGTGCTTTCGTCAACTGCATTCCTGTCTTCATCGCCTCTAGACCAGAATGGAGAAGAAGATTCGAACAAAGAGGTTTGCCATTGGTTGGTGACGACATCAAGTCTCAAGTTGGTGCTACCATTGTTCACAGATTGTTGGCTAACTTGTTCAGAGAAAGAGGTGTCAGAATTGACCGTACCTACCAATTGAACTTCGGTGGTAACACCGATTTCTTAAATATGTTGGAACGTGAAAGATTGGAATCCAAGAAGATCTCCAAGACTCAATCTGTCACTTCTCAATTAGACGTCCCATTGGAACCAGGTAATATCCATGTCGGTCCATCTGACCACGTTCCATGGTTGACTGACAGAAAGTGGGCTTATATTAGAGTTGAGGGTACCACCTTCGGTGGTGTCCCATTAAATGCTGAATTAAAGTTAGAGGTCTGGGACTCTCCAAACTCTGCTGGTGTTGTTATTGACGCTGTTAGATGTGCTAAATTGGCCTTGGACAGAGGTATTGCTGGTGCTTTAACCGGTCCTTGTTCCTACTTCATGAAGTCCCCACCAGAACAATTCACCGATGCTGAAGCCCGTCAACGTACCTTGGCTTTCATTGCTGGTAAGGATGAACCATTGTTGGACGCTGCTGAGTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Methanolobus psychrophilus* | K4ME48 | ATGTATTACTTCGACAGAGGTAACGTCATGGACAAGATCAAGATCGCTATTGCCGGTGTCGGTAACTGTGCTTCCTCTTTAATCCAAGGTATCGAATACTACAGAGACAAACATGAAAACGACGCTATCGGTTTGATGCACTGGGACATTGGTGGTTACCGTCCATCTGACATTGAAGTCGTCGCCGCTTTCGACATCGACAAGAGAAAGGTTGGTAAAGACATCTCTGAAGCCATCTTCGCCCCACCAAATTGTACTGCCATCTTCTGTTCTGACATTCCACAAAGAGGTGTCGTTGTTAAGATGGGTTGTATTTTGGACGGTTTCTCTGAACACATGATGGACTTTGACGAAAAAAGAACTTTCGTTCCATCCGATCAACCAGAGGCCTCTAAGGAAGGTGTTGTTCAAGCTTTGAAGGACTCTGGTGCTGAAATCTTGTTGAACTATTTACCAGTTGGTTCTGAACAAGCCACTCGTTTCTACATGGATTGTGCTTTGGACGCCGGTGTCGCCTGTGTCAACAACATGCCAGTTTTCATCGCTTCTGATCCAGAATGGGCTGCTAAGTTCGAGAAGCGTGGTATTCCTATCATTGGTGACGATATCAAAGCTCAATTAGGTGCTACCATTACCCATAGAATGTTGGCTGACTTGTTCAACAAGAGAGGTGTTAAGTTGGAAAGAACTTACCAATTGAACACTGGTGGTAATACTGACTTCTTGAATATGTTGAATAGATCTAGATTGGCTTCTAAGAAGACTTCCAAAACTGAAGCTGTTCAATCCGTTTTGGCTCAAAGATTGGACGACGACAACATTCATGTCGGTCCTTCCGACTACGTTCCATGGCAAAACGACAACAAGGTCTGTTTCTTGAGAATGGAAGGTAAGTTATTTGGTGATGTTCCAATGAACTTAGAGTTGCGTTTGTCTGTTGAAGACTCTCCAAACTCTGCTGGTGTCGTCATTGACGCTATTCGTTGTTGTAAGTTGGCCTTGGATAGAGGTATCGGTGGTGTCTTGTACTCCCCATCTGCCTACTTCATGAAACATCCACCAAAACAATTCACTGACGATGAAGCTCACAAGATGACCTCTGAATTCATCCACGGTGACAGAACTAACTGATAA |
| *Mucilaginibacter paludis* | H1Y1B6 | ATGAAGACTAACATTGAACCAGCTGAAGGTAAATTGGGTATCTTGATCCCTGGTTTGGGTGCTGTTGCTACTACTTTAATCGCTGGTGTCGAAGCTGTTAAGAAGGGTATTTCTAAGCCAATCGGTTCCTTGACCCAAATGTCCTCCATCCGTTTAGGTAAGAGAACCGATAATAGATACCCAAAGATCAAGGACTTCGTTCCATTGGCTGACTTAAACGACATTGTCTTCGGTGGTTGGGATGTCTACGCTGACAACGTTTACCAAGCTGCCTCCAACGCCAAGGTCTTGGACCAACACTTGTTGGACGCTGTTAAGGAACCTTTGGAAGCTATCGTCCCAATGAAGGCCGCTTTCGACCATAATTACGTTAAGAATTTGACCGGTACCCATATCAAGGAATTTACTACCAGATACGACTTAGCCCAACAAGTCATCGCCGACATTGAAAACTTTAAGGAAAAGCACAACTTAAACAGAGTCGTTTTGGTTTGGTGTGGTTCTACCGAAATTTACTTCGAAGAATCTGAAATTCACCAAAACTTGGCTAATTTCGAACAAGCTTTACAAAACAACGATGAACGTATCGCTCCATCTATGATTTACGCTTACGCTGCTTTGAAGTTGGGTATTCCATTCGCCAACGGTGCTCCAAATTTGACTGTTGACATTCCAGCTTTAGTCGAATTGTCCAAGTTGACCAACACTCCAATTGCCGGTAAGGACTTCAAGACCGGTCAAACTTTGATGAAGACTATTTTGGCTCCAGGTTTGACTGCTAGAGCCTTGGGTGTTAAGGGTTGGTTCTCTACCAACATTTTGGGTAACCGTGACGGTTGGGTTTTGGACGATCCAGACAATTTTAAAACTAAGGAGGTTTCTAAGTTGTCTGTTTTGGAAGAAATCTTCCAACCAGAAATTAACCCAGAATTATACGGTGACATGTACCACAAGGTTAGAATCAACTACTACCCACCACGTGGTGATAACAAGGAATCCTGGGACAACATTGACATCTTCGGTTGGTTGGGTTATGAAATGCAAATCAAGATCAACTTCTTGTGCAGAGATTCCATCTTGGCTGCCCCAATCGTTTTGGATTTGGCTTTGTTCATGGACTTGGCTAAGAGAGCTGATATGTCCGGTATCCAAGAATGGTTGTCCTTCTACTTAAAGTCCCCACAAACCGCTCCAGGTTTGAAGCCAGAACACGATATCTTTAAGCAATTGATTAAGTTGCAAAATACTTTGCGTCATATGATGGGTGAAGATTTAATTACCCACTTAGGTTTAGACTACTACCAAGAATTGGTTGAATCCATGTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Nocardia nova* | W5TTL7 | ATGTCTGATGTTAACCCAGCTGCCGAAATCAGAGTCGCTATTGTCGGTGTCGGTAACTGTGCCTCTTCCTTGGTTCAAGGTGTTCAATATTACAAGGACGCTGATGAAAACGCTACCGTCCCTGGTTTAATGCATGTTAAATTCGGTCAATACCATGTCAGAGACGTCAAGTTCGTTGCTGCTTTCGACGTTGATGCTAAGAAGGTTGGTTTCGACTTGTCTGACGCTATCTTTGCTTCCGAAAACAACACCATTAAGATCTCCGACGTCCCACCAACTGGTGTCACCGTTCAAAGAGGTCCAACCTTGGACGGTATCGGTAAGTACTACGCTCAAACCATCGAATTATCCGAAGCTGACCCAGTCGATGTCGTCCAAGCCTTGAAGGATGCCCAAGTTGACGTTTTGGTTTCTTACTTACCAGTCGGTTCTGAAGACGCTGACAAGTTCTACGCTCAATGCGCCATTGACGCCAATGTTGCTTTTGTCAACGCTTTGCCAGTTTTCATTGCTTCTGATCCAGCTTGGGCCCAAAAATTTGTTGACGCTGGTGTTCCTATTGTCGGTGATGACATCAAGTCTCAAGTTGGTGCTACTATCACTCACAGAGTCATGGCCAAGTTGTTCGAAGATCGTGGTGTTCAATTGGATAGAACCATGCAATTGAACGTTGGTGGTAACATGGATTTCAAAAACATGTTGGAAAGAGAACGTTTAGAATCTAAGAAGATTTCCAAGACCCAAGCTGTCACTTCTAACTTGAAGAAAGAATTGGGTGCCAACGATGTTCACATTGGTCCATCTGATCACGTTGGTTGGTTGGACGACCGTAAATGGGCTTACGTCAGATTGGAGGGTCGTGCTTTTGGTGACGTTCCATTGAACTTGGAATACAAGTTAGAAGTTTGGGACTCTCCAAATTCTGCTGGTATTATTATTGACGCCGTCAGAGCTGCTAAAATCGCCAAGGACAGAGGTATCGGTGGTCCAGTTATCCCTGCTTCCGCTTATTTGATGAAATCTCCACCAAAACAATTGGCTGACGACGTTGCTAGAACCCAATTGGAAGCTTTCATTATTGGTGCTGAATGATAA |
| *Phytophthora ramorum* | H3G8E9 | ATGGCTTCTTCTGATTTCTTTCAAGAACCTTTCACTGTTAACTCTAAGAACGTCGTTTACTCTGCTGACGAAATCACTTCTCAATACACCTATACTACTACTAGAGTCGAAGGTACCGTTGCTACTCCAGTTGAAGAAAAGTATACTTTTAAGACCCAAAGAAAGGTCCCAAAGTTGGGTGTTATGATTGTTGGTTTGGGTGGTAATAACGGTTCCACTTTGTTGGCCTCCATTATTGCTAATAAGCAACACATTACCTGGACTACTAAGGAAGGTGTTCAAGAGCCAAATTATTTCGGTTCTGTTACTCAAGCTTCTACTGTTAGATTGGGTACTAACGCTAACGGTGAAGGTGTTTACATCCCATTCCACAACTTGTTGCCAATGGTTGCTCCTAACGATTTGGTTATCGGTGGTTGGGATATCTCCTCTTTGAACTTGGCCGAGGCCATGAAGCGTGCCCAAGTCTTGGACCACGACTTACAAAGACAATTGGTCCCACATTTAGAACAAATTAAGCCATTGCCTTCCATTTACTACCCAGATTTCATTGCTGCTAACCAAGCTGACAGAGCTGATAACTTGTTAAAAGGTTCTAAACAAGAACACTTAGATGCCGTCAGACAACAAATCAGAGATTTCAAGCAATCCAACGGTTTGGACAAGGTTATCGTCTTGTGGTCTGCCAACACTGAACGTTTCTCCGACATCGTTGAAGGTGTTAATGACACCTCCGCTAACTTGTTGGAATCTATTAAGGCTGGTGAACCAGAAGTTTCTCCATCTACTGTTTTTGCTGTTGCTTCCATCTTGGAAGGTTGTTCTTACATCAACGGTTCTCCTCAAAACACCTTCGTTCCTGGTGTCTTGGATTTGGCTGAAGAGAAAAAGATTTTCGTCGGTGGTGACGATTTCAAGTCTGGTCAAACCAAAATGAAGTCTGTCTTAGTTGACTTTTTGGTTTCTGCTGGTATTAAGCCAACCTCTATCGTCTCCTACAACCATTTGGGTAATAACGATGGTAAGAACTTATCCGCTCCACAACAATTTAGATCTAAGGAAATCTCTAAGTCTAACGTCGTCGATGATATGGTCGCTTCCAATAGATTATTGTACAAGGAGAACGAACATCCAGATCACGTCGTTGTCATCAAGTACGTTCCATTCGTCGGTGACTCTAAAAGAGCTTTGGACGAGTACACTTCCAAAATCTTTATGAACGGTCAAAATACCATTTCTATGCATAACACCTGTGAGGATTCCTTGTTAGCTTCCCCTTTAATCTTGGATTTGGTTTTGGTTTGTGAATTGGCTGAAAGAATTACTTTGAAAAAGGAAGGTGCCAAAGATTTCGAACATTTGCACTCTATTTTGTCCATCTTGTCCTACATGTTGAAAGCTCCATTAGTCCCTCGTGGTACTCCTGTTGTTAACGCTTTGTTCGCCCAAAGAGAGTGTATGATCAATATCTTCAGAGCCTGTGTTGGTTTGACCCCAGAATCTCATATGTTGTTGGAAAATAAGTTGGCCTCCGAAATTGATGCTAGACAATGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Prevotella buccae* | D3HVK9 | ATGGAAAGAACCAACGTTAAGCCAGCTGAAGGTAAGTTGGGTATCATGGTTGTTGGTTGTGGTGCTGTTGCCACTACCTTCATGACTGGTGTTTTAATGGTCCGTAAGGGTTTGGCTAAGCCAGTCGGTTCTATGACCCAATACGACAAGATTAGAGTTGGTAAGGGTGATAACAAAAAGTACTTGTCCTACGCTGACATTGTTCCATTGGCTAAATTGGATGATATCGTTTTCGGTACTTGGGACGTTTACCCTCAAAATGCTTACCAAGCCGCTATGTACGCTGAAGTCTTGCAAGAAAAAGACATCAACCCAGTTCGTGATGAATTGGAAAAGATCGTCCCATTGAAGGCTGCCTTCGACAAGAACTATGCCAAGAGATTGGATGGTGATAACGTTAAGGACTGTAAAACCAGATGGGAGATGGTTGAAGAGTTACGTAGAGACATGAGAAGATTCAAGGAAGAAAACGGTTGCGCTAGAATCGTTGTCATCTGGGCTGCTTCTACTGAAATTTACGTTCCAGTTGATGAAAGAGTTCACGGTACTTTGGCTGCTTTGGAAGCTGCTATGAAAGCTGACGATAGAGAGCATGTTGCTCCATCCATGTGTTACGCTTACGCTGCCTTGAAAGAAGGTGCCCCTTTCATTATGGGTGCCCCAAATACCACCGTCGATATTCCTGCTATGTGGGAATTAGCTGAACAAACTAGAATGCCAATTTCTGGTAAGGACTTTAAGACTGGTCAAACTTTGGTTAAGTCTGGTTTCGCTCCAATCATCGGTACTAGATGTTTAGGTTTGAATGGTTGGTTTTCTACTAATATCTTGGGTAACAGAGACGGTTTGGTCTTGGATGAACCAGCTAACTTTCATACTAAGGAAGTCTCCAAGTTGTCTACTTTGGAGACTATTTTGAAGAAGGAAGACCAACCAGATTTGTACGGTGATATCTACCATAAAGTTAGAATCAACTACTATCCACCAAGAAACGACAACAAAGAAGGTTGGGATAACATCGACATCTTTGGTTGGATGGGTTACCCAATGCAAATCAAAATTAACTTTTTATGTAGAGACTCCATTTTAGCCGCTCCATTGTTGTTGGATTTGACCTTATTGTCTGATTTGGCTGCTAGAGCTGGTAGATATGGTATTCAAAGATTTTTGTCTTTCTTCTTGAAGTCTCCTATGCACGATTACACTCAAGGTGAAGAACCAGTTAACAACTTGTACCAACAATACACTATGTTGAAGAACGCTATCCGTGAAATGGGTGGTTACGAACCAGATGAAGAAATCGATTGATAA |
| *Sesamum indicum* | Q9FYV1 | ATGTTCATCGAATCCTTCAAGGTTGAATCTCCAAACGTTAAATACACCGAAGGTGAAATTCATTCTGTCTATAACTACGAAACCACCGAATTGGTTCACGAGTCCCGTAATGGTACTTATCAATGGATCGTCAAACCAAAGACTGTCAAGTACGAATTCAAGACTGATACTCACGTTCCAAAGTTAGGTGTCATGTTGGTTGGTTGGGGTGGTAACAACGGTTCCACTTTAACTGGTGGTGTTATTGCCAACAGAGAGGGTATTTCTTGGGCCACTAAGGATAAAGTTCAACAAGCTAATTACTTCGGTTCTTTGACCCAAGCTTCTTCTATTAGAGTTGGTTCTTTTAACGGTGAAGAAATCTACGCCCCATTTAAGTCTTTGTTGCCAATGGTTAACCCAGATGACGTTGTTTTCGGTGGTTGGGATATTTCTAACATGAACTTAGCTGACGCCATGGGTAGAGCCAAGGTTTTGGATATCGATTTGCAAAAGCAATTGAGACCATATATGGAACATATGGTTCCATTACCAGGTATCTACGATCCTGATTTCATCGCTGCCAACCAAGGTTCTAGAGCTAACAACGTTATCAAGGGTACCAAGAAGGAACAAGTTCAACAAATCATCAAGGACATGAGAGATTTCAAAGAACAAAACAAGGTCGACAAGGTCGTCGTCTTATGGACTGCTAATACTGAAAGATACTCCAACGTTGTTGTTGGTTTGAATGATACCGCTGAATCTTTGATGGCCTCTGTTGAACGTAACGAAGCTGAAATCTCTCCTTCTACTTTGTACGCCATCGCTTGTGTTTTCGAAAATGTTCCTTTCATTAACGGTTCTCCTCAAAACACTTTTGTTCCAGGTTTGATCGATTTAGCTATCCAACGTAACTCCTTGATCGGTGGTGACGACTTCAAGTCTGGTCAAACTAAGATGAAGTCCGTTTTGGTCGACTTCTTGGTTGGTGCCGGTATTAAGCCAACTTCTATTGTTTCCTACAACCACTTGGGTAACAATGACGGTATGAACTTGTCCGCCCCACAAACTTTCAGATCCAAGGAGATCTCTAAGTCTAACGTTGTCGACGATATGGTTGCTTCTAATGGTATTTTGTACGAACCAGGTGAACATCCAGATCATATTGTTGTCATCAAATACGTCCCATACGTCGGTGATTCCAAGAGAGCTATGGACGAATACACCTCTGAAATCTTCATGGGTGGTAAGTCTACCATTGTCTTGCACAATACTTGTGAAGACTCCTTGTTGGCCGCTCCAATCATTTTGGACTTGGTTTTGTTAGCTGAATTATCTACCAGAATCCAATTAAAGGCCGAAGGTGAAGGTAAATTTCATTCTTTCCATCCAGTCGCTACTATCTTGTCTTACTTGACTAAGGCTCCATTGGTTCCTCCAGGTACCCCAGTCGTTAACGCTTTGTCTAAACAACGTGCTATGTTAGAAAACATCTTGAGAGCTTGTGTTGGTTTAGCTCCAGAAAACAACATGATTTTGGAATACAAGTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Sphaerobacter thermophilus* | D1C4I3 | ATGTCTTCCAGAAAGATCAGAGTCGCCATCATCGGTGTCGGTAACTGTGCTTCTTCCTTAGTTCAAGGTGTTGAATACTACAGACACGCCGACCCAAATGACTTCGTCCCAGGTTTAATGCATGTCGACTTGGGTGGTTACCACGTTGGTGACATTGAATTCTCTGCTGCCATTGATATTGACAAGAACAAGGTCGGTAAGGACTTGTCTGAAGCCATCTTCACCTCCCCAAACAACACCTACAAGTTCTCTGATGTCCCACATTTGGGTGTTCCAGTCCACAGAGGTATGACTCACGACGGTTTGGGTAAGTACTTATCCCAAATTATCGAAAAAGCCCCTGGTTCCACTGCCGATATCGTCGGTATCTTAAAAGAGACTGGTACTGACGTCGTCGTTAACTTCTTGCCTGTTGGTTCTGAAATGGCTACTAAGTGGTACGTTGAACAAGTTTTGGAAGCCGGTTGTGCTTTCGTTAACTGTATTCCAGTCTTTATCGCTAGAGAGGAATACTGGCAAAACAGATTCAGAGAACGTGGTTTGCCAATTATCGGTGATGATATTAAGTCCCAAGTTGGTGCTACCATTACCCATAGAGTCTTAACCAGATTGTTCGCCGACAGAGGTGTCAGAATTGACCGTACTTACCAATTGAATTTCGGTGGTAACACTGATTTTTTGAACATGTTGGAAAGAGAAAGATTGGAATCTAAGAAGATTTCCAAGACCAATGCTGTTACTTCTCAAATTGATTACCCAGTTGACCCAGAAAACGTCCACGTCGGTCCATCTGACTACGTCCCATGGTTGCAAGACCGTAAGTGGTGTCATATCAGAATGGAAGGTACCACTTTCGGTGATGTTCCATTGAACATCGAATTGAAATTAGAAGTCTGGGACTCCCCAAACTCTGCCGGTGTCGTCATCGATGCCATCAGATGTGCCAAATTGGCCTTGGACACTGGTATCTCTGGTGCTTTGTTGGGTCCATCTGCTTACTTCATGAAGTCTCCACCAGTCCAATACCATGACGACCAAGCCAGAGAAATGGTCGAATCTTTCATTAGAGAAACTGTCGCTCACAGAGAAGCTGCTGAAGCTGCCGCTACTCCTGCTGAACAAGGTTGATAA |
| *Streptomyces cattleya* | F8JTE4 | ATGGGTTCTGTTAGAGTCGCTATTGTCGGTGTCGGTAACTGTGCCGCTTCTTTAGTTCAAGGTGTCGAATACTACAAGGATGCTGACCCAGATTCTAGAGTTCCAGGTTTGATGCACGTCCAATTTGGTGACTACCACGTTAGAGATGTCGAGTTTGTCGCCGCTTTCGATGTTGACGCTAAGAAGGTCGGTTTAGACTTGGCTGATGCCATCGGTGCTTCTGAAAACAACACTATTAAGATCTGTGACGTCCCACCATCTGGTGTTACTGTCCAAAGAGGTCACACTTTGGACGGTTTGGGTAGATACTATAGAGAAACTATTGAAGAGTCCGCCGAAGAACCTGTTGATGTCGTTCAAATTTTGAAAGATAGACAAGTTGATGTTTTGGTCTGTTATTTGCCAGTTGGTTCTGAAGAGGCTGCTAAGTTTTATGCTCAATGCGCCATCGACGCCAAGGTCGCCTTCGTTAACGCCTTGCCAGTCTTCATTGCTGGTACTAAGGAATGGGCTGATAAATTCACCGAAGCCGGTGTTCCAATCGTTGGTGACGATATCAAGTCTCAAGTTGGTGCTACCATTACCCACCGTGTCATGGCTAAGTTGTTCGAAGATCGTGGTGTTGTCTTGGATCGTACTATGCAATTGAATGTCGGTGGTAACATGGATTTCAAGAACATGTTGGAAAGAGATAGATTAGAATCCAAAAAGATCTCCAAGACTCAAGCTGTCACTTCTCAAATCCCAGATAGAGATTTAGGTGCCAAGAATGTCCACATCGGTCCATCTGATTACGTCGCTTGGTTGGATGATCGTAAATGGGCTTACGTTAGATTAGAAGGTAGAGCCTTCGGTGACGTCCCATTGAACTTGGAATACAAGTTGGAAGTCTGGGACTCTCCAAACTCTGCTGGTGTTATCATCGATGCCTTGAGAGCTGCCAAGATTGCCAAGGACCGTGGTATCGGTGGTCCAGTTTTATCTGCTTCTTCCTATTTCATGAAATCCCCACCTGTCCAATACTTTGACGATGAAGCCAGAGAAAATGTTGAAAAGTTCATCAGAGGTGAAGTTGAGAGATGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Thermococcus eurythermalis* | A0A097QQW8 | ATGGTTAAGGTTGTCATTTTGGGTCAAGGTTACGTTGCTTCCATCTTCGCTTCTGGTTTGGAAAAGATTAAGGCTGGTAAGATG<br>GAACCATATGGTGTCCCATTGGCTGATGAATTACCAATTAAGATCAAGGACATCGAAATCGTTGGTTCCTACGATGTTGACAA<br>AGCCAAGGTTGGTAAGGATTTGTATGAAGTCGTTAAGGCCTACGATCCAGAGGCCCCAGAATCTTTGAAGGGTATTACCATCA<br>GAAAAGGTGTCCATTTGAGATCTTTGAGAAACTTACCATTGGAAGCCACTGGTTTGGAAGATGAAATGACTTTGAAGGAAGC<br>CGTCGAACATTTGGTTTCTGAGTGGAAGGAATTGGGTGCTGAAGGTTTCATTAACGTCTGTACTACTGAAGCTTTCGTCCCATT<br>TGGTAACAAGGAAGAATTGGAAAAGGCCATTGCTGAGGACAACAGAGACAGATTGACTGCTACTCAAGTTTACGCTTATGCC<br>GTTGCTCAATACGCTAAAGAAGTCGGTGGTGCTGCCTTTGTTAACGCCATTCCAACCTTAATTGCCAACGATCCAGCTTTCGTT<br>GAATTAGCTAAAGAATCTAACATGGTTATCTTCGGTGATGATGGTGCCACCGGTGCTACCCCATTAACCGCCGATATCTTATCC<br>CACTTGGCTCAAAGAAACAGATATGTTTTGGATATTGCTCAATTCAACATCGGTGGTAACCAAGACTTCTTAGCCTTGACCGAC<br>AAAGAAAGAAACAAGTCTAAGGAATTCACCAAGTCCTCCATTGTTAAGGACTTGTTGGGTTACGACGCTCCACATTACATTAA<br>ACCAACTGGTTTCTTAGAACCTTTGGGTGATAAGAAATTCATCGCTATGCATATTGAATACGTCTCTTTCAACGGTGCTCACGA<br>CGAATTGGTTATTACTGGTAGAATTAACGATTCTCCAGCTTTGGCCGGTTTATTGGTCGACTTGGCCAGATTGGGTAAGATTGC<br>TTTGGAAAAGAAAGCTTTCGGTACTGTTTACGAAGTTAACGCTTTCTACATGAAGAACCCAGGTCCAAAGGAAATGCCAAACA<br>TTCCACGTATTATTGCTCACGAAAAGATGAGAACTTGGGCTGGTTTAAAACCTAGATGGTTGTGATAA |
| *Vigna radiata* | A8WEL5 | ATGTTCATCGAAAACTTTAAGGTTGAATGTCCAAACGTTAGATACACCGAGACTGAAATTCAATCTGTCTACAACTACGAAACC<br>ACTGAATTGGTTCACGAAAACCGTAACGGTACTTACCAATGGATTGTTAAGCCAAAGTCCGTTAAGTATGAATTCAAGACTGA<br>CACCCATGTCCCAAAGTTGGGTGTTATGTTGGTTGGTTGGGGTGGTAACAACGGTTCCACTTTGACCGGTGGTGTTATCGCCA<br>ACAGAGAAGGTATCTCTTGGGCTACTAAGGACAAGATCCAACAAGCCAACTACTTCGGTTCCTTGACTCAAGCTTCTGCTATCA<br>GAGTTGGTTCTTTCCAAGGTGAAGAAATCTACGCCCCATTCAAATCTTTATTACCTATGGTTAACCCAGATGACATTGTCTTCGG<br>TGGTTGGGACATCTCCAACATGAACTTGGCTGATGCTATGGGTAGAGCTAAGGTTTTCGATATCGACTTGCAAAAGCAATTGA<br>GACCATACATGGAATCCATGGTCCCATTACCAGGTATCTACGACCCAGATTTCATTGCTGCCAACCAAGAAGAGAGAGCTAAC<br>AACGTTATCAAGGGTACTAAGAAGGAACAAGTCCAACAAATCATCAAGGACATTAAGGAATTCAAGGCTGCTACTAAAGTTG<br>ATAAAGTTGTTGTTTTATGGACTGCTAATACCGAAAGATACTCCAACTTGGTTGTCGGTTTGAACGATACTTCCGAAAACTTGT<br>TGGCCGCTTTGGATAGAAACGAAGCTGAAATCTCCCCTTCTACCTTGTACGCTATCGCTTGCGTTATGGAGAATGTCCCATTCA<br>TTAACGGTTCCCCTCAAAACACCTTTGTTCCAGGTTTGATTGATTTCGCCATTGAAAAGAACTCCTTGATTGGTGGTGACGATTT<br>TAAGTCTGGTCAAACTAAGATGAAGTCCGTCTTGGTTGACTTCTTGGTTGGTGCTGGTATCAAGCCAACTTCTATTGTTTCTTAC<br>AACCATTTAGGTAATAACGATGGTATGAATTTATCCGCTCCTCAAACTTTCAGATCTAAAGAAATCTCCAAGTCCAACGTTGTT<br>GACGATATGGTCAACTCTAACGCTATCTTGTTTGAACCAGGTGAACATCCAGACCATGTCGTCGTTATCAAATACGTCCCATAT<br>GTCGGTGACTCTAAGAGAGCCATGGACGAATACACCTCTGAAATCTTTATGGGTGGTAAGAACACTATCGTTTTACACAACAC<br>CTGTGAAGACTCTTTGTTAGCCGCTCCTATCATTTTGGATTTGGTCTTATTGGCTGAATTATCTACTAGAATCCAATTTAAGGCT<br>GAAAACGAAGGTAAGTTCCACTTATTCCATCCTGTTGCTACTATTTTATCCTACTTGACTAAAGCTCCATTGGTCCCACCAGGTA<br>CTCCTGTTGTTAACGCTTTGTCTAAACAAAGAGCTATGTTGGAAAACATCTTACGTGCCTGTGTTGGTTTAGCTCCAGAAAACA<br>ACATGATTTTGGAATACAAGTGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Zygosaccharomyces bailii* | S6EIK9 | ATGACTACTGATTCTTACTTCACCCCATCTATTAAGGTTGCTAACGAAAATGTCCAATACTCCGAAACTGAATTAACCACCAAGTACACTTACGTCAACTCTATCGTTACCGAAAACCCATCCACTCAAACCTTGTCTGTTAACCCAGTTGAACAAACCTACCAATTCAAGGTCGATTTGAAGTTGCCAAAGGTTGGTGTTATGTTGGTCGGTTTCGGTGGTAACAACGGTACTGCTTTCTTGGCTTCCATTTTAGCCAACAGAGAAAAATTGAAGTTCAACACTAAGGAAGGTTTGTTGCAAGCTAACTACTACGGTTCCGTCACTCAATCTTCCACCTTAAAATTGGGTATCAGAGAAGACGGTTCTGATTACTACGTTCCATTTAACTCCTTATTACCATTTGTTTCTCCAAACGACTTCGAAGTTACTGGTTGGGATATCAACGGTTCCGATATGGGTAAGGCCATGACCAGAGCTCAAGTTTTGGAATATGACTTGCAAGATAAGTTGAGATCTGAAATGTCCAAGTATAAGCCATTGCCATCCATTTACTACCCAGATTTCATTGCTGCCAACCAAGACGACAGAGCCGATAATTGTATCAACAGACCAGACAACTCTGCCCCAGCTTCTACTAAGAACAAGTGGTCTCATTTGGAAAAGATTCGTTCCGATATCAGAAACTTTAAGGAAAAGAAGAACTTAGATAAGGTCTTGGTCTTGTGGACCGCTAATACTGAGAGATACGCTGATATCGTCCCAAACGTTAACGATACCGCTGATAACTTGTTGAATGCTATTAAGGAAGACAACGAAGAAATTGCTCCTTCCACTATCTTCGCTGTTGCTTCCATCTTGGAAAACGCCGTTTACATTAATGGTTCTCCTCAAAACACTTTCGTTCCAGGTGTTATTGAATTGGCTGAAAGAGAAGATACTTTTATCGCTGGTGATGACTTGAAGTCCGGTCAAACTAAAGTCAAGTCCGTTTTGGCTCAATTTTTGGTCGATGCCGGTATCAGACCAGTCTCTATCGCTTCTTACAACCACTTGGGTAATAATGATGGTTACAACTTGTCTTCCGAGCGTCAATTCAGATCCAAAGAAATCTCTAAAAAGTCCGTTGTTGATGATGTCATTGCTTCTAACCAAATTTTGTACAACGATAAATTGGGTAAGACCATTGACCATTGTATCGTTATCAAATACATGAACGCTGTCGGTGACTCTAAGGTCGCCATGGACGAATACTACTCTGAATTGATGTTAGGTGGTCACAACAGAATTTCCATCCACAACGTCTGTGAAGATTCTTTGTTGGCTACCCCATTGATCATTGACTTATTAATCATGGCTGAATTTTGTACTCGTGTTTCCTACAAGAAGGCTGGTGGTAACGACAATTACGAAAAATTCTACAACATTTTGTCTTTTTTATCCTACTGGTTGAAGGCCCCATTGACTAGAAAAGGTTACCAAACTATTAACGGTTTGAACAAGCAAAGAGCTGGTTTGGAAAACTTCATGAGATTGTTAATCGGTTTGCCACCACAAGACGAATTGCGTTTTGAAGAAAGATTAAAGTGATAA |
| *Nitrosopumilus maritimus* | A9A3B6 | ATGACTGGTAGAATTAAGGTTGGTTTGGTTGGTATCGGTAACTGTTTCTCCGGTTTGATCCAAGGTATTGAATACTATCGTAAGAACCCATCTCAAGAAGTTATTGGTATCATTCATGACAAGTTAGCCGGTTACGGTATTCACGATATTGACTTCGTTTGTGGTTTCGACGTCGGTGAAAACAAGGTCGGTAAATTGATTAACGAAGCCATTTATGAATACCCAAACATGGTTGATTGGATCCCAAAAGATGAAATGCCAAAGACCGATGGTAAGGTTTTCGAATCCCCAGTTTTAGATGGTGTTGGTTTGTGGGTTGAAAACAGAGTCAAGCCAATTAAGTCTGCCAAAACTGACGATGAGATCGCTGAAGAAGCTAAAAAAATTATTAAAGAAACTGGTGCCGAAATCATTGTTTCCTATTTGCCAGTCGGTTCTGACAAGGTTACCCAATTCTGGGCTCAAGTCTGTTTAGACACCAATACCGCTTTTGTCAATTGTATCCCTTCTTTTATTGCTTCTGATCCAGAGTGGGCTAAGAAGTTTGAAGAAAAGAACATTCCATGTATTGGTGATGATATCAAAGGTCAAGTTGGTGCTACCATTGTCCACAGAACTTTGGCTAAGTTATGTAATGACAGAGGTACTAAAATTGAAAAGACTTACCAAATCAACGTTGGTGGTAACACCGACTTCTTGAACATGAAGGAACAAGAAAGATTGGTTTCTAAAAAGATCTCCAAGACTGAATCTGTCCAATCTCAATTGGACGAAAGATTAGATGATGACCAAATCTACGTTGGTCCATCCGATTTTATCCCTTTCTTGGGTAACACTAAATTAATGTTTATGAGAATCGAAGGTAGACAATGGGCTAACATTCCTTACAACATGGAAGTTCGTTTAGACGTTGATGACAAGGCTAACTCCGCCGGTATTGTTATCGACGCCATCAGATTGGCTAAGATCGCTTTGGATAGAGGTGTTGGTGGTCCAATCAAGCCAGCTTCCGCTTACTTGATGAAGCATCCAATTGAACAAACTTCTGACGTTGCTGCCAAAACTGCTTGTGAAAAGTTCGTTGCTGGTGAATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Methanothermobacter thermautotrophicus* | T2GII1 | ATGGATAAGATTAAGATTGCTATTGTTGGTGTTGGTAACTGTGCCTCTTCCTTAATTCAAGGTATTTACTACTACAGAAACAAGGGTGCTGGTGACTCCATTGGTTTGATGCATTGGGATATTGGTGGTTACGAACCAGGTGACATCGAAGTTGTTGCCGCCTTCGACATCGATAGAAGAAAGGTTGGTAGAGATGTTTCTGAAGCTATTTTCGCTCCACCAAATTGTACCGCCGTTTTCTGTGACGACGTTCCAGAAATGGGTGTCGAAGTTTCCATGGGTCACGTCTTGGATGGTGTTGCTCCACACATGAAGGATTACCCAGAAAAGCAAACCTTCGTTGTTGCTGACGAAGAACCAGTTGACGTTGTTGAAGTTTTGAGAGAGTCTGGTGCCGAGATTTTGTTGAACTACTTGCCAGTTGGTTCTGAAGAAGCCGCTCGTTTTTATGCCAGATGTGCTTTGGAAGCTGGTGTTGCTTACATCAACAACATGCCAGTCTTCATTGCTTCCGATCCAGAATGGGCCGCTAGATTTCAAGAAAAGGGTATTCCAATTGTCGGTGATGACATTAAGGCTCAATTGGGTGCTACTATTACCCACAGAACCTTGACCAACTTATTCAAGAGAAGAGGTGTTAAGTTGGATAGAACTTACCAAATTAACACTGGTGGTAACACCGACTTTTTAAACATGTTGAACAGAGACAGATTGGACTCCAAGAAAGAATCTAAGACTGAAGCCGTCCAATCTATTTTAGGTGAAGACAGATTGGATGACGAAAACATTCACATCGGTCCATCTGACTATATTCCATGGCAAAAGGACAACAAAATTTGTTTTTTAAGAATGGAAGGTCGTTTGTTCGGTGATGTCCCAATGAACTTGGAATTGAGATTGTCCGTCGAAGACTCCCCTAACTCCGCTGGTTGTGTTATCGACGCTATTAGATGTTGTAAGTTAGCTATTGACAGAGGTATTGGTGGTCCATTGACTTCCATTTCTTCCTACACCATGAAGCACCCACCTGTCCAATATACCGACGACGTTGCTGCTAGAATGGTCGATGAATTTATTGCTGGTGAAAGAGAAAGATAA |
| *Thermocrinis albus* | D3SMX0 | ATGGCTGACAGAAAAATTAGAGTTGCTATCGTCGGTGTTGGTAACTGTGCTTCCGCTTTGGTCCAAGGTATTTACTACTATCAAAAGAGACAAAATTTGGACACTTCTGGTTTAATGTTTGAAGATGTTGGTGGTTACAAGCCATGGGATATCGAAATTGTTGCTGCCTGGGACATTGACGCTCGTAAGGTTGGTAAAGATGTCTCTGAAGCCATCTTTTCTCCACCAAACTGTACTACTGTCTTCGAACCAGAAGTTCCACATATGGGTGTCAAGGTCAGAATGGGTAAGGTTTTGGATGGTTATGCTCCACATATGGCTAATTACCCACCAGAGAGATCTTTCGTCTTGGCCCAAGAAAAGGAAGATGAATTAGAAGATGTTGTTTCTGTCTTGAAAGAAACTAGAGCTGACGTCTTGGTTAATTACGTTCCAGTCGGTTCTGAGCAAGCTGCTAGATTCTACGCTGAGGCCTGTTTGAGAGCTGGTGTTTCTTTCATCAATGGTATGCCAACCTTCATCGTTTCTGATCCAGAATGGGCTAAGAGATTTGAAGCTGAAGGTATCCCAGCTGTCGGTGACGATATTAAGTCCCAAGTCGGTGCTACTATCTTACACAGAACTTTGGTTCAATTATTCGTCGAAAGAGGTGTCAAGATCGATAGAACTTATCAATTGAATTTCGGTGGTAACACTGACTTCTTGAACATGTTAGAACGTTCTAGATTGCAAACCAAGAAGACCTCCAAAACTGAAGCTGTCTCCTCCTTGATCCCATATACCTTGGATTGGGAAAATATTCATATCGGTCCATCTGACTGGGTTCCATGGTTGAAAGATAGAAAGATTGCTTACATTAGATTGGAGGGTAGATTGTTCGGTGATGTCCCAATGTACGTCGAAGTTAAATTGGACGTCGAAGATTCCCCAAACTCTGCTGGTTCCATGATCGACGCTATTAGATGTTGTAAATTGGCCAGAGACAGAGGTATTGGTGGTCCATTATACTCCATTTCCGCTTACACTATGAAACACCCACCAGTCCAATACCCAGATTGGCAAGCTAGAAAGATGGTTGAAGAATTCATTAGAGGTGAAAGAGAAAGATAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| **Organism** | **UniProt ID** | **MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*)** |
|---|---|---|
| *Bacillus mycoides* | A0A076W5U7 | ATGACTTACCAAACTGGTGTTTTATTCGTTGGTATGTTGGGTGCTGTTGCTACTACCACCATTTCCGGTTTGTTCGCCGTTAACCAAAACTTGGCTCCTTTGAGAGGTGTCATCTCTTCTGAAAAGGAATTCGAAGTTTTGCAATTGACCCCATTAGATCAAATCGCTTTCGGTGGTTGGGACATTCAAAAAGATTCCTTGATCGAAGTTGTTAAGAGACATGGTATCATCCAAGAACCAATCTTGCAAAAGATTGAAATGAATTTGAATGATGTTCCAGTTTGGCAAGCCCCATTGGCCAATGTTAACGACTTCGTTAAAGGTGTCTACTCCTTGAAAGGTGAACCTGAAACTTTGGAATCTGCTGTTGACAAGATTCAAGAAGACATTGAAGCTTTCAGAAAGAAGTACGATTTAGAAAGAATCGTCGTCATCAACACCGCTTCCACCGAAGAAAAGACTAAGACTCACTCTTTATACCAATCCTTGAAGGCTTTCGAAACTGGTTTGAGAGAAAACTCCCCTGATATTAGACCTGGTATGTTGTATGCCTACGCTGCTATGAAGTCCAAATGTGCTTACGTCAATTTTACCCCATCCGTTACCGCCGAAATCCCAGCCTTACAAAAGTTGGCTGAAACCCAAGGTGTTCCAACCGCCGGTAAGGACGGTCGTACTGGTCAAACCTTGTACAAGCATGTTTTGGGTAAGATGTTCAAGCAAAGAGGTTTGAACATTGTCGGTTGGTACTCTACTAACATCTTGGGTAACCAAGATGGTGCCATTTTGGATCATCCAAGACACTCCTCTACTAAGATTGATTCCAAGTCCATCGGTTTGGAAAGAATTTTAGGTTACTCTCACTTCGACCATAAGGTCAGAATTGACTACTTCCCAGTCCGTGGTGACAGAAAGGAAGCCTGGGACACTGTCGATTTCGAAGGTTGGTTGGGTGAAAGAATGACTATGAAAATCAATTGGTTGGGTATTGATTCTATCTTGGCTGCTCCATTGATTATCGACTTGTCTAGATTCATGGACCACGCCTTGCAAAAAGGTAAAGCTGGTATCATGGAACACTTGTCTTTGTTCTTTAAGTCCCCAATCGGTACTGACGAATACGCTTTGGATCAACAATACCAAACCTTGTTGGAATATGTCAAGCACTTTGAATACAACGCCTAA |
| *Bradyrhizobium sp. WSM1253* | I2QG71 | ATGCACTCCAGATTGCAAGATAGAAGAAGAGTCAGAGTCGGTATTGTCGGTGTTGGTAACTGCGCTTCCTCCTTCGTCCAAGGTTTATCTTACTACCGTGATGCTAAGTCTAATGAACCAGTCCCAGGTTTGATGAACGCCGATTTGGGTGGTTACCACATCTCCGACATTCAAGTCGCTTCCGCCTTTGATGTTCATGCCGGTAAGGTTGGTCGTGATGTTGCCGAAGCCATTTTCGCTGCTCCAAATAATACCCACAGATTCTCCGACGTCGCCCCAACTGGTGTCATGGTTCAACGTGGTCCAGTCATGGACGGTGTCGGTCAATACTTGAAGGACGACGTTCCAATTGCCGATGTCCCAGAAGCCGATGTCTCTGAAGTCTTAGCTACCTCTAGAACTGATGTCTTGGTCTCTTACTTGCCAGTCGGTTCCCAACGTGCTTCTGAATTCTACGCTGCTAGAGCTATCGAAGCTGGTTGTGGTTACGTTAACTGTATTCCAGTTTTCATCGCTTCTAACCCAGACTGGAGAAGAAGATTCGAAGATGCTGGTTTGCCAATCGTTGGTGATGACATTAAGTCCCAAGTTGGTGCTACTATTTTACACAGAGTTTTAGCTAACTTGTTCAGAGAACGTGGTGTTAGATTGGACAGAACCTACCAATTAAACGTCGGTGGTAACACTGACTTCAAGAACATGTTGGAAAGAGAGAGATTGACTTCCAAGAAGATTTCCAAAACTCAAGCCGTTACCTCTCAATTCGACGTCCCAATGGACGCCGACAACATCCACGTTGGTCCATCCGATCATGTTCCTTGGTTGACCGACAGAAAATTGGCTTTCATCCGTTTGGAAGGTACCACCTTTGGTGGTGTTCCATTATCTGCTGAAGTCAAATTAGAAGTTTGGGATTCTCCAAATTCTGCCGGTGTTGTTATCGATGCCGTCAGATGTGCTAAGTTGGCTATGGACCGTGGTCAAGCTGGTGCCTTGACTGGTCCATCTTCTTACTTCATGAAGTCCCCACCACAACAATTCACCGACGAGGAAGCCGGTAGAAGAACCAGAGCCTTCATCGACGATAAGGCTTACGCTTAA |

Table B.2. MIPS DNA sequences from JGI (cont.)

| Organism | UniProt ID | MIPS DNA Sequence (Codon-Optimized for *S. cerevisiae*) |
|---|---|---|
| *Haemonchus contortus* | U6NKU3 | ATGAACGGTTACGCTAACGGTACTGACGCTAATCATCAAAAGCACAAGAGAGTTATCGTTGATTCTCCTTATGTTAGATGTGACGGTAAAGAAATGGAAACTAGATTCTGTTATAGAAAGAATCATTTCTCCCACACCGCTGACGGTTTGAAGGTCACCCCAAAGGAACATGAGTATATTTTCAAGACTCAATTGAAGCCAAAGAAGACCGGTTTGATGTTGGTTGGTATCGGTGGTAACAATGGTTCTACTTCTGTCGGTGCCATTTACGCTAATAAGAAACACATGACCTGGCGTACCAAAGAAGGTATTCAAACTGCTAACTACTTTGGTTCCGTTACTCAATCTTCCACCATTCACTTGGGTTGGGATGGTCAACAACAAATTCATGTCCCATTCAACGAGATCATTCCAATCTTGTCTCCAAACGACTTGATTATTGACGGTTGGGATATCAACAACGCTAACTTGTACCAAGCTATGGTCAGAGCTAAAGTTTTTGAACCAGAATTGCAAGAAAAGTTGAGACCTTACATGGAACCAATTGTTCCAATGCCATCTATCTACTACCCAGATTTCATCGCTGCTAATCAAGGTGACAGAGCTAACAACACTATTCCAGGTACTGACAAGAAGGAACACTTAGAACACATCAGAAGAGACATTAGAAACTTCAAGGCTAAGCATGACTTGGAGTGTGTCATCGTTTTGTGGACCGCTAACACCGAAAGATACACTGATGTTGTTGATGGTTTAAATATGAACGCTGAACAAATCTTGGCTTCTGTTGACGCTTCCGCTGATGAAATCTCTCCATCCAATATTTTCGCTATTGCTGCTATCTTAGAAGGTGCCCACTACATCAACGGTTCCCCACAAAATACTTTGGTCCCAGGTATTATCGATTTGGCTCACAAGCACAATGTCTTTGTCGGTGGTGACGACTTCAAATCTGGTCAAACTAAGATCAAGTCTGCTTTGGTTGATTTCATGGTTTCTTCTGGTTTGAAGCCAGAATCCATTGTCTCCTACAACCACTTGGGTAACAACGACGGTAAGAACTTGTCTGAAGCCAGACAATTTAGATCTAAGGAAATCTCCAAGTCTTCTGTTGTTGATGACATGGTTGAAGCTAACAAGATCTTATACCCTACCGGTCAAAAGCCAGACCACTGTATTGTTATTAAGTATGTCCCATTTGTTGGTGATTCTAAGCGTGCTATGGATGAATACATTTGTTCTATTTTCATGGGTGGTCGTCAAACTTTTGTCATCCACAATACCTGTGAAGACTCCTTATTAGCTACTCCTTTGATCTACGACTTAGCTATCTTGACTGAATTGGCTACCCGTATTCGTTACGCTGATGCCAACGACGGTGAATTCAGATCCTTTCACGAAGTCTTATCTATCTTGTCTTTGTTGTTAAAGGCTCCAGTTGTTCCACCAGGTACTCCAGTTTCCAACGCTTTCATGCGTCAATTCGCTTCCTTAACCAAGTTGATTACCGCCTTGGCCGGTATTTCCGCTGATACTGATATGCAAATTGAATTTTTCACTCAATTACCAAAAGCTAACTAA |

Table B.3. Sequences codon-optimized for *E. coli*

| Name | DNA Sequence (Codon-Optimized for *E. coli*) |
|---|---|
| At4-MIPS-opt | CAGCCAGGATCCGAATTCGTTTATCGAGAGCTTTAAAGTTGAAAGCCCGAACGTGAAATATACCGAAAACGAAATTAACAGCGTGTATGACTATGAAACCACCGAAGTTGTTCATGAAAATCGCAATGGCACCTATCAGTGGGTTGTTAAACCGAAAACCGTGAAATACGATTTCAAAACCGATACACGTGTTCCGAAACTGGGTGTTATGCTGGTTGGTTGGGGTGGTAATAATGGTAGCACCCTGACCGCAGGCGTTATTGCAAATAAAGAAGGTATTAGCTGGGCCACCAAAGATAAAGTTCAGCAGGCAAACTATTTTGGTAGTCTGACCCAGGCAAGCAGCATTCGTGTTGGTAGCTATAATGGCGAAGAAATCTATGCACCGTTTAAAAGCCTGCTGCCGATGGTTAATCCGGAAGATGTTGTTTTTGGTGGTTGGGATATTAGCGATATGAATCTGGCAGACGCCATGGCACGTGCGCGTGTTCTGGATATTGATCTGCAGAAACAGCTGCGTCCGTATATGGAAAATATGATTCCGCTGCCTGGTATCTATGATCCGGATTTTATTGCAGCAAATCAGGGTAGCCGTGCAAATAGCGTTATTAAAGGCACCAAAAAAGAACAGGTGGACCACATCATTAAAGATATGCGCGAATTTAAAGAAAAAAACAAAGTGGATAAACTGGTGGTTCTGTGGACCGCAAATACCGAACGTTATAGCAATGTTATTGTGGGCCTGAATGATACCACAGAAAATCTGCTGGCAAGCGTGGAAAAAGATGAAAGCGAAATTAGCCCGAGCACACTGTATGCAATTGCCTGCGTTCTGGAAGGTATTCCGTTTATTAACGGTAGTCCGCAGAATACCTTTGTTCCGGGTCTGATTGAACTGGCCATTAGCAAAAATTGTCTGATTGGTGGTGATGACTTTAAAAGCGGTCAGACCAAAATGAAAAGCGTCCTGGTTGATTTTCTGGTTGGTGCAGGTATTAAACCGACCAGCATTGTGAGCTATAATCATCTGGGCAATAATGATGGCATGAATCTGAGCGCACCGCAGACCTTTCGTAGCAAAGAAATTAGCAAATCCAACGTGGTTGATGATATGGTTGCAAGCAATGGCATTCTGTTTGAACCGGGTGAACATCCTGATCATGTTGTGGTTATCAAATATGTTCCGTATGTGGCAGATAGCAAACGTGCAATGGATGAATATACCAGCGAAATCTTTATGGGTGGTCGTAATACCATTGTGCTGCATAATACCTGTGAAGATAGCCTGCTGGCAGCACCGATTATTCTGGATCTGGTTCTGCTGGCCGAACTGAGCACCCGTATTCAGTTTAAAGCAGAAGGTGAAGGCAAATTCCATAGCTTTCATCCGGTTGCCACCATTCTGAGCTATCTGACCAAAGCACCGCTGGTTCCGCCTGGTACACCGGTTGTTAATGCACTGAGCAAACAGCGTGCAATGCTGGAAAACATTCTGCGTGCATGTGTTGGTCTGGCACCGGAAAATAACATGATTATGGAATACAAATAATGAAAGCTTGCGGCCGC |
| Bt6-MIPS-opt | CAGCCAGGATCCGAATTCGAAACAAGAAATTAAACCGGCAACCGGTCGTCTGGGTGTTCTGGTTGTTGGTGTTGGTGGTGCAGTTGCAACCACCATGATTGTTGGCACCCTGGCAAGCCGTAAAGGTCTGGCAAAACCGATTGGTAGCATTACCCAGCTGGCAACCATGCGTATGGAAAATAATGAAGAGAAACTGATCAAAGATGTTGTGCCGCTGACCGATCTGAATGATATTGTTTTTGGTGGCTGGGATATCTTTCCGGATAATGCATATGAAGCAGCAATGTATGCAGAAGTGCTGAAAGAAAAAGATCTGAACGGTGTGAAAGATGAACTGGAAGCCATTAAACCGATGCCTGCAGCATTTGATCATAATTGGGCAAAACGTCTGAATGGCACCCATATCAAAAAAGCAGCAACCCGTTGGGAAATGGTTGAACAGCTGCGTCAGGATATTCGTGATTTCAAAGCAGCCAATAATTGCGAACGTGTTGTTGTTCTGTGGGCAGCAAGCACCGAAATCTATATTCCGCTGAGTGATGAACATATGAGCCTGGCAGCACTGGAAAAAGCAATGAAAGATAATAACACCGAAGTGATTAGCCCGAGCATGTGTTATGCATATGCCGCAATTGCAGAAGATGCACCGTTTGTAATGGGTGCACCGAATCTGTGTGTTGATACACCGGCAATGTGGGAATTTAGCAAACAGAAAAATGTTCCGATTAGCGGCAAAGACTTTAAAAGCGGTCAGACCCTGATGAAAACCGTTCTGGCACCGATGTTTAAAACCCGTATGCTGGGTGTTAATGGTTGGTTTAGCACCAATATTCTGGGTAATCGTGATGGTGAAGTTCTGGATGATCCGGATAACTTTAAAACCAAAGAAGTGAGCAAACTGAGCGTGATCGATACCATTTTTGAGCCTGAGAAATATCCGGACCTGTATGGTGATGTTTATCATAAAGTGCGCATCAACTATTATCCGCCTCGCAAAGACAATAAAGAAGCCTGGGATAACATTGATATCTTTGGTTGGATGGGTTATCCGATGGAAATCAAAGTTAATTTCCTGTGCCGTGATAGCATTCTGGCTGCACCGATTGCACTGGATCTGGTTCTGTTTAGCGATCTGGCAATGCGTGCAGGTATGTGTGGTATTCAGACCTGGCTGAGCTTTTTTTGTAAAAGCCCGATGCATGATTTTGAACATCAGCCGGAACATGACCTGTTTACCCAGTGGCGTATGGTTAAACAGACCCTGCGTAATATGATTGGTGAAAAAGAACCGGATTATCTGGCCTGATAAAAGCTTGCGGCCGC |

Table B.3. Sequences codon-optimized for *E. coli* (cont.)

| Name | DNA Sequence (Codon-Optimized for *E. coli*) |
|---|---|
| Cg7-MIPS-opt | CAGCCAGGATCCGAATTCGACCGTTAATAAAGGTATTAGCATTCGCGTGAATAACGTGGGTGATAAAGTGAGCTATAAAGAAAATGAACTGCTGACCAACTATACCTATCATACCAATGTTGTGCATACCAACAGCGATAAAACCCAGTTTGAAGTTACACCGCTGGATAAAAACTACCAGTTTAAAGTGGATCTGAACAAACCGGAACGTCTGGGTGTGATGCTGGTTGGTTTAGGTGGTAATAATGGTAGCACCATGATGGCAGCAGTTCTGGCAAATAAACACAATGTTTGTTTTCGCACCCGTGATAAAGAAGGTCTGACCGAACCGAACTATTATGGTAGTCTGACCCAGAGCAGCACCATTAAACTGGGTGTTGATAGCAAAGGCAAAGATGTTTATGTGCCGTTTAATAGCCTGGTTCCGATGGTTAATCCGAATGATTTTGTTGTTAGCGGCTGGGATATTAATGGCGCAACCATGGATCAGGCAATGGAACGTGCAAGCGTTCTGGAAGTTGATCTGCGTAATAAACTGGCACCGATGATGAAAGATCATAAACCGCTGAAAAGCGTGTATTACCCGGATTTTATTGCAGCCAATCAGGATGAACGTGCAGATAATTGTCTGAATGTTGATCCGCAGACCGGTAAAGTTACCACCACCGGTAAATGGGAACATCTGAATCATATTCGCAATGATATCCGCACCTTTAAACAGCAGAATGATCTGGACAAAGTGATTATTCTGTGGACCGCAAATACCGAACGTTATGTTGAAATTCTGCCTGGTGTTAACGATACCATGGAAAATCTGCTGGAAGCCATCAAAAATGATCATACCGAAATTGCACCGAGCACCATTTTTGCAGCAGCCAGCATTCTGGAACATTGTCCGTATATCAATGGTAGTCCGCAGAATACCTTTGTTCCGGGTCTGATTGAACTGGCCGAAAAAAATGATAGCCTGATTGCCGGTGATGATTTCAAAAGTGGTCAGACCAAAATGAAAAGTGTTCTGGCACAGTTTCTGGTTGATGCAGGTATTCGTCCGGTTAGCATTGCAAGCTATAATCATCTGGGCAATAACGATGGTTACAATCTGAGCAGTCCGCAGCAGTTTCGTAGCAAAGAAATTAGCAAAGCAAGCGTGGTGGATGATATTATTGAAAGCAATCCGATCCTGTACAACGATAAACTGGGCAACAAAATTGATCACTGCATCGTGATCAAATATATGCATGCAGTTGGTGACAGCAAAGTTGCAATGGATGAATATTACAGCGAACTGATGTTAGGTGGCCATAATCGCATTAGCATCCATAATGTTTGTGAAGATAGCCTGCTGGCAACACCGCTGATTATTGATCTGATTGTTATGACCGAATTTTGCAGCCGTGTTACCTATCGTAATGTTGATGGTCAGGATGGTGCCGAAGCAAAAGGTGATTTTGAAAACTTTTATCCGGTGCTGAGCTTTCTGAGCTATTGGCTGAAAGCACCGCTGACCAAACCGGGTTATCAGCCGATTAATGGTCTGAATAAACAGCGTACCGCACTGGAAAACTTTCTGCGTCTGCTGATTGGTCTGCCTGCAATTGATGAACTGCGTTTTGAAGAACGCCTGAAGTAATAAAAGCTTGCGGCCGC |
| Mps15-MIPS-opt | CAGCCAGGATCCGAATTCGTATTATTTCGATCGTGGTAATGTGATGGATAAAATCAAAATTGCCATTGCCGGTGTTGGTAATTGTGCAAGCAGCCTGATTCAGGGCATTGAATATTATCGTGATAAACACGAAAACGATGCCATTGGTCTGATGCATTGGGATATTGGTGGTTATCGTCCGAGCGATATTGAAGTTGTTGCAGCCTTTGATATCGACAAACGTAAAGTGGGTAAAGATATTAGCGAAGCCATTTTTGCACCGCCTAATTGTACCGCAATTTTTTGTAGCGATATTCCGCAGCGTGGTGTTGTTGTTAAAATGGGTTGTATTCTGGATGGCTTTAGCGAACACATGATGGATTTTGATGAAAAACGTACCTTTGTGCCGAGCGATCAGCCGGAAGCAAGCAAAGAAGGTGTTGTTCAGGCACTGAAAGATAGCGGTGCAGAAATTCTGCTGAACTATCTGCCGGTTGGTAGCGAACAGGCAACCCGCTTTTATATGGATTGTGCACTGGATGCGGGTGTTGCATGTGTGAATAATATGCCGGTTTTTATTGCAAGCGATCCGGAATGGGCAGCCAAATTTGAAAAACGCGGTATTCCGATTATTGGCGACGATATTAAAGCACAGCTGGGTGCAACCATTACACATCGTATGCTGGCAGACCTGTTTAACAAACGTGGTGTTAAACTGGAACGTACCTATCAGCTGAATACCGGTGGTAATACCGATTTTCTGAATATGCTGAATCGTAGCCGTCTGGCAAGCAAAAAAACCAGCAAAACCGAAGCAGTTCAGAGCGTTCTGGCACAGCGTCTGGATGATGATAACATTCATGTTGGTCCGAGTGATTATGTTCCGTGGCAGAATGATAATAAAGTGTGCTTTCTGCGCATGGAAGGTAAACTGTTTGGTGATGTTCCGATGAATCTGGAACTGCGTCTGAGCGTTGAAGATAGCCCGAATAGCGCAGGCGTTGTTATTGATGCAATTCGTTGTTGTAAACTGGCACTGGATCGTGGCATTGGTGGTGTTCTGTATAGCCCGAGCGCCTATTTTATGAAACATCCGCCTAAACAGTTCACCGATGATGAAGCACACAAAATGACCAGCGAATTTATTCATGGTGATCGCACCAACTGATAAAAGCTTGCGGCCGC |

Table B.3. Sequences codon-optimized for *E. coli* (cont.)

| Name | DNA Sequence (Codon-Optimized for *E. coli*) |
|---|---|
| Si20-MIPS-opt | CAGCCAGGATCCGAATTCGTTTATCGAGAGCTTTAAAGTTGAAAGCCCGAACGTGAAATATACCGAAGGTGAAATTCATAGCGTGTATAACTATGAAACCACCG<br>AACTGGTTCATGAAAGCCGTAATGGCACCTATCAGTGGATTGTTAAACCGAAAACCGTGAAGTATGAGTTCAAAACCGATACACATGTTCCGAAACTGGGTGTT<br>ATGCTGGTTGGTTGGGGTGGTAATAATGGTAGCACCCTGACCGGTGGTGTTATTGCAAATCGTGAAGGTATTAGCTGGGCCACCAAAGATAAAGTTCAGCAGG<br>CAAACTATTTTGGTAGTCTGACCCAGGCAAGCAGCATTCGTGTTGGTAGCTTTAATGGCGAAGAAATCTATGCACCGTTTAAAAGCCTGCTGCCGATGGTTAAT<br>CCGGATGATGTTGTTTTTGGTGGTTGGGATATTAGCAATATGAATCTGGCAGATGCAATGGGTCGTGCAAAAGTTCTGGATATTGATCTGCAGAAACAGCTGCG<br>TCCGTATATGGAACATATGGTTCCGCTGCCTGGTATTTATGATCCGGATTTTATTGCAGCAAATCAGGGTAGCCGTGCCAATAATGTTATTAAAGGCACCAAAAA<br>AGAACAGGTGCAGCAGATCATTAAAGATATGCGCGATTTTAAAGAACAGAACAAAGTGGATAAAGTGGTTGTTCTGTGGACCGCAAATACCGAACGTTATAGC<br>AATGTTGTTGTGGGTCTGAATGATACCGCAGAAAGCCTGATGGCAAGCGTTGAACGTAATGAAGCAGAAATTAGCCCGAGCACACTGTATGCAATTGCCTGTG<br>TTTTTGAAAACGTGCCGTTTATTAACGGTAGTCCGCAGAATACCTTTGTTCCGGGTCTGATTGATCTGGCAATTCAGCGTAATAGCCTGATTGGTGGTGATGATT<br>TCAAAAGCGGTCAGACCAAAATGAAAAGCGTTCTGGTTGATTTTCTGGTTGGTGCAGGTATTAAACCGACCAGCATTGTTAGCTATAATCATCTGGGCAATAAC<br>GATGGCATGAATCTGAGCGCACCGCAGACCTTTCGTAGCAAAGAAATTAGTAAAAGCAACGTGGTGGATGATATGGTTGCAAGCAATGGTATTCTGTATGAAC<br>CGGGTGAACATCCTGATCATATTGTGGTTATCAAATATGTGCCGTATGTGGGTGATAGCAAACGTGCAATGGATGAATATACCAGCGAAATCTTTATGGGTGGC<br>AAAAGCACCATTGTTCTGCATAATACCTGTGAAGATAGCCTGCTGGCAGCACCGATTATTCTGGATCTGGTTCTGCTGGCCGAACTGAGCACCCGTATCCAGCT<br>GAAAGCAGAAGGTGAAGGTAAATTTCATTCATTTCATCCGGTTGCCACCATTCTGAGCTATCTGACCAAAGCACCGCTGGTTCCGCCTGGTACACCGGTTGTTAA<br>TGCACTGAGCAAACAGCGTGCAATGCTGGAAAATATTCTGCGTGCATGTGTTGGTCTGGCACCGGAAAATAACATGATCCTGGAATACAAATAATGAAAGCTT<br>GCGGCCGC |
| Hc31-MIPS-opt | CAGCCAGGATCCGAATTCGAATGGTTATGCAAATGGCACCGATGCCAATCATCAGAAACATAAACGTGTTATTGTGGATAGCCCGTATGTTCGTTGTGATGGTA<br>AAGAAATGGAAACCCGTTTTTGCTACCGCAAAAACCATTTTAGCCATACCGCAGATGGTCTGAAAGTTACCCCGAAAGAACACGAGTATATCTTTAAAACCCAG<br>CTGAAACCGAAAAAGACAGGTCTGATGCTGGTTGGTATTGGTGGTAATAATGGTAGCACCAGCGTTGGTGCAATTTACGCAAACAAAAAACATATGACCTGGC<br>GCACCAAAGAAGGTATTCAGACCGCAAACTATTTTGGTAGCGTTACCCAGAGCAGCACCATTCATTTAGGTTGGGATGGTCAGCAGCAGATTCATGTTCCGTTT<br>AATGAAATTATCCCGATTCTGAGCCCGAACGATCTGATTATTGATGGTTGGGATATTAACAACGCCAATCTGTATCAGGCAATGGTTCGTGCAAAAGTTTTTGAA<br>CCGGAACTGCAAGAAAAACTGCGTCCGTATATGGAACCGATTGTTCCGATGCCGAGCATCTATTATCCGGATTTCATTGCAGCAAATCAGGGTGATCGTGCCAA<br>TAATACCATTCCGGGTACAGATAAAAAAGAGCATCTGGAACACATTCGTCGTGATATCCGTAACTTTAAAGCCAAACATGATCTGGAATGCGTTATTGTTCTGTG<br>GACCGCAAATACCGAACGTTATACCGATGTTGTTGATGGCCTGAATATGAATGCAGAGCAGATTCTGGCAAGCGTTGATGCAAGCGCAGATGAAATTAGTCCG<br>AGCAACATTTTTGCAATTGCCGCAATTCTGGAAGGTGCCCATTATATCAATGGTAGTCCGCAGAATACCCTGGTTCCGGGTATTATCGATCTGGCACATAAACAC<br>AATGTTTTCGTTGGTGGTGATGACTTTAAAAGCGGTCAGACCAAAATCAAAAGCGCACTGGTTGATTTTATGGTTAGCTCAGGTCTGAAACCGGAAAGCATTGT<br>TAGCTATAATCATCTGGGCAACAACGATGGTAAAAATCTGAGCGAAGCACGTCAGTTTCGTAGCAAAGAAATTAGCAAAAGCAGCGTGGTTGATGATATGGTT<br>GAAGCCAACAAAATTCTGTATCCGACCGGTCAGAAACCTGATCATTGTATCGTTATCAAATATGTGCCGTTTGTGGGTGATAGCAAACGTGCAATGGATGAATA<br>TATCTGCAGCATTTTTATGGGTGGTCGTCAGACCTTTGTGATTCATAATACCTGTGAAGATAGCCTGCTGGCAACACCGCTGATTTATGATCTGGCCATTCTGACC<br>GAACTGGCAACCCGTATTCGTTATGCAGATGCAAATGATGGTGAATTTCGCAGCTTTCATGAAGTTCTGAGCATTCTGAGCTTACTGCTGAAAGCACCGGTTGTT<br>CCGCCTGGTACACCGGTTAGCAATGCATTTATGCGTCAGTTTGCAAGCCTGACCAAACTGATTACCGCACTGGCAGGTATTAGCGCAGATACCGATATGCAGAT<br>TGAATTTTTCACCCAGCTGCCGAAAGCCAACTAAAAGCTTGCGGCCGC |

Table B.4. Amino acid residue positions for INO1 used for sequence analysis

| **Group** | **Absolutely conserved** | **Conserved eukaryotic** | **Within 5Å of active site** | **5-10 Å from active site** | **Conserved "blocks"** | **Conserved eukaryotic "block"** |
|---|---|---|---|---|---|---|
| **Source** | Dastidar & Chatterjee, 2006 | Dastidar & Chatterjee, 2006 | Jin et al., 2004 | Jin et al., 2004 | Basak et al., 2017 | Basak et al., 2017 |
| **Residues** | | | | | | |
| | 325 | 146 | 71 | 69 | 292 | 66 |
| | 352 | 147 | 72 | 70 | 293 | 67 |
| | 354 | 148 | 74 | 73 | 294 | 68 |
| | 360 | 149 | 75 | 79 | 295 | 69 |
| | 369 | 241 | 76 | 80 | 296 | 70 |
| | 400 | 242 | 77 | 81 | 297 | 71 |
| | 402 | 243 | 78 | 146 | 298 | 72 |
| | 412 | 244 | 147 | 151 | 299 | 81 |
| | 438 | 245 | 148 | 152 | 300 | 82 |
| | 489 | 246 | 149 | 153 | 346 | 83 |
| | | 247 | 150 | 154 | 347 | 84 |
| | | 248 | 160 | 156 | 348 | 85 |
| | | 249 | 184 | 157 | 349 | 86 |
| | | 250 | 185 | 158 | 350 | 87 |
| | | 293 | 186 | 161 | 351 | |
| | | 294 | 191 | 162 | 352 | |
| | | 295 | 198 | 163 | 353 | |
| | | 296 | 243 | 181 | 354 | |
| | | 297 | 244 | 182 | 355 | |
| | | 298 | 245 | 183 | 356 | |
| | | 299 | 246 | 187 | 357 | |
| | | 300 | 247 | 188 | 366 | |
| | | 302 | 248 | 189 | 367 | |
| | | 303 | 277 | 190 | 368 | |
| | | 304 | 281 | 192 | 369 | |
| | | 308 | 295 | 194 | 370 | |
| | | 309 | 296 | 195 | 371 | |
| | | 318 | 297 | 196 | 372 | |
| | | 319 | 320 | 197 | 373 | |
| | | 320 | 321 | 199 | 374 | |
| | | 322 | 322 | 200 | 410 | |
| | | 323 | 323 | 201 | 411 | |
| | | 324 | 324 | 202 | 412 | |
| | | 325 | 325 | 203 | 413 | |

Table B.4. Amino acid residue positions for INO1 used for sequence analysis (cont.)

| Group | Absolutely conserved | Conserved eukaryotic | Within 5Å of active site | 5-10 Å from active site | Conserved "blocks" | Conserved eukaryotic "block" |
|---|---|---|---|---|---|---|
| **Source** | Dastidar & Chatterjee, 2006 | Dastidar & Chatterjee, 2006 | Jin et al., 2004 | Jin et al., 2004 | Basak et al., 2017 | Basak et al., 2017 |
| **Residues** | | | | | | |
| | | 326 | 326 | 204 | 414 | |
| | | 348 | 327 | 211 | 415 | |
| | | 349 | 350 | 223 | 416 | |
| | | 350 | 352 | 226 | 417 | |
| | | 351 | 354 | 242 | 418 | |
| | | 352 | 355 | 249 | 433 | |
| | | 353 | 356 | 275 | 434 | |
| | | 354 | 360 | 276 | 435 | |
| | | 355 | 369 | 278 | 436 | |
| | | 356 | 373 | 279 | 437 | |
| | | 357 | 402 | 280 | 438 | |
| | | 360 | 410 | 292 | 439 | |
| | | 369 | 412 | 293 | 440 | |
| | | 372 | 438 | 294 | 441 | |
| | | 373 | 439 | 298 | | |
| | | 374 | 442 | 300 | | |
| | | 376 | 489 | 319 | | |
| | | 378 | | 328 | | |
| | | 379 | | 329 | | |
| | | 400 | | 330 | | |
| | | 402 | | 335 | | |
| | | 412 | | 348 | | |
| | | 438 | | 349 | | |
| | | 489 | | 351 | | |
| | | | | 353 | | |
| | | | | 357 | | |
| | | | | 358 | | |
| | | | | 359 | | |
| | | | | 361 | | |
| | | | | 365 | | |
| | | | | 366 | | |
| | | | | 367 | | |
| | | | | 368 | | |
| | | | | 370 | | |
| | | | | 371 | | |

Table B.4. Amino acid residue positions for INO1 used for sequence analysis (cont.)

| Group | Absolutely conserved | Conserved eukaryotic | Within 5Å of active site | 5-10 Å from active site | Conserved "blocks" | Conserved eukaryotic "block" |
|---|---|---|---|---|---|---|
| **Source** | Dastidar & Chatterjee, 2006 | Dastidar & Chatterjee, 2006 | Jin et al., 2004 | Jin et al., 2004 | Basak et al., 2017 | Basak et al., 2017 |
| **Residues** | | | | | | |
| | | | | 372 | | |
| | | | | 374 | | |
| | | | | 376 | | |
| | | | | 398 | | |
| | | | | 400 | | |
| | | | | 401 | | |
| | | | | 403 | | |
| | | | | 404 | | |
| | | | | 405 | | |
| | | | | 408 | | |
| | | | | 409 | | |
| | | | | 411 | | |
| | | | | 413 | | |
| | | | | 414 | | |
| | | | | 415 | | |
| | | | | 416 | | |
| | | | | 418 | | |
| | | | | 421 | | |
| | | | | 428 | | |
| | | | | 434 | | |
| | | | | 435 | | |
| | | | | 436 | | |
| | | | | 437 | | |
| | | | | 440 | | |
| | | | | 441 | | |
| | | | | 443 | | |
| | | | | 444 | | |
| | | | | 445 | | |
| | | | | 446 | | |
| | | | | 449 | | |
| | | | | 486 | | |
| | | | | 487 | | |
| | | | | 488 | | |
| | | | | 490 | | |
| | | | | 503 | | |

Table B.5. His-tagged MIPS protein expression at 30°C as measured by volume normalized to total protein

| JGI # | Volume normalized to total protein |
|---|---|
| 0 | 213207851 |
| 1 | 215223459 |
| 2 | Not detected |
| 3 | 289432274 |
| 4 | 186416266 |
| 5 | 37779991 |
| 6 | 55036810 |
| 7 | 356345438 |
| 8 | 376738280 |
| 9 | 48405086 |
| 10 | Not detected |
| 11 | 95667834 |
| 12 | 108602900 |
| 13 | Not detected |
| 14 | Not detected |
| 15 | 36777581 |
| 16 | 266534670 |
| 17 | 282095224 |
| 18 | 284547034 |
| 19 | 137201046 |
| 20 | 288600190 |
| 21 | 86304966 |
| 22 | 151584031 |
| 23 | 65518722 |
| 24 | 90964213 |
| 25 | 198848391 |
| 26 | 39818545 |
| 27 | Not detected |
| 28 | 23925600 |
| 29 | Not detected |
| 30 | 6523368 |
| 31 | 119320926 |

Table B.6. Selected amino acid differences relative to INO1 and At4 MIPS

| | MIPS Variants | | | | | | | | | | | Factor $R^2$ | | Selected Mutations | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 4 | 5 | 7 | 8 | 11 | 18 | 20 | 24 | 25 | 31 | Stability | Titer | INO1 | At4 | Pr18 |
| 30°C MI titer normalized by expression | 0.45 | 0.43 | 0.90 | 0.27 | 0.17 | 0.42 | 0.24 | 0.26 | 0.68 | 0.33 | 0.88 | | | | | |
| Relative MI titer at 37°C vs. 30°C | 0.34 | 0.07 | 0.57 | 0.39 | 0.56 | 0.10 | 0.01 | 0.53 | 0.37 | 0.10 | 0.49 | | | | | |
| Optimum growth temperature | 30 | 25 | 24 | 25 | 22.5 | 40 | 20 | 25 | 35 | 22.5 | 30 | 0.00 | 0.13 | | | |
| **Amino acid differences relative to INO1** | | | | | | | | | | | | | | | | |
| L66T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.07 | 0.30 | | | |
| I68M | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.15 | 0.34 | | | |
| M69L | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.14 | 0.15 | | | |
| L81M | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0.06 | | | |
| L81F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.11 | 0.03 | | | |
| L81S | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.07 | 0.30 | | | |
| V82M | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.13 | 0.21 | V82M | | |
| V82T | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0.00 | 0.09 | | | |
| A83G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0.18 | 0.14 | A83G | A79G | |
| S84G | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.05 | 0.12 | | | |
| S84A | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.00 | 0.03 | | | |
| L86E | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.12 | 0.00 | | | |
| L86Y | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.07 | 0.30 | | | |
| N150S | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0.00 | 0.01 | | | |
| S184G | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.00 | 0.00 | | | |
| L242M | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.15 | 0.34 | | | |
| Y250F | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0.38 | 0.06 | Y250F | | F233Y |
| F281Y | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.00 | 0.00 | | | |
| Y292F | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.05 | 0.12 | | | |

Table B.6. Selected amino acid differences relative to INO1 and At4 MIPS (cont.)

| | **MIPS Variants** | | | | | | | | | | | **Factor R$^2$** | | **Selected Mutations** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 4 | 5 | 7 | 8 | 11 | 18 | 20 | 24 | 25 | 31 | Stability | Titer | INO1 | At4 | Pr18 |
| L308F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.01 | 0.08 | | | |
| L321F | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.05 | 0.02 | | | |
| S374A | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0.06 | | | |
| S374Q | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.14 | 0.15 | | | |
| S374K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.11 | 0.03 | | | |
| 411N | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.15 | 0.34 | | | |
| V413R | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0.01 | 0.23 | V413R | | |
| M415L | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0.01 | 0.05 | | | |
| H433F | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.15 | 0.34 | | | |
| V435T | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0.02 | 0.02 | | | |
| **Amino acid differences relative to At4** | | | | | | | | | | | | | | | | |
| N24G/T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0.00 | 0.01 | | | |
| D31N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.09 | 0.00 | | | |
| Y120F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.09 | 0.00 | | | |
| D151N | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0.17 | 0.12 | N151D | D146N | |
| A159G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.09 | 0.00 | | | |
| N178H/S | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.09 | 0.00 | | | |
| D221Q | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.08 | 0.00 | | | |
| H222Q | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0.00 | 0.00 | | | |
| D271N | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0.06 | 0.06 | | | |
| S273A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.09 | 0.00 | | | |
| L287F/M | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.09 | 0.00 | | | |
| G289N | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0.23 | 0.00 | | | |
| S311Q/E | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0.07 | 0.03 | | | |
| A409G | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.16 | 0.00 | | | |
| M528L | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0.00 | 0.03 | | | |

B.2. Python Scripts

B.2.1. Python script to calculate and return the number of amino acid IMGT differences relative to a given sequence for all sequences in a given multiple alignment file and a given list of amino acid positions indexed relative to the sequence of interest.

```
#file name for FASTA file containing multiple alignment
align_file = "promals3d_mult_align_20181010_formatted.fasta"

#reference sequence name (string) within multiple alignment file
ref_seq_name = "sp_P11986_INO1_YEAST_Inos"

#file name containing amino acid residue position numbers
select_file = "Ino1_5A_resi_positions.txt"

#Output node attribute file name
outfile_name = "INO1_5A_diff_IMGT_20181013_corrected.na"
output_text = "\t5A diff IMGT from Ino1" #node attribute column name

from Bio import SeqIO
import numpy

#Import position numbers
literature_pos = numpy.loadtxt(select_file, dtype = int)

#Get aligned positions from literature positions
alnList = []
count = 0

for seq_record in
    SeqIO.parse(align_file,"fasta"):
    if ref_seq_name in seq_record.id:
        for pos in range(len(seq_record.seq)):
            char_check = seq_record.seq[pos]
            if char_check.isalpha(): #if char is letter, count it
                count = count +1 #this is literature position
                if count in literature_pos:
                   #add position to list
                   alnList.append(pos)

#initialize reference residue ID dictionary (position residue ID dictionary)
posResIDDict = {}

#Make dictionary of reference residues at positions of interest
for position in alnList:
    #Get residue(s) from reference sequence(s) at given position
    for seq_record in SeqIO.parse(align_file,"fasta"):
        #Get residue at position of interest
        ResID = seq_record.seq[position]

        #Convert to IMGT class
            if (ResID == 'A' or ResID == 'I' or ResID == 'L' or ResID == 'V'):
                ResID = 1
            elif (ResID == 'N' or ResID == 'Q'):
                ResID = 2
            elif (ResID == 'M' or ResID == 'C'):
                ResID = 3
            elif (ResID == 'S' or ResID == 'T'):
                ResID = 4
            elif (ResID == 'R' or ResID == 'K' or ResID == 'H'):
                ResID = 5
            elif (ResID == 'D' or ResID == 'E'):
                ResID = 6
            elif (ResID == 'F'):
                ResID = 7
            elif (ResID == 'W'):
                ResID = 8
            elif (ResID == 'Y'):
                ResID = 9
            elif (ResID == 'P'):
                ResID = 10
            elif (ResID == 'G'):
                ResID = 11

        #print seq_record.id,ResID

        #Populate residue dictionary for reference sequence
        ResIDList = []
        if ref_seq_name in seq_record.id:
            if ResID in ResIDList: #useful for residues in more than one sequence
                continue
            else:
                ResIDList.append(ResID)
                posResIDDict[position] =
                ResIDList
                #print seq_record.id,position,ResIDList

#print posResIDDict

#Initialize dictionary of number of differences from reference sequence(s)
#(distance count dictionary)
DistanceCountDict = {}
count = 0
for seq_record in SeqIO.parse(align_file,"fasta"):
      DistanceCountDict[seq_record.id] = count

#Get number of conserved residues from reference sequence(s)
for position in alnList:
    for seq_record in SeqIO.parse(align_file,"fasta"):
        #Get residue at position of interest and check if it matches reference
        ResID = seq_record.seq[position]

        #Convert to IMGT class
            if (ResID == 'A' or ResID == 'I' or ResID == 'L' or ResID == 'V'):
                ResID = 1
            elif (ResID == 'N' or ResID == 'Q'):
                ResID = 2
            elif (ResID == 'M' or ResID == 'C'):
                ResID = 3
            elif (ResID == 'S' or ResID == 'T'):
                ResID = 4
            elif (ResID == 'R' or ResID == 'K' or ResID == 'H'):
                ResID = 5
            elif (ResID == 'D' or ResID == 'E'):
                ResID = 6
            elif (ResID == 'F'):
                ResID = 7
            elif (ResID == 'W'):
                ResID = 8
            elif (ResID == 'Y'):
                ResID = 9
            elif (ResID == 'P'):
                ResID = 10
            elif (ResID == 'G'):
                ResID = 11

        #Count number of residues that match reference
        for ResType in posResIDDict[position]:
            #print ResType
            if ResID != ResType:
                continue
            else:
                #Get number of residue matches found previously and add 1
                count = DistanceCountDict[seq_record.id]
                count = count+1
                DistanceCountDict[seq_record.id] = count

#print DistanceCountDict

#Maximum number of matches/differences is number of residues in alignment list
MaxScore = len(alnList)

#print number of differences to node attribute file
output = open(outfile_name,'w')
output.write(output_text)
output.write('\n')

for seq_record in SeqIO.parse(align_file,"fasta"):
    spliter = seq_record.id.split('_') #match to shared name/ID in network
    if 'sp' in seq_record.id: #Swiss-prot sequences
        ID = spliter[1]
        output.write(ID)
        output.write('\t')
        #Convert number of matches to number of differences
        SpecificScore = DistanceCountDict[seq_record.id]
        Score = MaxScore - SpecificScore
        output.write(str(Score))
        output.write('\n')
    elif 'tr' in seq_record.id: #TrEMBL sequences
        ID = spliter[1]
        output.write(ID)
        output.write('\t')
        #Convert number of matches to number of differences
        SpecificScore = DistanceCountDict[seq_record.id]
        Score = MaxScore - SpecificScore
        output.write(str(Score))
        output.write('\n')
    elif 'zzz' in seq_record.id: #JGI/user-added sequences
        ID = spliter[0]
        output.write(ID)
        output.write('\t')
        #Convert number of matches to number of differences
        SpecificScore = DistanceCountDict[seq_record.id]
        Score = MaxScore - SpecificScore
        output.write(str(Score))  output.write('\n')

output.close()
```

```
empty_str = ""

for seq_record in SeqIO.parse(align_file,"fasta"):
        DistanceCountDict[seq_record.id] = empty_str

#Get different residues from reference sequence(s)
for seq_record in SeqIO.parse(align_file,"fasta"):
    differences = [] #initialize array
    for position in alnList:
        #Get residue at position of interest and check if it matches reference
        ResID = seq_record.seq[position]
        ResType = posResIDDict[position]

        #If not identical, encode in IMGT and check again
        if ResID != ResType:
            if (ResID == 'A' or ResID == 'I' or ResID == 'L' or ResID == 'V'):
                ResID_IMGT = 1
            elif (ResID == 'N' or ResID == 'Q'):
                ResID_IMGT = 2
            elif (ResID == 'M' or ResID == 'C'):
                ResID_IMGT = 3
            elif (ResID == 'S' or ResID == 'T'):
                ResID_IMGT = 4
            elif (ResID == 'R' or ResID == 'K' or ResID == 'H'):
                ResID_IMGT = 5
            elif (ResID == 'D' or ResID == 'E'):
                ResID_IMGT = 6
            elif (ResID == 'F'):
                ResID_IMGT = 7
            elif (ResID == 'W'):
                ResID_IMGT = 8
            elif (ResID == 'Y'):
                ResID_IMGT = 9
            elif (ResID == 'P'):
                ResID_IMGT = 10
            elif (ResID == 'G'):
                ResID_IMGT = 11

            #convert ResType to IMGT class
            ResType_IMGT = ResType
            if (ResType == 'A' or ResType == 'I' or ResType == 'L' or ResType == 'V'):
                ResType_IMGT = 1
            elif (ResType == 'N' or ResType == 'Q'):
                ResType_IMGT = 2
            elif (ResType == 'M' or ResType == 'C'):
                ResType_IMGT = 3
            elif (ResType == 'S' or ResType == 'T'):
                ResType_IMGT = 4
            elif (ResType == 'R' or ResType == 'K' or ResType == 'H'):
                ResType_IMGT = 5
            elif (ResType == 'D' or ResType == 'E'):
                ResType_IMGT = 6
            elif (ResType == F'):
                ResType_IMGT = 7
            elif (ResType == 'W'):
```

B.2.2 Python code to extract amino acid IMGT differences relative to a given sequence for all sequences in a given multiple alignment file and a given list of amino acid positions indexed relative to the sequence of interest.

```
#file name for FASTA file containing multiple alignment
align_file = "Mult_align_31var_only_20181011.fasta"

#reference sequence name (string) within multiple alignment file
ref_seq_name = "sp_P11986_INO1_YEAST_Inos"

#file name containing amino acid residue position numbers
select_file = "Ino1_5A_resi_positions.txt"

#Output node attribute file name
outfile_name = "31var_only_5A_diff_IMGT.txt"


from Bio import SeqIO
import numpy

#Import position numbers
literature_pos = numpy.loadtxt(select_file, dtype = int)
literature_pos = literature_pos.tolist()

#Get aligned positions and IDs from literature positions
#Initialize list of corresponding position numbers wrt alignment file
alnList = []
#Initalize dictionary of literature positions that correspond to aligned positions
orig_pos = {}
#Initialize reference residue ID dictionary (position residue ID dictionary)
posResIDDict = {}

count = 0
for seq_record in SeqIO.parse(align_file,"fasta"):
    if ref_seq_name in seq_record.id:
        for pos in range(len(seq_record.seq)):
            ResID = seq_record.seq[pos]
            if ResID.isalpha(): #if char is letter, count it
                count = count +1 #this is literature position
                if count in literature_pos:
                    #add position to list
                    alnList.append(pos)
                    #get index of orig position in list and associate with entry
                    index_orig = literature_pos.index(count)
                    orig_pos[pos] = literature_pos[index_orig]
                    #add aa char to posResIDDict
                    posResIDDict[pos] = ResID

#Print statements for verification
#print sorted(orig_pos.items(), key=lambda x: x[0])
#print sorted(posResIDDict.items(), key=lambda x: x[0])

#Initialize dictionary of strings of differences from reference sequence(s)
DistanceCountDict = {}
```

```
                ResType_IMGT = 8
            elif (ResType == 'Y'):
                ResType_IMGT = 9
            elif (ResType == 'P'):
                ResType_IMGT = 10
            elif (ResType == 'G'):
                ResType_IMGT = 11

            #Check whether IMGT classes match. If not, add difference to list.
            if ResID_IMGT != ResType_IMGT:
                #save difference in mutation format, using original index
                str_resi = ResType + str(orig_pos[position]) + ResID
                differences.append(str_resi)

    #convert entries to single string
    DistanceCountDict[seq_record.id] = ", ".join(differences)

#print DistanceCountDict
output = open(outfile_name,'w')

for seq_record in SeqIO.parse(align_file,"fasta"):
    spliter = seq_record.id.split('_') #match to shared name/ID in network
    if 'sp' in seq_record.id: #Swiss-prot sequences
        ID = spliter[1]
    elif 'tr' in seq_record.id: #TrEMBL sequences
        ID = spliter[1]
    elif 'zzz' in seq_record.id: #JGI/user-added sequences
        ID = spliter[0]
    output.write(ID)
    output.write('\t')
    output.write(DistanceCountDict[seq_record.id])
    output.write('\n')

output.close()
```