Screening of a library of putative maleate isomerases for the engineering of a synthetic maleic acid-producing *Escherichia coli* strain

Mindy Melgar

A Thesis

in

The Department

of

Biology

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Science (Biology) at

Concordia University

Montreal, Quebec, Canada

November 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

by: Mindy Melgar

entitled: Screening of a library of putative maleate isomerases for the engineering of a synthetic maleic acid-producing *Escherichia coli* strain.

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Biology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

____ Chair

Dr. Grant Brown

____ Examiner

Dr. Malcolm Whiteway

____ Examiner

Dr. Christopher Brett

____ Supervisor

Dr. Vincent Martin

Approved by ____

Chair of Department or Graduate Program Director

____

Dean of Faculty

Date ____

ABSTRACT

Screening of a library of putative maleate isomerases for the engineering of a synthetic maleic acid-producing *Escherichia coli* strain

Mindy Melgar

Efforts to stimulate economic independence while remaining environmentally conscious have motivated researchers to replace non-renewable resources through the engineering of suitable microbes. The efficient generation of polymer precursors, such as maleic acid, through engineered microbial strains will contribute to this aim. This work proposed an engineered strain of *Escherichia coli* to produce the non-native dicarboxylic acid, maleic acid. The strategy included the design of a synthetic metabolic pathway to overproduce fumaric acid, which would serve as a substrate for the heterologously expressed maleate isomerase. The modifications to the *E. coli* strain included 6 knockouts, that resulted in a 10-fold increase in fumaric acid production (7.46mg/l vs. 0.78mg/l). The kinetic characteristics of maleate isomerases had not previously been explored in great detail, which prompted the inclusion of a screening of putative isomerases for optimal activity in the synthetic pathway. A library of 55 maleate isomerase candidates was generated from published literature and bioinformatics databases. Activity of the candidates was screened; the favoured forward reaction (maleic acid to fumaric acid) was observed in 23 variants, and the desired reverse reaction was observed in 13 variants. Maleic acid was successfully produced *in vivo* at a titre of about 100µg/l, which will require significant improvement before industrial implementation is possible. This is however, to our knowledge, the first time a host has been engineered to produce maleic acid using maleate isomerases. Future optimization of this production strain will contribute to a reduced dependence on petroleum as a resource for the polymer precursor, maleic acid.

ACKNOWLEDGEMENTS

Thank you, Dr. Vincent Martin, for taking a chance and granting me the opportunity to study in your lab. I value the entire experience and appreciate your teaching style, which allowed me to learn many skills and challenged me to grow as a student, a scientist, and as an individual. You are always ready to listen and advise, as well as give inspiration to excel while remaining real. Thank you.

I would like to thank my committee members, Dr. Malcolm Whiteway and Dr. Christopher Brett for their guidance, encouragement and thoughtful feedback.

The Martin Lab team, past and present, has been great. Audrey Morin, you hold the whole show together and are always willing to help or consult on matters in and out of the lab, thank you. Kane Larue, thank you for everything you taught me and for making sure I had a solid foundation of lab skills. Nicholas Gold, you have been great to bounce ideas off of and you have been a pronounced source of assurance, thank you. Damien Biot-Pelletier and Elena Fossati, thank you both for all of your encouragement throughout my master's, you were both abundant informational resources. Michael Pyne, Lauren Narcross, Leanne Bourgeois, Mohamed Nasr, Meghan Davies, and Alex Vallerand, thank you for listening and helping me troubleshoot, you all provided valuable perspectives and motivation.

To all of the awesome GE third floor scientists, thank you for being cool.

Marcos Di Falco, thank you for your persistence in helping to optimize my HPLC method.

Thanks goes out to the guys at Hyasynth Bio: Alex Campbell, Zach Wiltshire, and Shoham Mookerjee. You were always willing to help when you were at Concordia and your advice is always appreciated. Shoham, you incite confidence, thank you for being you.

To my friends: You have inspired me and put up with me, thank you.

To my siblings: You have always been there, I know I can turn to you for anything.

To my parents: I have learned so much from you both, thank you for believing in me.

Finally, I would like to deeply thank my husband and best bud, David Tan. You've always supported my goals and you keep seeing me through it all. I couldn't have done it without you.

AUTHOR CONTRIBUTIONS

Mindy Melgar and Vincent JJ Martin conceived the experimental design. Shuyi Lu and Radhakrishna Mahadevan developed the 6-knockout model. Abdelhamid Elsherbiny performed select isomerase over-expressions in BL21(DE3), including corresponding SDS-PAGE analyses. Marcos Di Falco helped optimize the LC-MS detection method of maleic acid and fumaric acid. Mindy Melgar performed all other experiments and data analyses.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The manufacturing of industrially useful compounds using alternatives to non-renewable resources is becoming increasingly important. Numerous useful products are derived from petroleum and crude oil, which are not only finite resources, but often include potentially harmful extraction or modification procedures. These issues have prompted synthetic biology researchers to develop production strategies using metabolically engineered microbial hosts.

One such example would be the production of synthetic polymer precursors from a source other than crude oil or natural gas. An alternative to using non-renewable resources is to use biological hosts to metabolically produce the desired precursors using a renewable feedstock [1, 2]. The metabolism of these biological hosts, such as the bacterium *Escherichia coli*, can be engineered to improve native pathways or gain heterologous enzyme expression pertinent to the compound production [3].

Historically, microorganisms have been modified by directed evolution using selective pressure or screening [3]. This approach can be time-consuming and can result in an accumulation of deleterious mutations alongside the desired ones [3, 4]. The advancements of genetic modification techniques allows for more focused directed evolution of genes of interest and the introduction of specific genetic changes without affecting the rest of the organism's genome, through an approach called rational design [3-5]. Metabolic engineering often includes both focused directed evolution and rational design resulting in the creation of optimal biosynthetic expression pathways. Biosynthetic pathways can grant host microorganisms the ability to produce industrially useful compounds that may only be produced in small amounts by the cell or that are otherwise only available from non-renewable resources or include potentially harmful reaction conditions [2]. **The efforts herein aim to synthetically engineer an *E. coli* strain capable of producing the non-native dicarboxylic acid, maleic acid.**

Maleic acid is a four-carbon dicarboxylic acid that is a suitable substrate for numerous industrially important compounds using simple chemical reactions. For example, maleic acid can be used to generate maleic anhydride by dehydration, fumaric acid by isomerization, malic acid by hydration, and glyoxylic acid by ozonolysis [6]. These products are used in the production of polyester resins, pharmaceuticals, and food additives. Maleic acid is generally derived from maleic anhydride, which had a global market size of USD 2.3 billion in 2015 [7]. Maleic anhydride is produced from benzene or C4 hydrocarbons [6]. In addition to its costs being continuously on the rise [8], benzene is a carcinogenic substance and both it and C4-hydrocarbons are currently sourced from petrochemicals [6]. Given the environmental impacts and the unpredictable market values of crude oil and the petroleum industry, it can be reasoned that it is industrially useful to investigate a renewable and safe alternative for the production of petroleum's downstream products.

E. coli is a suitable host for metabolic engineering and high-titre production of industrially relevant products such as isobutyraldehyde [2] and 1,4-butanediol [9]. In addition to the suitability of *E. coli* as a high production platform, genetic modification techniques have been and continue to be optimized for *E. coli*, making its metabolism easily amenable to change. The genome and native metabolism of multiple *E. coli* strains have been extensively studied and are well documented [10, 11]. Because *E. coli*'s genome is not considered to be redundant, it is possible to use metabolic modelling algorithms to predict the flux of synthetically engineered strains *in silico* [12]. The synthetic pathway models generated from the algorithms can then be implemented *in vivo* to assess the accuracy of model predictions and make adjustments to the biosynthetic pathway based on the findings.

Biosynthetic pathways often include the knockout, knockdown, or overexpression of native genes in order to focus the energy of the organism on the creation of a single compound. While *E. coli* does not natively produce maleic acid, it does produce fumaric acid, a component of the tricarboxylic acid (TCA) cycle [11]. We hypothesized that because fumaric acid is the *trans* isomer of maleic acid, it would serve as a suitable precursor to maleic acid production via a single reaction performed by a heterologous enzyme. In order to make use of this strategy we explored suitable heterologous enzyme candidates for the production of maleic acid. The pursuit of candidate enzymes started with organisms that produce maleic acid natively.

Naturally, maleic acid usually exists in various soil-dwelling bacteria [13-15] as an intermediate metabolite of the pyrrolidine pathway during nicotine degradation [13, 14] or in the nicotinate/nicotinamide metabolism pathways [15]. The bacteria express a maleate isomerase (MI, EC 5.2.1.1) enzyme, which catalyzes the conversion of maleic acid to fumaric acid, thus directing carbon flux into the TCA cycle. Once the carbon is available in the TCA cycle, it can be utilized in numerous pathways, which branch away from central metabolism. There is no other known use for maleic acid in the cell [15]. MIs appear to have evolved to efficiently favour the forward conversion of maleic acid to fumaric acid (Figure 1). The isomerization reaction equilibrium (K_{eq}) favours fumaric acid, which is more thermodynamically stable than maleic acid. The K_{eq} value has been estimated to be between a theoretical value of 47 [16], and an *in vitro* estimation of 500 (with the *Alcaligenes faecalis* MI) [17], suggesting that only up to about 2% conversion to maleic acid could be observed. Furthermore, it has been hypothesized by Dokainish *et al*. that in the forward reaction, after the maleic acid substrate is subject to a nucleophilic attack to form the enzyme-substrate complex, and following intermediate steps, the dissociation of the fumaric acid product is concerted by the release of non-covalent bonds [18]. Although the reverse reaction was not discussed in that study, the conformation of fumaric acid may result in a decreased chance of nucleophilic attack to initiate the enzyme-substrate complex. Forward activity was previously confirmed for MIs from 7 bacteria using *in vitro* activity assays (Table 1), however investigation into the reversibility of MIs was generally overlooked.

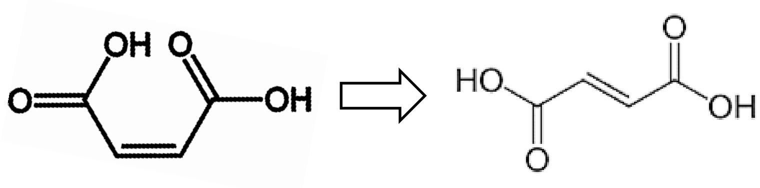


Figure 1: Forward reaction of maleate isomerases.

Maleate isomerases have been characterized as having the ability to catalyze the isomerization of the maleic acid (left) to fumaric acid (right). Examination of the reverse reaction was largely omitted from previous research.

Knowledge of enzyme activity is crucial for the creation of efficient biosynthetic pathways. Enzymes are often annotated as having a particular activity based on enzyme assays with specifically chosen substrates, as can be seen online in The Comprehensive Enzyme Information System, BRENDA [19]. For example, the enzyme called malease (or maleate hydratase, 4.2.1.31) was discovered with an assay containing prepared corn slurry and maleic acid as a substrate resulting in the formation of malic acid [20]. Initially we considered the possibility of utilizing a malease for the production of maleic acid. However, we found a later report that described isopropylmalate isomerase (4.2.1.33) which primarily acts on isopropylmalate and can also accept maleic acid as a substrate [21], even though the compound and enzyme do not naturally exist together [15]. We hypothesized that the catalytic activity observed with the corn slurry was most likely a side reaction due to the promiscuity of isopropylmalate isomerase. Because it is not feasible to assay all conceivable substrates, there is the possibility that enzymes have some promiscuous activities, which may be similar to the functions of other distinct enzymes. These reasons led us to believe that the isopropylmalate isomerase would not be a suitable candidate for the production of maleic acid because if the desired specific catalytic activity is a side reaction, there may be an increased chance in the formation of other non-specific products.

Returning to MI candidates, it should be noted that enzyme homologs found in the NCBI database are often annotated based on amino acid sequence similarity meaning that the actual activity is unknown. Because the annotation might not match the catalytic activity, for the purposes of designing a synthetic metabolic pathway, it is often necessary to test different candidate enzymes. Considering these factors, we hypothesized that MI homologs may vary in kinetic characteristics and specifically in the degree to which fumaric acid is accepted as a substrate for MI's reverse reaction of maleic acid formation. For this reason, we constructed and screened a library of putative MIs to identify a candidate with the highest fumaric acid to maleic acid activity.

At the time that potential MI candidates were being investigated (early 2015), no consensus sequence existed for MIs with the exception of 2 catalytic cysteines, which is characteristic of 4 of the 5 members of Aspartate/Glutamate racemase superfamily, PF01177, to which MIs belong [22]. In this work, we have defined a consensus sequence for the breathing loop regions of active MIs and have found additional conserved residues throughout the protein. Previously described MIs are typically around 250 amino acids long. Using this information along with the 7 previously published MI sequences, we identified from the NCBI database 48 putative MIs originating from a variety of prokaryotic organisms (Table 2). The bioinformatics query suggested that MIs do not exist in eukaryotic organisms.

Certain organisms, such as *Cupriavidus basilensis*, harboured 2 putative MIs [23]. To resolve whether one was better suited for this experiment than the other, an attempt was made to determine if either existed within an operon or cluster of genes belonging to the same pathway as MIs. The MI from *Pseudomonas putida* S16 (MI-PpuS, one of the 7 previously characterized MIs) is found within a genomic cluster of genes called *nic*2 that mediate the later steps of nicotine degradation [14]. The *nic*2 cluster described for *P. putida S16* includes MI-PpuS and four other genes in close proximity: hydroxy-succinoylpyridine hydroxylase (HSP), N-formylmaleamate deformylase (NFO), dihydroxy pyridine dioxygenase (HPO), and maleamate amidase (AMI). Many of the genomes from which the putative MIs originated were vaguely annotated as having coding sequences belonging to broad families or simply as “hypothetical ORFs”, making it difficult to discern which MIs might be promising based solely on their genetic environment.

We hypothesized that there might be a correlation between the forward and reverse MI reactions and that if the forward reaction could be demonstrated, the selection of candidates for the reverse reaction would require less guesswork. The idea to test the forward activity was also pursued because its detection is possible by simple use of a spectrophotometer with absorbance readings at 290nm. This activity assay was first reported in 1969 [24], and has been used numerous times since [13, 17, 25, 26].

Where possible, we expressed and purified putative MIs and performed *in vitro* activity assays with the goal of discovering those that favour the production of maleic acid. The isomerization of fumaric acid to and from maleic acid is not a spontaneous reaction due to the high activation energy required to break the double carbon bond about which the isomerization occurs [18]. Although MIs have been shown to favour the forward reaction (maleic acid → fumaric acid), enzymes may be able to favour the production of one species over the other in reversible reactions by having more favourable reverse maximum velocity (V_{maxREV}) and Michaelis-Menten constant (K_{MREV}) [27]. We therefore screened MIs in both directions. Our *in vitro* assays showed that the reverse reaction is indeed unfavoured in MIs and that reversible

MIs can convert up to about 2% fumaric acid overnight. The optimal candidate enzymes were then subject to *in vivo* assays via expression in the synthetic *E. coli* strain.

We incorporated a model for the overproduction of fumaric acid to serve as a base strain for the production of non-native maleic acid. A six knockout (6KO) *E. coli* model overproducing fumaric acid was designed for this research by our collaborators (Lu & Mahadevan, University of Toronto, 2015 communications) using a minimal cut set algorithm. The six KOs are pyruvate kinase I (Δ*pykF*), pyruvate kinase II (Δ*pykA*), NAD-requiring malate dehydrogenase (Δ*sfcA*), NADP-requiring malate dehydrogenase (Δ*maeB*), fumarate reductase (Δ*frdA*), and glutamate dehydrogenase (Δ*gdhA*). These knockouts should cause the metabolic flux to evade pyruvic acid production and ensure fumaric acid production. We hypothesized that the expression of a MI in a fumaric acid-producing *E. coli* strain would result in the reverse isomerization reaction to produce maleic acid.

Previously MIs have received attention as possible candidates for the biosynthetic production of fumaric acid [13]. However, this research was primarily interested in whether or not MIs can be employed in a synthetically engineered organism to exploit the reverse, often overlooked reaction to produce maleic acid. If maleic acid is successfully sustainably produced in a biosynthetic platform strain, it could serve as a precursor to and facilitate the production of many downstream molecules such as fumaric acid, maleic anhydride, and glyoxylic acid [6].

This research provides an initial proof-of-concept for the production of maleic acid in a synthetically engineered *E. coli* strain. This biosynthetic pathway is now available for further optimization and may be attempted in other, more robust organisms capable of dealing with high titres of acid production, such as *Saccharomyces cerevisiae* [28]. The implications of an improved maleic acid-production platform microorganism will be beneficial in efforts to reduce reliance on the petroleum industry in the production of maleic acid and any downstream applications.

Table 1: List of partially characterized MIs and their published characteristics.

ID	Accession	Species	Opt. pH	Crystal structure	k_{cat}	K_M (µM)	k_{cat}/K_M (s^{-1},M^{-1})	Rev. rxn?	Cofactors	Stabilizer/ activator	Ref(s)
MI-PpuS	4FQ7_A	*Pseudomonas putida* S16	8.4	complete							[13]
MI-Rjo	ABG92314	*Rhodococcus jostii* RHA1	8		96.7 ± 1.3	149.6 ± 9.2	6.5 x 10^5				[29]
MI-Nfa	2XEC_D	*Nocardia farcinica*	7.5	complete	2.8 ± .7	4.6 ± .5	6.1 x 10^5	unk.			[30]
MI-Bst	BAA77296.1	*Bacillus stearothermophilus*		Cys80 & 198					None		[26]
MI-Sma	BAA96747.1	*Serratia marcescens*		Cys80 & 198							[25]
MI-Afa	BAA23002.1	*Alcaligenes faecalis*	8	three Cys		40		weak (<2%)	None	reducing agents	[17, 31]
MI-Pfl	YP_002872377.1	*Pseudomonas fluorescens*	~8.4?		30	~300?		no evid	mercaptans/ thiols	ethylene	[24]

Abbreviations: Opt. pH: optimal pH; Rev.rxn?: presence of reverse reaction; unk.: specified as unknown. Empty cells denote information that was not included in the publication(s).

*Definition of units used: * "One unit of maleate cis–trans isomerase activity was defined as the amount of enzyme that catalyzed the formation of 1 µmol of fumaric acid from maleic acid in 1 min."* [17, 31] *** "One unit of maleate cis-trans isomerase corresponds to the amount of enzyme which increases 0.001 of the optical density at 295 mu in 1 min at 28± 1 °C."* [24]

Table 2: List of MIs selected for analysis in this study.

Library #:	Accession #:	Isolation	Location	Species	ID	*nic*2 genes
MI01	4FQ7_A	soil	China	*Pseudomonas putida* S16	MI-PpuS	4
MI02	ABG92314	soil	Canada	*Rhodococcus jostii* RHA1	MI-Rjo	2
MI03	2XEC_D	mammal	Japan	*Nocardia farcinica* IFM 10152	MI-Nfa	1
MI04	BAA77296.1	soil	Japan	*Bacillus stearothermophilus*	MI-Bst	1
MI05	BAA96747.1	soil	Japan	*Serratia marcescens*	MI-Sma	3
MI06	BAA23002.1	NA	NA	*Alcaligenes faecalis*	MI-Afa	0
MI07	CAY49026.1	leaf	UK	*Pseudomonas fluorescens* SBW25	MI-Pfl	4
MI08	WP_043876743.1	soil	China	*Ralstonia solanacearum* FQY_4	MI-Rso	2
MI09	WP_010813501.1	soil	Australia	*Ralstonia sp.* GA3-3	MI-R.G	2
MI10	WP_009989870.1	NA	NA	*Sulfolobus solfataricus* P2	MI-Sso	1
MI11	AEI76235.1	soil	USA	*Cupriavidus necator* N-1	MI-Cne	4
MI12	AJG21767.1	groundwater	USA	*Cupriavidus basilensis*	MI-Cba1	4
MI13	AJG22759.1	groundwater	USA	*Cupriavidus basilensis*	MI-Cba2	2
MI14	AEQ53768.1	seawater	China	*Pelagibacterium halotolerans* B2	MI-Pha	3
MI15	KHQ54613.1	algae	Malaysia	*Ponticoccus sp.* UMTAT08	MI-P.U	2
MI16	EKE69327.1	ocean	Arctic	*Celeribacter baekdonensis* B30	MI-Cbae	1
MI17	AJE48609.1	ocean	India	*Celeribacter indicus*	MI-Cin	1
MI18	WP_019826135.1	river	USA	*Pseudomonas sp.* CF149	MI-P.C	2
MI19	WP_016498886.1	NA	NA	*Pseudomonas putida* NBRC 14164	MI-PpuN	2

Library #:	Accession #:	Isolation	Location	Species	ID	*nic*2 genes
MI20	EPZ44103.1	soil	Germany	*Alicyclobacillus acidoterrestris*	MI-Aac	0
MI21	ELQ90977.1	free	Japan	*Mycobacterium smegmatis* MKD8	MI-Msm	2
MI22	WP_037812099.1	NA	NA	*Streptomyces sp.* NRRL F-3213	MI-S.N1	1
MI23	WP_037812690.1	NA	NA	*Streptomyces sp.* NRRL F-3213	MI-S.N2	2
MI24	WP_031470027.1	sediment	Malaysia	*Sciscionella sp.* SE31	MI-S.S	1
MI25	WP_013678325.1	sludge	USA	*Pseudonocardia dioxanivorans*	MI-Pdi	1
MI26	WP_020658292.1	mammal	NA	*Amycolatopsis benzoatilytica*	MI-Abe	1
MI27	WP_033414103.1	soil	Thailand	*Actinomycetospora chiangmaiensis*	MI-Ach	0
MI28	WP_037370708.1	NA	NA	*Amycolatopsis orientalis*	MI-Aor	2
MI29	GAB90686.1	root	Japan	*Gordonia rhizosphera* NBRC 16068	MI-Grh	2
MI30	WP_007323639.1	mammal	Japan	*Gordonia araii*	MI-Gar	2
MI31	WP_043451626.1	stream	S. Korea	*Gordonia kroppenstedtii*	MI-Gkr	1
MI32	AHH19904.1	root	Brazil	*Nocardia nova* SH22a	MI-Nno	2
MI33	WP_040834905.1	mammal	NA	*Nocardia brevicatena*	MI-Nbr	0
MI34	ETD31745.1	soil	Antarctic	*Williamsia sp.* D3	MI-W.D	2
MI35	EHI45093.1	soil	NA	*Rhodococcus opacus* PD630	MI-Rop	2
MI36	WP_037176300.1	plant	USA	*Rhodococcus fascians*	MI-Rfa	2
MI37	ELB89809.1	NA	France	*Rhodococcus wratislaviensis* IFP 2016	MI-Rwr	1
MI38	EXG81230.1	soil	Japan	*Cryptosporangium arvum*	MI-Car	2
MI39	WP_019202278.1	mammal	NA	*Tsukamurella sp.* 1534	MI-T.1	0

Library #:	**Accession #:**	**Isolation**	**Location**	**Species**	**ID**	***nic*2 genes**
MI40	ENZ80135.1	sediment	USA	*Ralstonia pickettii* OR214	MI-Rpi	0
MI41	WP_020925465.1	mammal	Denmark	*Achromobacter xylosoxidans*	MI-Axy	0
MI42	BAP38945.1	soil	Japan	*Acinetobacter guillouiae*	MI-Agu	3
MI43	EJL92704.1	root	USA	*Herbaspirillum sp.* CF444	MI-H.C	3
MI44	WP_016348741.1	insect	Japan	*Burkholderia sp.* RPE64	MI-B.R	3
MI45	EXF45907.1	soil	Hungary	*Pseudomonas sp.* BAY1663	MI-P.B	0
MI46	WP_016451831.1	mammal	NA	*Delftia acidovorans*	MI-Dac	0
MI47	YP_006969192.1	mammal	USA	*Bordetella bronchiseptica* 253	MI-Bbr	2
MI48	XP_004528795.1	insect	Italy	*Ceratitis capitata*	MI-Cca	4
MI49	ADX72232.1	soil	Greece	*Arthrobacter phenanthrenivorans* Sphe3	MI-Aph	1
MI50	WP_024934567.1	mammal	Mexico	*Actinomadura madurae*	MI-Ama1	0
MI51	WP_033330551.1	mammal	Mexico	*Actinomadura madurae*	MI-Ama2	0
MI52	WP_028930650.1	animal	NA	*Pseudonocardia asaccharolytica*	MI-Pas	2
MI53	EBA81999.1	marine	Pacific	marine metagenome	MI-meB	0
MI54	ECH22039.1	marine	Pacific	marine metagenome	MI-meC	0
MI55	EDH12629.1	marine	Pacific	marine metagenome	MI-meD	0

The number of genes potentially matching members of the nic2 *cluster from* P. putida *S16 is denoted out of a possible maximum of 4. Accession numbers are current as of April 2017 with the exception of record* XP_004528795.1 *(MI48), which has been removed from the NCBI database.*

MATERIALS AND METHODS

Strains, media, and culture conditions

Chemically competent *E. coli* cells (DH5α and BL21(DE3)) were prepared to $OD_{600} \cong 0.6$ using the $CaCl_2$ method [32] to enable plasmid transformations. The *E. coli* strain MG1655 (K-12, CGSC 7740: λ^-, rph^{-1}) was used as wild-type and as the background for the 6KO strain, which was built using iterative P1 phage transductions [33]. Single KO strains for each of the 6KOs were obtained from the Keio collection [34]. A complete list of strains built in this study is available in Appendix Table 1.

E. coli cells were incubated at 37°C with shaking (200rpm for shake flask cultures, 300rpm for 2ml deep-well 96-well plates), in either Luria Bertani (LB) media for protein expression and for starter cultures, or in minimal M9 media with 2% glucose (w/v) when metabolite analyses were to be performed. Media was supplemented with kanamycin (50µg/ml) or ampicillin (50µg/ml) when plasmid selection was necessary. LB agar plates were prepared with antibiotics at the above concentrations except for those used in the P1 phage transductions, which were prepared at half kanamycin concentration (25µg/ml).

Saccharomyces cerevisiae strain CEN.PK113-10C with histidine auxotrophy was used for preliminary MI expression screening. *S. cerevisiae* cells were incubated at 30°C, 200rpm, in yeast nitrogen base without amino acids (YNB, 6.8g/l), with yeast synthetic drop-out medium supplements without histidine (HIS^-, 1.92g/l), and 2% glucose (w/v).

Plasmid construction

The nucleotide sequences of the 55 selected MIs were codon optimized for expression in *E. coli* while also excluding the following restriction enzyme recognition sites: *Aar*I, *Asc*I, *Bsa*I, *Kas*I, *Nco*I, *Not*I, *Sap*I, and *Xho*I, using IDT's codon optimization tool (idtdna.com/CodonOpt) and the Benchling online research tool (Benchling.com). The optimized MI open reading frames (ORFs) were synthesized and cloned into our pBOT expression vector containing a green fluorescent protein (GFP) tag and a HIS complementation marker [35]. The MI coding sequences were amplified from the pBOT vectors for ligation using primers designed to add restriction enzymes recognition sites to the candidate ORFs. *Nde*I or *Ase*I recognition sites were added to the 5` ends of ORFs while *Xho*I sites were added to the 3` ends. Inserts built with *Nde*I sites also contained an *Eco*RI site as a part of the 5' flanking sequence. The appropriate restriction enzymes (NEB) were then used to digest the ORF inserts (*Nde*I/*Ase*I/*Eco*RI and *Xho*I) and backbone plasmids (*Nde*I and *Xho*I for pET41b, *Eco*RI and *Sal*I for pTRC99a). The inserts were then ligated at ambient temperature for a minimum of 20min using T4 Ligase (NEB) into the expression plasmids. The pET41b vector was used in strain BL21(DE3) for IPTG-inducible overexpression and subsequent purification of the MIs via a C-terminal 8His-tag, while pTRC99a was used for *in vivo* heterologous MI expression in the 6KO strain. Assembled plasmids were designated pETMI-(3-4 letter abbreviated origin species) for the pET41b-derived plasmids and pTRCMI-(same abbreviated origin species) for the pTRC99a-derived plasmids. The pTRC99a-derived plasmids were designed post codon-optimization of the MI ORFs and thus only certain pTRCMI- variants were built due to the presence of an *Eco*RI site within the ORFs. Furthermore, the pTRCMI- variants corresponding to inactive or insoluble MI variants were not pursued. All clones were confirmed by colony PCR before being sequence verified (Operon). The complete list of primers used in this study for plasmid construction is available in Appendix Table 2.

DNA manipulations

Plasmids were purified from clonal *E. coli* colonies using the GeneJET Plasmid Miniprep Kit (Thermo). Linearized plasmids were extracted from 0.8% (w/v) agarose gel after electrophoresis and purified using the GeneJET Gel Extraction Kit (Thermo). Vectors were transformed into the appropriate *E. coli* strains either by the heat-shock method [36] for chemically competent cells or by electroporation [37].

Bioinformatics and phylogenies of candidate enzymes

To date, the activity of only seven MIs have been published (Table 1), with two of those having their structures elucidated [13, 30]. The sequences of four other characterized MI variants were also provided in the corresponding publications [17, 25, 26, 29]. The seventh MI discussed in the literature comes from *Pseudomonas fluorescens* ATCC 23728 [24], which did not have a published genome at the start of this project (early 2015). Therefore the putative MI from *P. fluorescens* SBW25 was selected as a stand-in for the seventh partially characterized MI. The 7 sequences were used as BLASTp [38] queries to generate the 55 MI candidate library (Table 2). Care was taken to include candidates from a wide range of origin species with the expectation that phylogenetic relatedness would often correlate to enzyme activity, substrate affinity, or specificity as has been seen with other enzyme library analyses [39, 40].

The presence of *nic*2 cluster genes surrounding the candidate MIs in their native genomes was determined using the Nucleotide Graphics information provided for each MI candidate in the National Center for Biological Information (NCBI) database. The annotations from ±4-ORF positions starting at the candidate MI genomic location were recorded (summary in Table 2, complete list in Appendix Table 4) and the functions described were matched to those of the cluster. Many genomes encountered had not been fully annotated at the latest time of access (May 2017). ORFs with vague identifiers could not be hypothesized to be a part of the *nic*2 cluster. For example, the identifier “alpha/beta hydrolase” was considered to possibly be the NFO member of the *nic*2 cluster, but those simply called “hypothetical ORF” were not considered as *nic*2 members. Other trends such as the presence of transcriptional regulators within the ±4-ORFs were notable in Appendix Table 4.

The amino acid sequences (without any tags) were aligned using MEGA6 software [41] with the default parameters of MUSCLE multiple sequence alignment. The putative MI of *Rhodococcus jostii* RHA1 (ABG92314.1) was reported to be soluble only upon removal of 48 N-terminal residues [29] (Del48), therefore the Del48 version was used here (MI-Rjo). A neighbour-joining phylogenetic tree was built using the Jones-Taylor-Thornton model with 100 bootstrap replicates to visualize the relatedness of the 55 putative MIs (Figure 2).

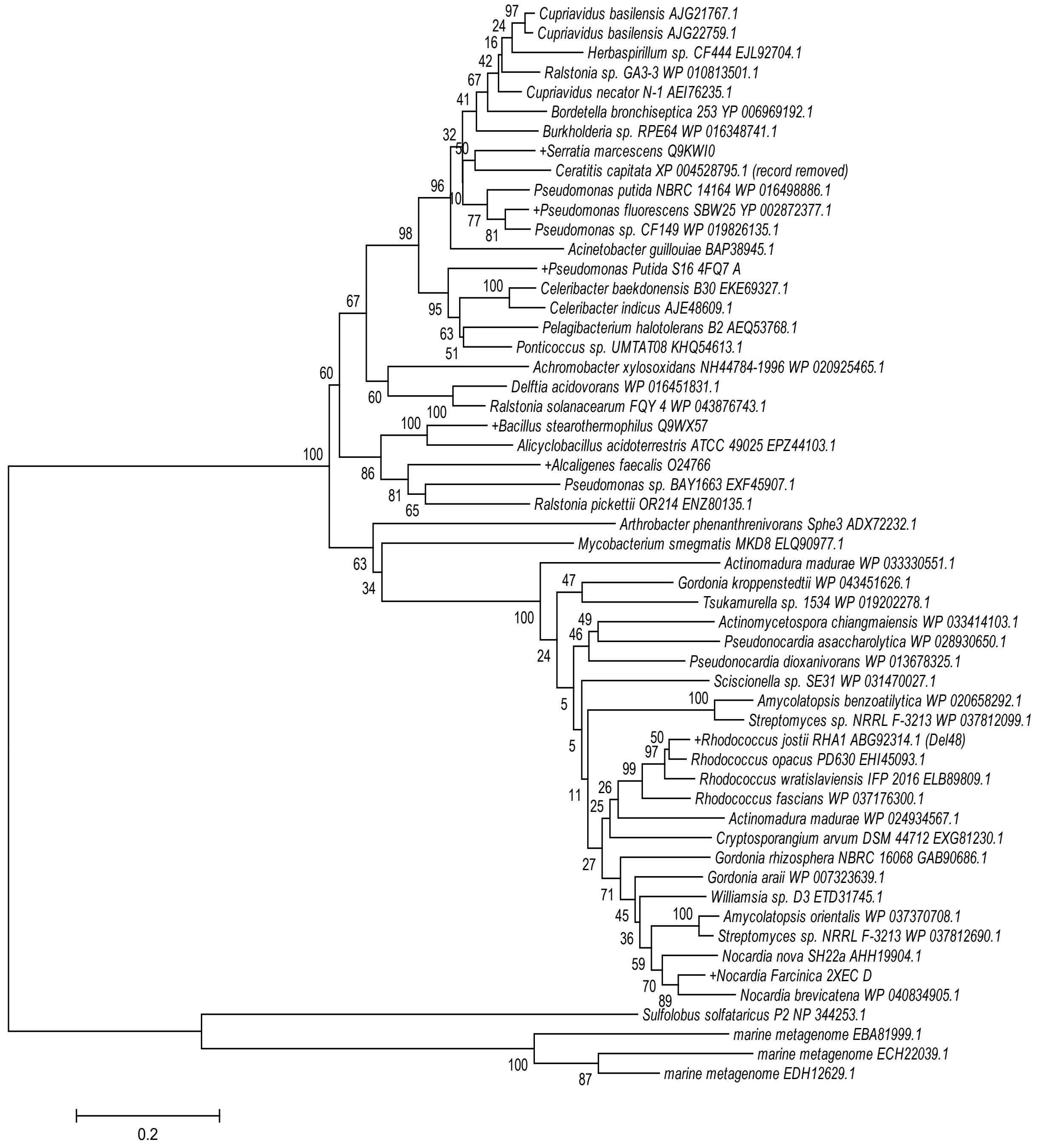

97
24
Cupriavidus basilensis AJG21767.1
Cupriavidus basilensis AJG22759.1
16
Herbaspirillum sp. CF444 EJL92704.1
42
Ralstonia sp. GA3-3 WP 010813501.1
67
Cupriavidus necator N-1 AEI76235.1
41
Bordetella bronchiseptica 253 YP 006969192.1
Burkholderia sp. RPE64 WP 016348741.1
32
50
+Serratia marcescens Q9KWI0
Ceratitis capitata XP 004528795.1 (record removed)
96
10
Pseudomonas putida NBRC 14164 WP 016498886.1
+Pseudomonas fluorescens SBW25 YP 002872377.1
77
81
Pseudomonas sp. CF149 WP 019826135.1
98
Acinetobacter guillouiae BAP38945.1
+Pseudomonas Putida S16 4FQ7 A
100
Celeribacter baekdonensis B30 EKE69327.1
67
95
Celeribacter indicus AJE48609.1
Pelagibacterium halotolerans B2 AEQ53768.1
63
51
Ponticoccus sp. UMTAT08 KHQ54613.1
Achromobacter xylosoxidans NH44784-1996 WP 020925465.1
60
Delftia acidovorans WP 016451831.1
60
Ralstonia solanacearum FQY 4 WP 043876743.1
100
100
+Bacillus stearothermophilus Q9WX57
Alicyclobacillus acidoterrestris ATCC 49025 EPZ44103.1
100
+Alcaligenes faecalis O24766
86
Pseudomonas sp. BAY1663 EXF45907.1
81
Ralstonia pickettii OR214 ENZ80135.1
65
Arthrobacter phenanthrenivorans Sphe3 ADX72232.1
Mycobacterium smegmatis MKD8 ELQ90977.1
63
Actinomadura madurae WP 033330551.1
34
47
Gordonia kroppenstedtii WP 043451626.1
Tsukamurella sp. 1534 WP 019202278.1
100
49
Actinomycetospora chiangmaiensis WP 033414103.1
46
Pseudonocardia asaccharolytica WP 028930650.1
24
Pseudonocardia dioxanivorans WP 013678325.1
Sciscionella sp. SE31 WP 031470027.1
5
100
Amycolatopsis benzoatilytica WP 020658292.1
Streptomyces sp. NRRL F-3213 WP 037812099.1
50
+Rhodococcus jostii RHA1 ABG92314.1 (Del48)
5
97
Rhodococcus opacus PD630 EHI45093.1
99
Rhodococcus wratislaviensis IFP 2016 ELB89809.1
26
11
Rhodococcus fascians WP 037176300.1
25
Actinomadura madurae WP 024934567.1
Cryptosporangium arvum DSM 44712 EXG81230.1
27
Gordonia rhizosphera NBRC 16068 GAB90686.1
Gordonia araii WP 007323639.1
71
Williamsia sp. D3 ETD31745.1
45
100
Amycolatopsis orientalis WP 037370708.1
36
Streptomyces sp. NRRL F-3213 WP 037812690.1
Nocardia nova SH22a AHH19904.1
59
+Nocardia Farcinica 2XEC D
70
89
Nocardia brevicatena WP 040834905.1
Sulfolobus solfataricus P2 NP 344253.1
marine metagenome EBA81999.1
marine metagenome ECH22039.1
100
87
marine metagenome EDH12629.1
0.2

Figure 2: Phylogeny of library of 55 MIs.

The protein sequences of 7 partially characterized MI proteins (denoted by '+' ahead of the species name) were used as BLASTp queries to generate this set totalling 55 MI candidates. The origin species or sample and the accession number for each putative MI are included as labels. The neighbour-joining clustering method was used with the Jones-Taylor-Thornton model and 100 bootstrap replicates to build the phylogenetic tree. Bootstrap replicate support is shown on nodes.

Maleate isomerase expression

Expression of the MIs was initially tested in *S. cerevisiae* with the synthesized pBOT vectors containing GFP fusions. Cells were grown to exponential phase and their fluorescence analysed using the BD Accuri C6 flow cytometer with excitation of 488nm and emission of 533nm to determine the percent of GFP-expressing cells (Table 2).

MI overexpression was performed using the *E. coli* expression strain BL21(DE3) transformed with the pETMI- vectors the day prior. All of the resulting transformants from the plates were inoculated into 2ml of LB Kan, which was used to inoculate 50-250ml cultures in Erlenmeyer flasks. Cells were grown at 37°C, 200rpm until $OD_{600} \cong 0.6$ was reached. The flasks were then chilled on ice with swirling to 16°C, at which point MI protein expression was induced with 1mM IPTG. The cells were then subject to a 16-hour incubation at 16°C. Cells were harvested at 9000 x g for 15min, and then suspended in a volume of buffer equal to 1/25 of the starting culture volume. Storage buffer (25mM Tris-HCl, pH 8.0, 300mM NaCl, 15% glycerol (w/v)) was used for whole cell lysate analysis and lysis buffer (NPI10: 50mM NaH_2PO_4, 300mM NaCl, 10mM imidazole, adjusted to pH 8.0 with NaOH) was used for MI purification. Cells suspended in storage buffer were either stored at -20°C for up to 6 months or lysed immediately using sonication on ice (probe amplitude of 30%, sonication on:off ratio of 8s:24s, total sonication time 40-80s). Debris from lysed cells was separated from the supernatant by centrifugation at 30000 x g, 20min, 4°C. For SDS-PAGE analysis, insoluble pellets were suspended in the same volume of storage buffer as the volume of supernatant removed. The soluble portion from lysed cells was used for crude activity analysis or MI purification. Total protein in samples was normalized using values obtained by Bradford Protein quantification assay (Bio-Rad). SDS-PAGE samples were prepared for the whole cell lysate and for the insoluble and soluble portions using 2X SDS-PAGE loading buffer (100mM Tris-HCl, pH 6.8, 4% SDS (w/v), 0.25% bromophenol blue (w/v), 20% glycerol (w/v), 200mM β-mercaptoethanol), boiled for 3min, and vortexed before loading onto a 12.5% acrylamide gel. Gels were run on the Mini-PROTEAN® Tetra Cell System (Bio-Rad) for approximately 50min at 150V before processing with Coomassie Safe Stain (Invitrogen) according to the manufacturer's direction. Alternatively, unstained gels were used for Western blot analysis, carried out as described in the Sambrook textbook [32] using a 1:1500 dilution of primary antibody for the C-terminal 8His tag (His-probe (G-18) rabbit polyclonal IgG, Santa Cruz Biotechnology) and a 1:1000 dilution of secondary antibody with alkaline phosphatase (Anti-rabbit IgG-AP, Sigma) after the proteins were transferred to nitrocellulose membranes (Whatman). The His-tagged MIs were visualized with 30µl BCIP (50mg/ml, in 100% DMF) and 60µl NBT (50mg/ml, in 70% DMF (v/v)) in 10ml AP buffer (100mM Tris-HCl, pH 9.5, 100mM NaCl, 5mM $MgCl_2$).

Maleate isomerase purification

His-tag purification was carried out using Ni-NTA resin reconstituted as described in the QIAgen handbook [42] and stored in lysis buffer. Induced BL21(DE3) cells suspended in lysis buffer were lysed by sonication as described above. One millilitre of reconstituted Ni-NTA resin was added to the soluble portion of lysed cells and incubated in 15ml conical tubes overnight with light agitation at 4°C. The solution was loaded onto a gravity-flow chromatography column fitted with porous 30µm polyethylene filter bed (Bio-Rad) and then washed with 8ml wash buffer (NPI20: 50mM NaH_2PO_4, 300mM NaCl, 20mM imidazole, adjusted to pH 8.0 with NaOH). Elution buffer (NPI250: 50mM NaH_2PO_4, 300mM NaCl, 250mM imidazole, adjusted to pH 8.0 with NaOH) was used to collect samples of 0.5-1.5ml. All flow through was collected and kept separately according to each step for SDS-PAGE analysis and/or troubleshooting purposes. Protein concentration of the elution fractions was determined using the Bradford Protein quantification assay.

Qualitative activity assays

Presence or absence of the forward MI reaction (maleic acid to fumaric acid) was determined spectrophotometrically by following the change in absorbance readings at 290nm [24] on the Infinite® 200 PRO multimode microplate reader (Tecan). Reactions were set up at 30°C in 96-well UV-Star (Greiner) plates and the absorbance was measured every 30s over a period of 30min. The total reaction volume was 200µl (denoting a path length of 0.56cm), consisting of 50µl soluble portion of induced cell lysate, 50mM Tris-HCl, pH 8.0, 5mM β-mercaptoethanol, and 50mM ammonium maleate, pH 8.0. A positive change in absorbance at

290nm over time signified the production of fumaric acid in the reaction mixture. The negative control was set up using crude lysates from BL21(DE3) with induced empty pET41b vector and did not show a change in absorbance over time.

Alternatively, end-point assays were used to monitor the final equilibrium of the reverse MI reaction (fumaric acid to maleic acid). Five micrograms of purified MI in elution buffer was used in 200µl reactions consisting of 2.5mM sodium fumarate and activity buffer (20mM potassium-phosphate, pH 7.6, 5mM β-mercaptoethanol). End-point assays were analyzed by HPLC-UV (Thermo-Finnegan) using the method described below for metabolite analysis. Standard curves for maleic acid were set up using samples prepared in the reaction mixture in order to keep elution times consistent. Standard samples were set up to represent fumaric acid to maleic acid conversions of 0, 1, 2, 3, 4, 5, and 10%, which produced a linear curve (R^2 = 0.9993).

Quantitative activity assays

Overexpressed MIs were purified from cells lysed in lysis buffer for *in vitro* enzymatic activity assays. Reactions were set up at 30°C in 96-well UV-Star (Greiner) plates and the absorbance (OD_{290}) was measured every 8-9.6s over a period of 90-120s. The total reaction volume was 200µl (denoting a path length of 0.56cm), consisting of 1-10µg purified MI, 20mM potassium-phosphate buffer, pH 7.6, 5mM β-mercaptoethanol, and varying concentrations (between 1-5500µM) of ammonium maleate, pH 8.0. Reactions were performed in triplicate. Initial velocity (V_o) values were calculated during the linear phase of the reaction using the conversion equation:

$$[\mu\text{M fumaric acid formed}]/\text{min} = \Delta\text{Abs290/s}*60/7.27\text{e-}5$$

where ΔAbs290/s is the slope based on the change in absorbance measured at 290nm per second. The conversion factor 60/7.27e-5 is based on a standard curve of 4-400µM sodium fumarate giving a linear curve (R^2 = 0.9678) with a slope of 7.27e-5 meaning that each absorbance increase of this amount denotes the formation of 1µM fumaric acid. The value is multiplied by 60 to give the rate of reaction per minute for Michaelis-Menten analysis. Initial velocities (V_o) were fit to Michaelis-Menten curves (enzyme kinetics – Substrate vs. Velocity) using non-linear regression with GraphPad Prism version 6 for Windows (GraphPad Software, La Jolla California USA, www.graphpad.com), which generated kinetic values. The substrate maleic acid also absorbs at 290nm but at a rate 10-fold lower than fumaric acid (slope=8.7e-6, R^2 = 0.9437, 4-400μM maleic acid) and was therefore considered negligible for the sake of V_o calculations concerning less than 10% conversion of maleic acid to fumaric acid.

Model design and strain construction

A 6KO *E. coli* metabolic model for increased fumaric acid production was designed by our collaborators (Lu and Mahadevan, University of Toronto, 2015 communications) using in house algorithms employing minimal cut sets. Their strategy was to design a strain with as few KOs as possible while still demonstrating a significant increase in desired product.

Using the *E. coli* strain MG1655 as a background strain, the 6KO strain was built by iterative P1 phage transductions as described by Thomason *et al.* [33]. The complete list of strains constructed in this study is available in Appendix Table 1. KO strains from the Keio collection containing a kanamycin selection cassette in the place of each knockout [34] served as template for the transductions. Kanamycin selection was removed from each introduced KO using flippase expressed from pCP20 to recombine the flippase recognition target (FRT) sites flanking the selection cassette. Removal of pCP20 was achieved through the temperature sensitivity of the plasmid. pCP20 can be propagated at 30°C on selection plates, which allows expression of the flippase, and can be cured from cells at 42°C. Curing of the plasmid was achieved on LB agar plates for up to 8hrs. Each KO was confirmed by colony PCR before sequencing confirmation. With each additional KO introduced, the previously introduced KOs were confirmed to remain in place. The list of primers used to confirm KO strains is listed in Appendix Table 3.

Metabolite analysis

The supernatant of cultures grown in minimal media (M9 + 2% glucose (w/v), 37°C) was used for metabolite analysis. Culture samples were taken at various time points throughout growth, their OD_{595} was measured, and then the cells were removed by centrifugation for at least 2min at 9000 x g. The supernatant was then either passed through 0.45µm filter syringes or centrifuged a second time before dispensing into sampling vials or 96-well microtitre plates for analysis by HPLC. Twenty microlitres of each supernatant sample was separated on an Aminex HPX-87H column (300 x 7.8 mm, 9 µm) using 10mM sulfuric acid as isocratic mobile phase at a flow rate of 0.4ml/min. Samples were run for 30min each. Detection was by UV/Vis (absorbance at wavelength 210nm) for aconitic acid, α-ketoglutaric acid, fumaric acid, pyruvic acid and acetic acid, or by refractive index (Waters 3100) for glucose. Concentrations of metabolites were determined using linear standard curves calculated by measurements from reference standard samples. LC-MS was performed on Agilent UHPLC Q-TOF with 10µl of supernatant samples mixed with 0.5% formic acid injected into the Phenomenex Synergi^{TM} 4µm Hydro-RP 80Å column using 0.1% formic acid as the mobile phase in either water (A) or acetonitrile (B). A gradient was applied as follows: 2min: 98.5% A at 0.25ml/min, 0.5min: 2% A at 0.25ml/min, 0.1min: 2% A at 0.35ml/min, 1.5min: 2% A at 0.35ml/min, 0.1min: 98.5% A at 0.35ml/min, 3.5min: 98.5% A at 0.35ml/min, 0.1min: 98.5% A at 0.25ml/min, 1.5min: 98.5% A at 0.25ml/min. Maleic acid was detected in negative mode with a m/z of 115.0032 and a retention time of about 1.9min compared to the isomer fumaric acid which eluted at about 2.6min. Although polymer species of both isomers were present in the chromatogram spectra, it should be noted that quantification was based solely on the presence of monomer species.

RESULTS

Diverse maleate isomerase candidate library generation

Previous studies characterizing MIs were available for 7 different bacterial species (Table 1). These studies focused on characterizing the forward (maleic acid to fumaric acid) reaction with only one study suggesting about 0.2% conversion in the reverse direction [17], and one study finding no evidence of the reverse reaction [24]. Since this project required the reverse reaction, we decided to build a library of putative MIs for activity screening. In order to increase the likelihood of finding an appropriate MI for expression in the proposed metabolically engineered *E. coli* strain, and because of the small number of previously characterized MIs, the library is comprised of 55 candidates from a diverse set of origin species (Table 2). The 7 previously characterized MIs are included in the library as well as an additional 48 putative MIs.

Occasionally, two putative MIs appeared for a single species. This was the case with *Cupriavidus basilensis* (MI-Cba1 and MI-Cba2), *Streptomyces sp.* NRRL F-3213 (MI-S.N1 and MI-S.N2), and *Actinomadura madurae* (MI-Ama1 and MI-Ama2). Because the genomic environment surrounding the MIs was mostly uninformative (described in the next section), both candidate MIs were included with the hypothesis that subsequent activity assays would reveal more information about the two candidates.

All of the candidate MI sequences are stated to originate from prokaryotic organisms with the exception of MI-Cca, which was reportedly from *Ceratitis capitata*, the Mediterranean fruit fly. However, the NCBI entry corresponding to this sequence (accession: XP_004528795.1) has since been withdrawn due to suspected bacterial contamination within the genome assembly. A basic local alignment search using the amino acid sequence of MI-Cca as a query yields 99% identity (100% coverage) to an Asp/Glu Racemase from the bacteria *Pluralibacter gergoviae* (accession: KMK11847.1).

The complete library of 55 MIs originate from various species ranging four Phylum (27 Proteobacteria, 25 Actinobacteria, 2 Firmicutes, 1 Crenarcheaota) and six Classes (25 Actinobacteria, 12 Betaproteobacteria, 8 Gammaproteobacteria, 7 Alphaproteobacteria, 2 Bacilli, 1 Thermoprotei). The Gram stain of 26 species from which MIs originated is positive, 14 negative, and 15 were not described. Two of the MIs come from species considered to be symbionts, 15 come from pathogenic species, and 18 were found in species isolated from contaminated environmental sites. The specific distribution of the origin species' traits is outlined in Appendix Table 5.

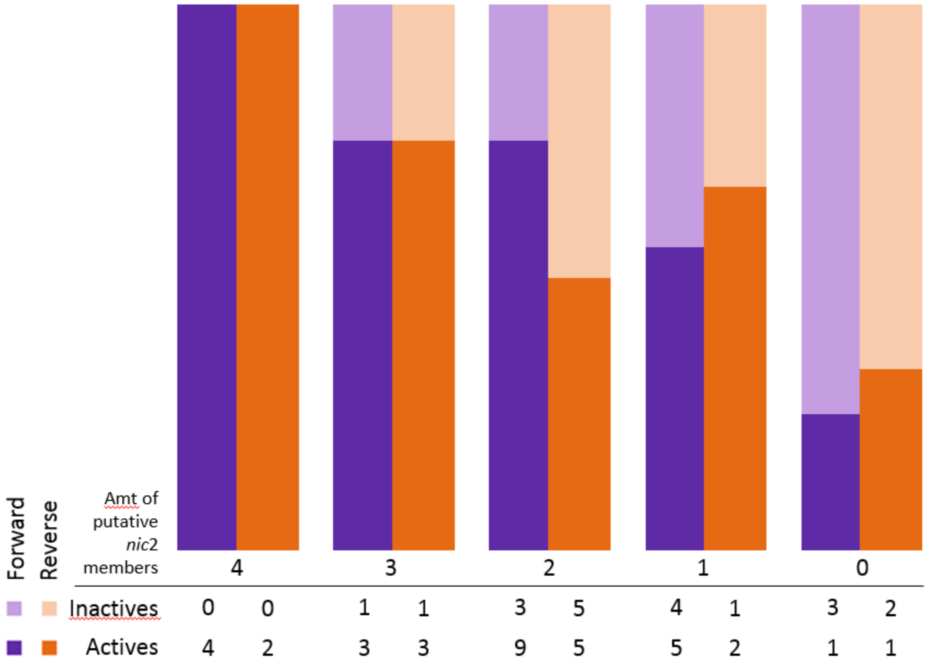
Genetic features and phylogeny of the maleate isomerase library

In addition to the MI candidates having a diverse origin, their genetic makeup displayed diverse trends as well. Interestingly, MI-Cca was the only MI besides MI-PpuS found in a genomic environment containing all four fully specified *nic*2 cluster members within the ±4-

ORF range. The four ORFs comprising the cluster are: HSP, NFO, HPO, and AMI. Information regarding the *nic*2 cluster was otherwise mostly inconclusive due to the lack of complete annotations. The available annotations were interpreted liberally, for example, ORFs given a vague annotation such as “alpha/beta hydrolase” or “oxidoreductase” were considered to be potential NFOs or HPOs, respectively. When the annotation given was “hypothetical protein”, an attempt was made to further clarify the ORF by performing basic local alignment with the “hypothetical protein” sequence, and any similar sequence hits were used as clues to the classification of the ORFs (the complete ±4-ORF classification is available in Appendix Table 4. These investigations led to the distinction that 3 additional MIs (MI-Pfl, MI-Cne, and MI-Cba1) were potentially surrounded by the 4-gene *nic*2 cluster. Another five MIs (MI-Sma, MI-Pha, MI-Agu, MI-H.C, and MI-B.R) appeared to be surrounded by 3 potential cluster members. The majority (19) of the remaining putative MIs were potentially surrounded by 2 members of the *nic*2 cluster, 14 did not yield any similarities with the cluster (often due to short contig lengths or hypothetical ORFS with no definitive hits), and 12 revealed a single possible cluster member. The ratios of active and inactive MIs are presented according to the number of putative *nic*2 members in Figure 3. Since it was not possible to declare the *nic*2 cluster absent due to ambiguous annotations, the 55 MI library includes 3 pairs of candidate MIs originating from three species as mentioned above.

The pairs of MIs originating from the same species grouped together for *C. basilensis* and *A. madurae*, but not for *S. sp.* NRRL F-3213. The phylogeny of candidate MIs in Figure 4 reveals some clustered groups, which may have some correlation to differences observed in activity or possibly some other aspect. The 7 partially characterized MI variants are spread throughout the tree and different predicted clusters. Cluster VII includes MI-Sso from *Sulfolobus solfataricus* P2 and the three marine metagenome MIs (MI-meB, MI-meC, and MI-meD) that are the least related to rest of the library.

A closer look at the alignment of putative MI sequences divulged that 16 residues are conserved throughout the library (Figure 5). This expands on the previously published list of 8 conserved MI residues: P14, N17, V51, C82, V84, Y139, N169, and C200 (Protein Data Bank Accession: 4FQ7 [13]). The additional ones found in this work were: E21, R45, L56, M59, P138, S198, L232, and G245 using the numbering published for MI-PPuS. Conserved residues are located primarily close to the active site in the solved structure. Possible amino acid sequence trends are examined in more detail in the Discussion section.



Forward	Reverse	Amt of putative *nic*2 members	4		3		2		1		0	
■	■	Inactives	0	0	1	1	3	5	4	1	3	2
■	■	Actives	4	2	3	3	9	5	5	2	1	1

Figure 3: Summary of putative *nic*2 cluster members of MIs for which activity was tested.

The presence or absence of MI activity (dark or light shading, respectively) in both the forward and reverse directions (purple and orange, respectively) correlates positively with the amount of putative nic2 cluster members found in the ±4-ORF space. Actual counts of assayed variants are included in the table, a total of 33 MIs were assayed for forward activity and 22 for reverse activity.

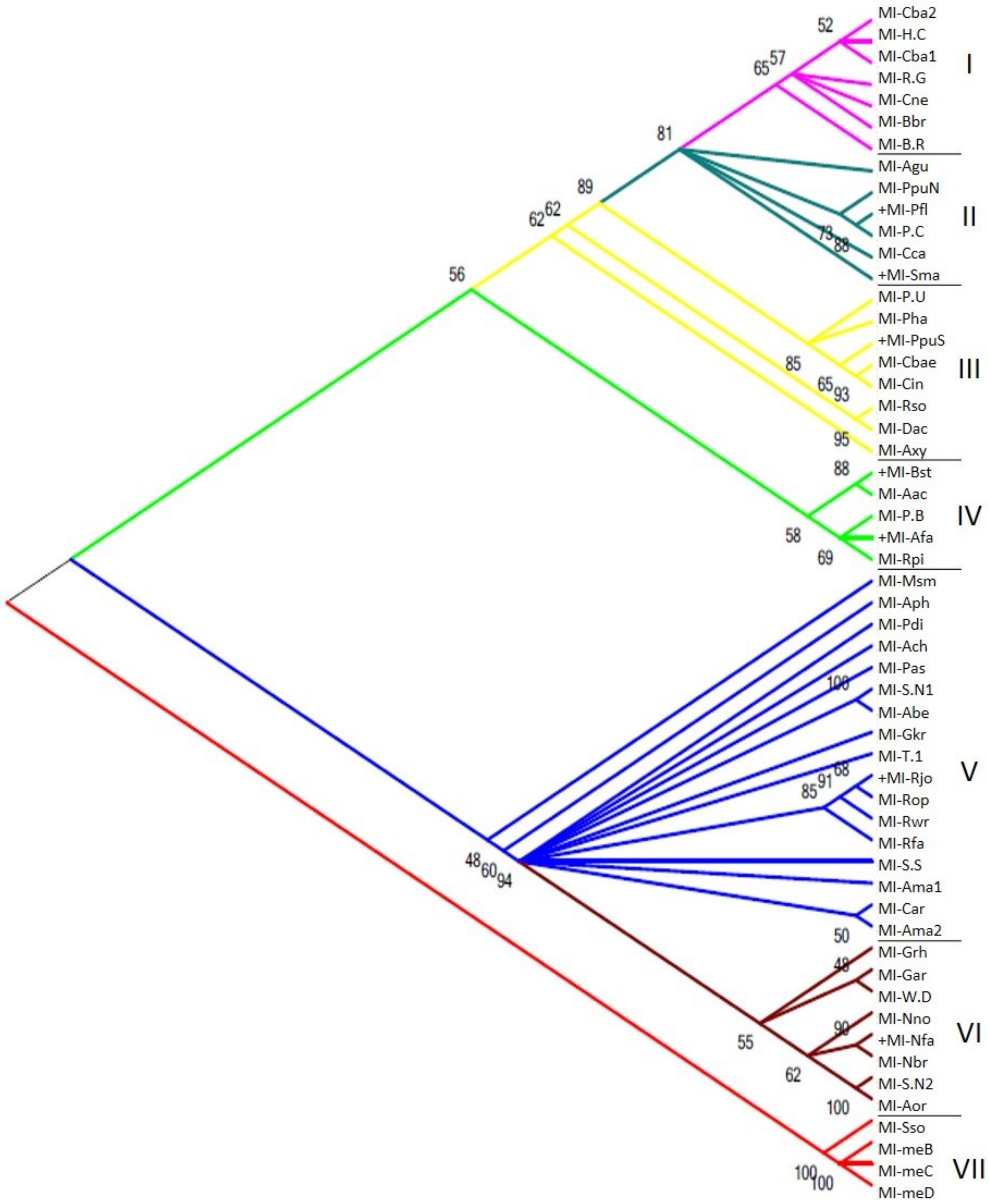


Figure 4: Phylogenetic clusters predicted for MIs.

The phylogenetic tree from Figure 2 is reproduced here to show seven MI clusters (I-VII), which were predicted by utilizing topologies with a 50% cut-off for bootstrap support. MIs are identified by their abbreviated IDs with the 7 previously characterized MIs denoted by a '+'. Bootstrap replicate support is shown on nodes.

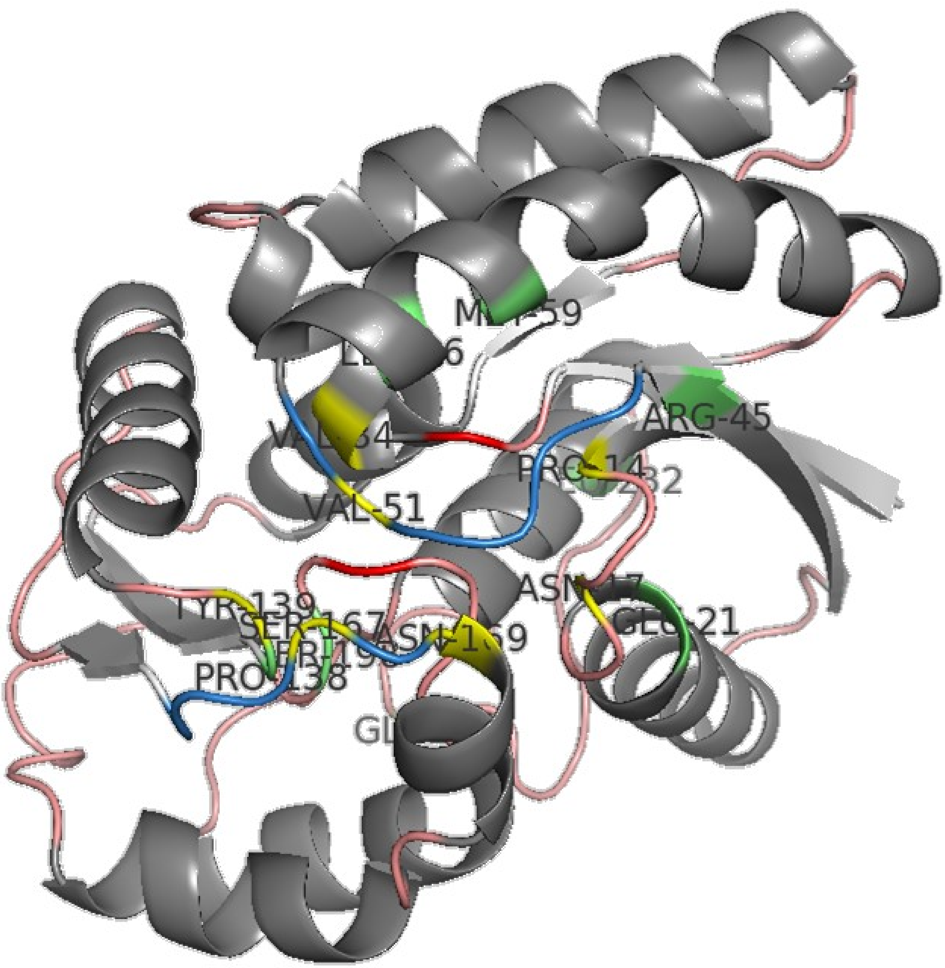


Figure 5: Amino acid sequence analysis of MIs

A monomer of the MI_PpuS structure (PDB: 4FQ7 [13]) is shown with 16 conserved residues apparent in the MI library. Residues in yellow were declared conserved when the structure was published, residues in green were found to be conserved in this study. The conserved residues are: P14, N17, E21, R45, V51, L56, M59, C82, V84, P138, Y139, N169, S198, C200, L232, and G245, the underlined residues were deduced in this study. Residues are largely conserved around the active pocket. Two loop domains (β2-α2 and β6-α7) are described as having a breathing motion and are responsible for locking the substrate in place and preventing side reactions. The structure's helices and sheets are grey, breathing loop domains are blue, other loops are pale pink, and the catalytic cysteines are indicated in red.

Maleate isomerase expression

In order to further characterize the putative MIs, we attempted to express the ORFs in two different ways. First we utilized the *S. cerevisiae* pBOT expression vectors onto which GFP-tagged MI ORFs were synthesized. Flow-cytometry revealed that the percentage of fluorescing *S. cerevisiae* cells harbouring the pBOT vectors was low overall (below 50%, Figure 6). Only 17 of the 55 variant cultures displayed fluorescence in >10% of their cell population. Twenty-two cultures displayed fluorescence in between 1 to 10% of their cells and the remaining 16 had <1% of their cell population fluorescing. An overlay of the GFP fluorescence data and the phylogeny shows that some groups of MIs have higher average fluorescence than other groups, for example, cluster I shows relatively high levels of fluorescence from 4 out of 7 members while the members of cluster II showed only poor fluorescence. The fluorescence levels were meant to be a preliminary test of expression.

A major goal of the MI library screen was to measure the activity of the enzymes *in vitro*. Protein overexpression and SDS-PAGE analysis of cell lysates revealed that 80% (44/55) of the MIs were expressed in a soluble fraction with the conditions used. Six MIs were expressed in the insoluble portion and 3 did not yield any detectable expression at all. We are uncertain of the expression of 2 variants (MI-P.C, MI-Axy) due to unclear SDS-PAGE results. Some example gel results are shown in Figure 7. Of the 44 MIs with soluble expression, half were visible as a wide band around 30kD with total protein Coomassie-stained SDS-polyacrylamide gel separation, 13 required Western blot for visualization, 7 were confirmed soluble by purification from lysate, and 2 were discovered to be soluble during preliminary crude lysate activity assays (Figure 6).

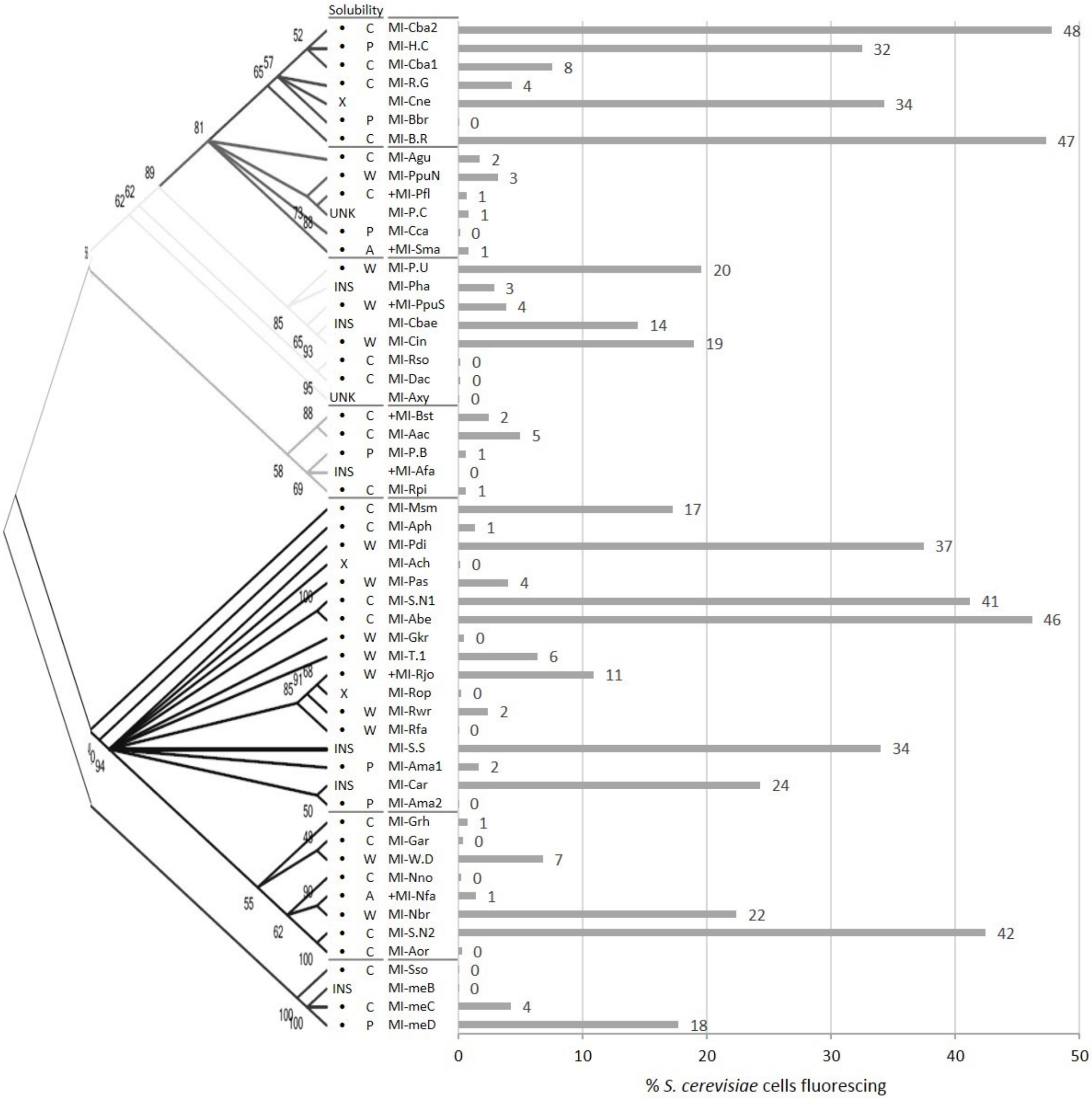


Figure 6: Expression of MIs by GFP levels in *S. cerevisiae* and by overexpression in *E. coli*.

The MI expression was assessed and presented with the phylogeny in order to illustrate trends related to the clustering of the MIs. The percentage of fluorescing *S. cerevisiae* cells when MIs are expressed with GFP attached is shown to the right of the abbreviated MI identifier. To the left the solubility of the overexpressed MIs in *E. coli* is indicated as •: soluble, INS: insoluble, X: not expressed, or UNK: solubility unknown. Soluble MIs were discovered to be as such by C: Coomassie-stained SDS-PAGE, W: Western blot, A: activity of crude lysate, or P: purified.

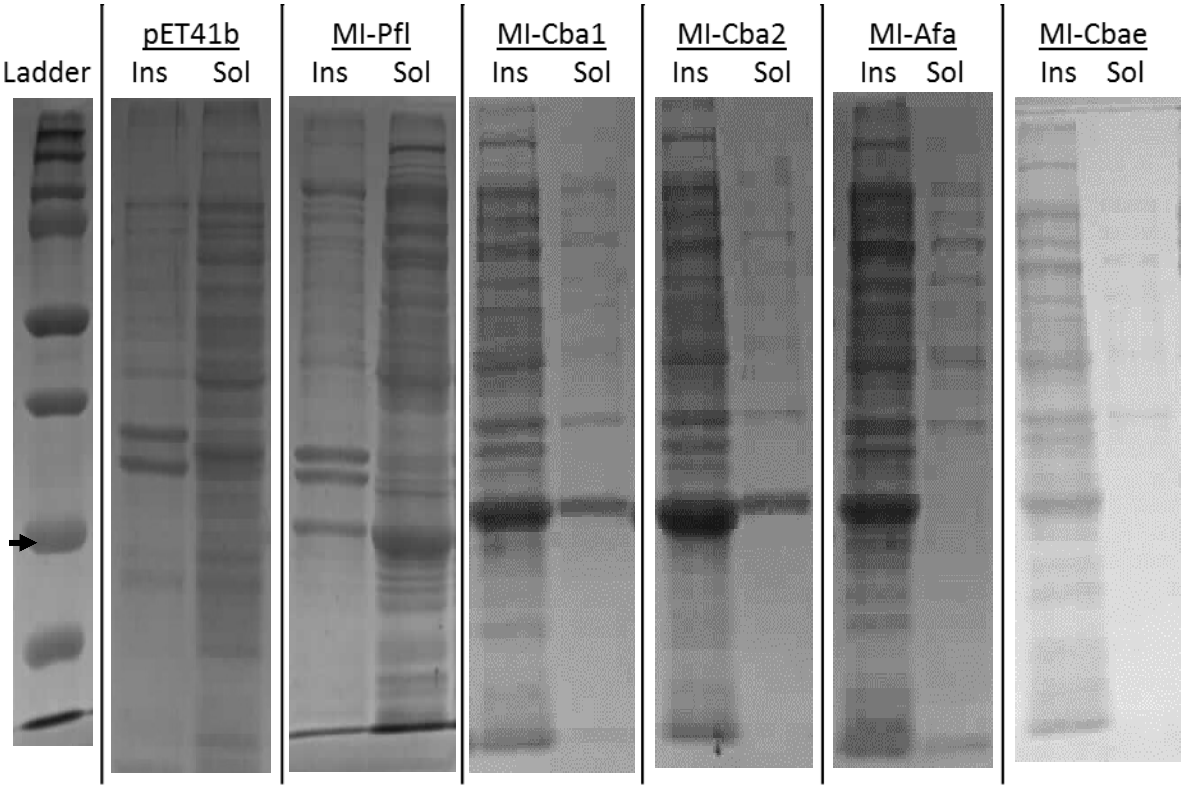


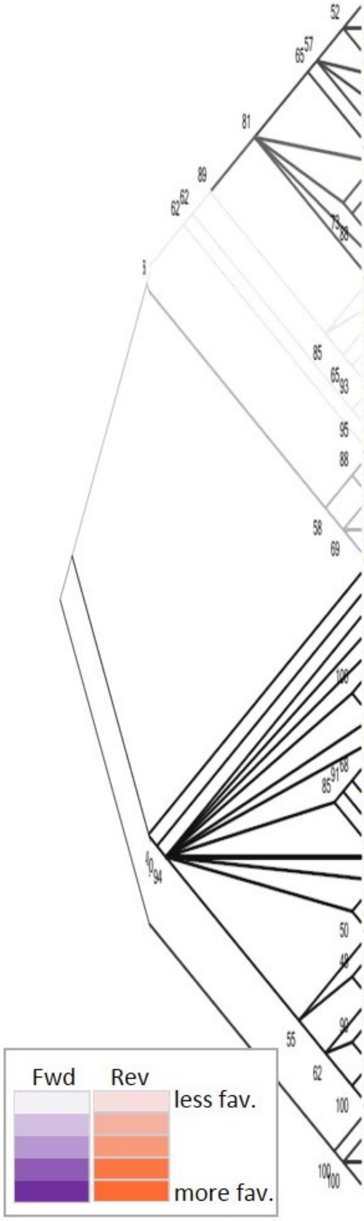
Figure 7: Representative SDS-PAGE results of overexpressed MIs in cell lysate portions.

The expression of pET41b empty vector and 5 MI variants from IPTG-induced pET vectors in *E. coli* BL21(DE3) is shown. Cell extract samples were prepared as either insoluble pelleted solids (Ins) or soluble supernatant (Sol). The arrow indicates the approximate migration of a 30kDa protein, the average size of the MIs. Variants MI-Pfl, MI-Cba1, and MI-Cba2 show soluble MI expression. MI-Afa shows insoluble expression only. The MI-Cbae result shown was inconclusive, thus a Western blot (not shown) was performed revealing that the MI was not visibly expressed.

Maleate isomerase activity screening

We performed detailed time course forward activity assays and qualitative reverse activity assays with 32 purified MIs for kinetic characterizations (Figure 8). Quantifiable forward activity was observed for 18 MIs and while we suspect an additional 3 MIs were active, their activity levels were not high enough for quantitation. The range of maximum velocity (V_{max}) calculated was about 50-fold, from 37µM/min for the MI from *Acinetobacter guillouiae* (MI-Agu) to 1739µM/min for the MI from *Gordonia araii* (MI-Gar). The Michaelis-Menten constant (K_M) values showed a range of almost 180-fold, the least optimal K_M being from MI-Gar (2998µM) and the best measured 16.7µM for the MI from *Nocardia farcinica* IFM 10152 (MI-Nfa). The turnover numbers (k_{cat}) proved to be less varied, ranging only about 10-fold from about $2s^{-1}$ for MI-Rjo to $18.7s^{-1}$ for MI-Cba2. The final kinetic value for the forward reaction which was calculated was the specificity constant, the k_{cat}/K_M ratio, which exhibited an almost 1000-fold range, from $8.7x10^2s^{-1},M^{-1}$ for MI-Rjo to $7.7x10^5s^{-1},M^{-1}$ for MI from *Pseudomonas sp.* BAY1663 (MI-P.B).

In order to assess the ability of MIs to perform the reverse reaction, we screened 22 of the purified MIs with *in vitro* and *in vivo* assays. The percent conversions measured *in vitro* were low, with the highest conversion measured for MI-Cba1 at 1.988%. Many of the enzymes showing forward activity were indeed active in the reverse direction as well. Of the variants showing forward activity, two thirds (10/15) were positive for reverse activity as well. Only 3 MIs (MI-W.D from *Williamsia sp.* D3, MI-H.C from *Herbaspirillium sp.* CF444, and MI-Pas from *Pseudonocardia asaccharolytica*) showed reverse activity without any detectable forward activity. Another 3 MIs (MI-Gkr from *Gordonia kroppenstedtii*, MI-meC, and MI-meD) displayed neither forward nor reverse activity (Figure 8).



	Forward			Forward					Reverse		
ID	Crude assay	Crude act.	Purified	Purified act.	V_{max} (µM/min)	K_m (µM)	K_{cat} (s^{-1})	K_{cat}/K_m (s^{-1}, M^{-1})	Rev assay	% conversion *in vitro*	*in vivo* MA (µM)
Mi-Cba2	•		•	•	989.8	2698	18.714	6936.18	•	1.731	0.901
Mi-H.C			•						•	0.500	
Mi-Cba1	•		•	•	464	680.4	8.946	13148.03	•	1.988	0.660
Mi-R.G	•	•	•	•	545.9	637.3	10.343	16229.38	•	0	
Mi-Cne	•										
Mi-Bbr			•	•	31.89	1898	2.972	1565.94	•	0	
Mi-B.R			•		0	0	0	0	•	0	
Mi-Agu			•	•	37.06	183	0.709	3873.41	•	0.496	
Mi-PpuN	•										
Mi-Pfl	•	•	•	•	143.2	292.8	1.355	4626.61			0.598
Mi-P.C	•										
Mi-Cca			•	•	28.5	151.3	2.7	17844.68	•	0.521	0.368
Mi-Sma	•	•	•	•	341.9	546.6	6.448	11796.99	•	1.949	
Mi-P.U			•		0	0	0	0			
Mi-Pha	•										
Mi-PpuS	•		•		0	0	0	0			
Mi-Cbae	•										
Mi-Cin	•										
Mi-Rso	•	•	•	•	95.12	356.1	1.814	5094.79	•	1.874	0.042
Mi-Dac											
Mi-Axy											
Mi-Bst	•		•								
Mi-Aac			•								
Mi-P.B			•	•	134	6.915	5.34	772284.41	•	1.009	
Mi-Afa	•										
Mi-Rpi											
Mi-Msm	•	•	•	•	132.7	365.5	13.106	35858.65	•	0.512	
Mi-Aph	•	•	•	•	691.2	1179	13.46	11416.43	•	1.527	
Mi-Pdi											
Mi-Ach											
Mi-Pas			•						•	0.509	
Mi-S.N1	•	•	•	•	387	1153	7.273	6307.91	•	1.008	
Mi-Abe	•	•									
Mi-Gkr			•						•	0	
Mi-T.1											
Mi-Rjo			•	•	216	2364	2.055	869.24			
Mi-Rop											
Mi-Rwr	•	•									
Mi-Rfa											
Mi-S.S											
Mi-Ama1											
Mi-Car											
Mi-Ama2											
Mi-Grh	•	•	•	•	510.7	1390	5.054	3636.14	•	0	
Mi-Gar	•	•	•	•	1739	2998	16.329	5446.7	•	0	
Mi-W.D	•		•						•	0.542	0.249
Mi-Nno											
Mi-Nfa	•	•	•	•	115	16.67	10.657	639272.15			
Mi-Nbr											
Mi-S.N2	•		•								
Mi-Aor	•	•	•	•	788.7	1562	7.148	4576.35	•	0	
Mi-Sso	•		•								
Mi-meB											
Mi-meC			•						•	0	
Mi-meD			•						•	0	

Figure 8: Summary of MI catalytic activity.

MI variants were assessed for phylogenetic trends related to catalytic activity. The set of MIs that were subject to the crude assay and those subsequently found to be active are indicated in the first and second columns respectively (described in Appendix I). Kinetics values were obtained with purified MIs. A value of '0' for forward kinetics denotes that minimal activity was detected, and was insufficient for quantitation. The reverse reaction was quantified using end-point assays to assess the percent of 2.5mM fumaric acid converted to maleic acid *in vitro*, where a value of '0' signifies that no maleic acid was detected in the overnight sample. *In vivo* maleic acid production values were determined with the culture supernatant of the 6KO strain expressing select MIs in shake flasks (Appendix Figure 1).

Synthetic model design and strain construction

The reverse MI reaction was previously shown to be very weak *in vitro* [17]. In order to produce the desired product, we hypothesized that the low favourability of maleic acid production could be reconciled by a saturated bioavailability of substrate, fumaric acid. The metabolic algorithm used predicted a model with 6 KOs that would be optimal for the production of fumaric acid while maintaining viability. The 6 genes to be knocked out in the model were: *pykF, pykA, sfcA, maeB, frdA,* and *gdhA* (Figure 9). The metabolic flux was predicted to avoid pyruvic acid production, preferring oxaloacetic acid production from phosphoenolpyruvate, which would lead to L-malic acid, and ultimately to fumaric acid.

The 6KO strain was successfully built using iterative P1 phage transductions. The order by which the KOs were introduced was not by design, as numerous versions were attempted. This was to increase the chance that some KO combinations would work and avoid the possibility that certain KO combinations may have reduced fitness resulting in problematic future P1 phage transductions. The final successful order of KO introduction was: *pykF*, *maeB*, *frdA*, *pykA*, *sfc*, then *gdhA*. The resulting 6KO strain was phenotypically distinct from the WT, which was apparent from culture appearance (Figure 10), and it did experience some lag in growth (Figure 11A) which could be due to the change in metabolic flux.

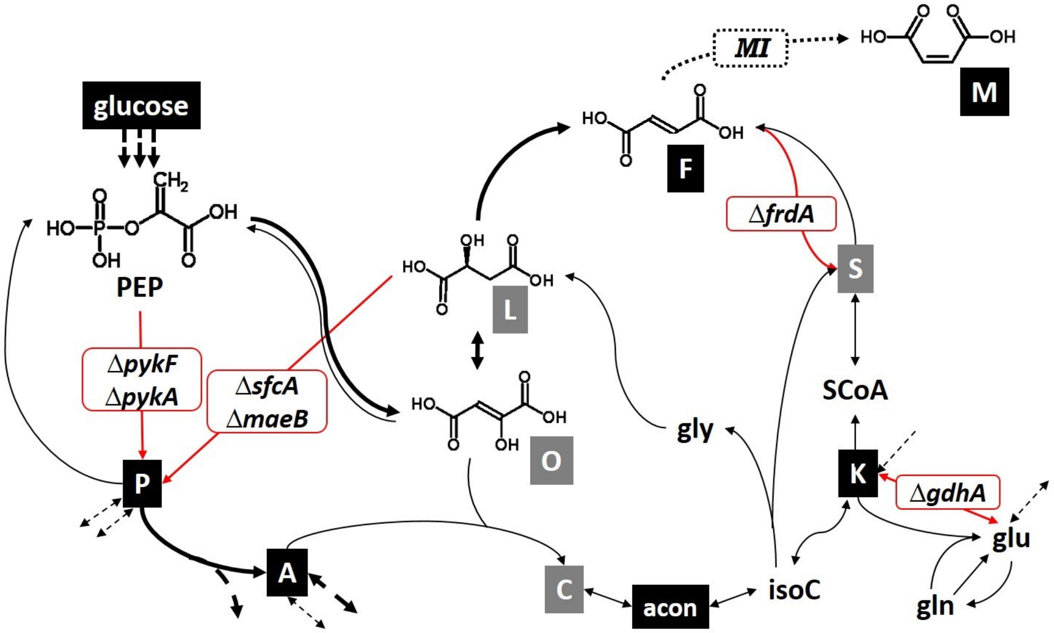


Figure 9: Model design and predicted flux for the 6KO strain.

The 6 gene KOs predicted by the minimal cut set algorithm are shown in the metabolic flux diagram (based on the metabolic model map provided by Lu and Mahadevan, 2015 correspondence). The KOs are pykF, pykA, sfcA, maeB, frdA, and gdhA shown in red. The flux is hypothesized to travel from glucose through the bold arrows to result in an increased production of fumaric acid. The dotted arrow represents the predicted conversion of fumaric acid to maleic acid by MI. Dashed arrows represent other reactions entering or exiting central metabolism. Metabolites that were investigated through HPLC or LC-MS are abbreviated as: P: pyruvic acid, A: acetic acid/acetyl CoA, O: oxaloacetic acid, L: L-malic acid, F: fumaric acid, M: maleic acid, S: succinic acid, K: α-ketoglutaric acid, acon: aconitic acid, and C: citric acid, and are shown in either black or grey boxes signifying those intermediates affected by the 6KOs in vivo or those where no change was observed, respectively. Other labelled metabolites that were not investigated are: PEP: phosphoenolpyruvate, isoC: isocitrate, SCoA: succinyl-CoA, glu: glutamic acid, gln: glutamine, and gly: glyoxylic acid.

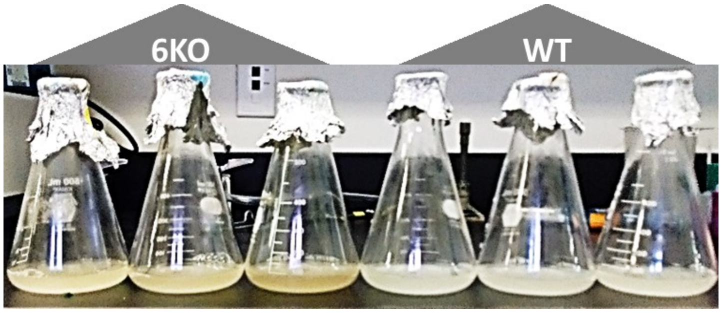


Figure 10: Cultures of 6KO and WT.

When grown in minimal M9 media, the 6KO culture turns beige (left flasks) while the WT culture is off-white (right flasks).

Maleic acid production in vitro *and* in vivo

We originally planned the expression of MIs to be from inducible overexpression vector (pET41b) for *in vitro* testing, and as chromosomally integrated in the engineered background strain for *in vivo* testing. However, chromosomal integration would remove the possibility of controlling expression with an inducible plasmid and would require multiple MI insertions for multiple variant testing. We decided to express MIs *in vivo* from a different inducible vector (pTRC99a) as MG1655 does not express the T7 polymerase required for pET vectors. The ORF sequences were initially codon-optimized to exclude the restriction cut sites necessary for the synthesis and pET cloning steps (*Aar*I, *Asc*I, *Bsa*I, *Kas*I, *Nco*I, *Not*I, *Sap*I, and *Xho*I). Nineteen MI variants were excluded from *in vivo* testing because they contained an *Eco*RI cut site within the ORF, which was necessary for pTRC cloning. An additional 12 MI variants were not pursued once colony-PCR or sequencing of selected colonies revealed empty backbone pTRC99a plasmid, or that the ORF amplicon contained deleterious mutations. A total of 24 of the original 35 MIs tested *in vitro* were cloned into the pTRC99a background, which was sufficient for our purpose of screening a set of putative MIs.

In vivo testing in the 6KO background *E. coli* strain was carried out with 5 MIs (MI-Rso, MI-Cba1, MI-Cba2, MI-W.D, and MI-Cca) confirmed to have reverse activity and which represented the most phylogenetic diversity of the library. MI-Pfl was included in the *in vivo* studies even though it had not previously been tested for reverse activity *in vitro*. This inclusion was logical because MI-Pfl was used as the control in all of the overexpression assays and it was used to select the appropriate activity buffer for all *in vitro* experiments [19]. Maleic acid production *in vivo* was measured below 1µM by LC-MS using culture supernatants of the 6KO strain expressing various MIs. The maximum maleic acid measured was 0.901 µM (105µg/l, molecular weight of maleic acid: 116.07g/mol) when MI-Cba2 was expressed in the 6KO strain in shake flasks (Figure 8 and Appendix Figure 1).

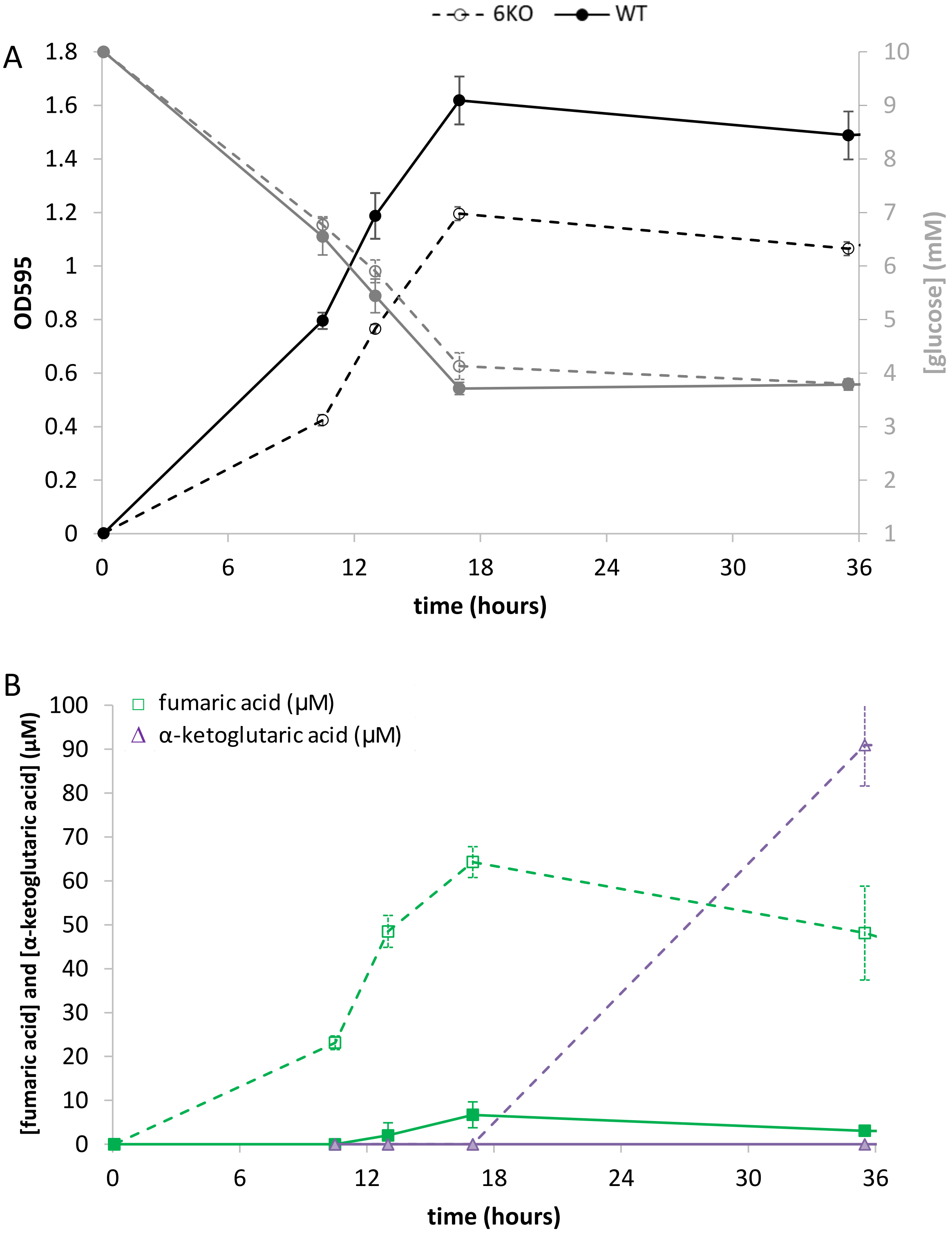
E. coli *metabolite analysis*

In addition to monitoring the growth of the 6KO strain, select metabolite profiles were investigated for the 6KO and WT strains without (Figure 11) and with MIs expressed (Figure 12). Compared to the WT strain, the 6KO strain exhibited some lag in growth, however the glucose uptake rate was comparable (Figure 11A). The 6KO produced more fumaric acid as desired, at a rate of about 10-fold that of WT (Figure 11B). There was also a notable and unexpected increase in α-ketoglutaric acid. In line with the model predictions, 6KO production of aconitic acid, acetic acid, and pyruvic acid decreased (Figure 11C). Citric acid, succinic acid, shikimic acid, and ethanol levels were investigated however no discernable differences were observed between WT and 6KO. L-malic acid was not detectable with the HPLC method employed. It should be noted that oxaloacetic acid levels could not be accurately measured due to its spontaneous decomposition into pyruvate and CO_2 [43], which would therefore affect the quantification of pyruvic acid.

The expression of reversible MIs did not appear to alter the levels of the key metabolites affected in the 6KO, but did result in the production of maleic acid (Figure 12). In an effort to accurately represent the triplicates, samples where there was no peak detected for maleic acid were included with a value of “0”. Because of this inclusion, high standard deviations for maleic acid production were calculated as represented by the error bars. In the 96-well *in vivo* experiment, the MI-Cca variant appears to be the best maleic acid producer, producing 553nM maleic acid (64.2µg/l). Compared to 6KO with other MIs or no MI, the MI-Cca expression resulted in less fumaric acid, more pyruvic acid, comparable acetic acid, less aconitic acid (Figure 12C-F), and possibly less α-ketoglutaric acid (Appendix Figure 2). A preliminary MI *in vivo* expression test in shake flasks suggested that the MI-Cba2 variant was possibly the optimal reversible MI (Appendix Figure 1) with 901nM maleic acid (104µg/l) which was the basis for its selection for expression in WT. However, that experiment included IPTG induction of the pTRC-MI vectors which appeared to have negatively affected maleic acid production.

Unexpectedly, the presence of a MI in WT results in maleic acid production as well, at a potentially higher amount than any of the other 6KO-MI combinations tested (Figure 12B). The MI appears to have affected the overall fitness of the WT as seen in the growth lag (Figure 12A).

The other metabolite levels measured corresponded with the previous 6KO vs. WT observations with the exception of aconitic acid which was lower in WT with the MI but was previously higher in WT with no MI compared to 6KO.



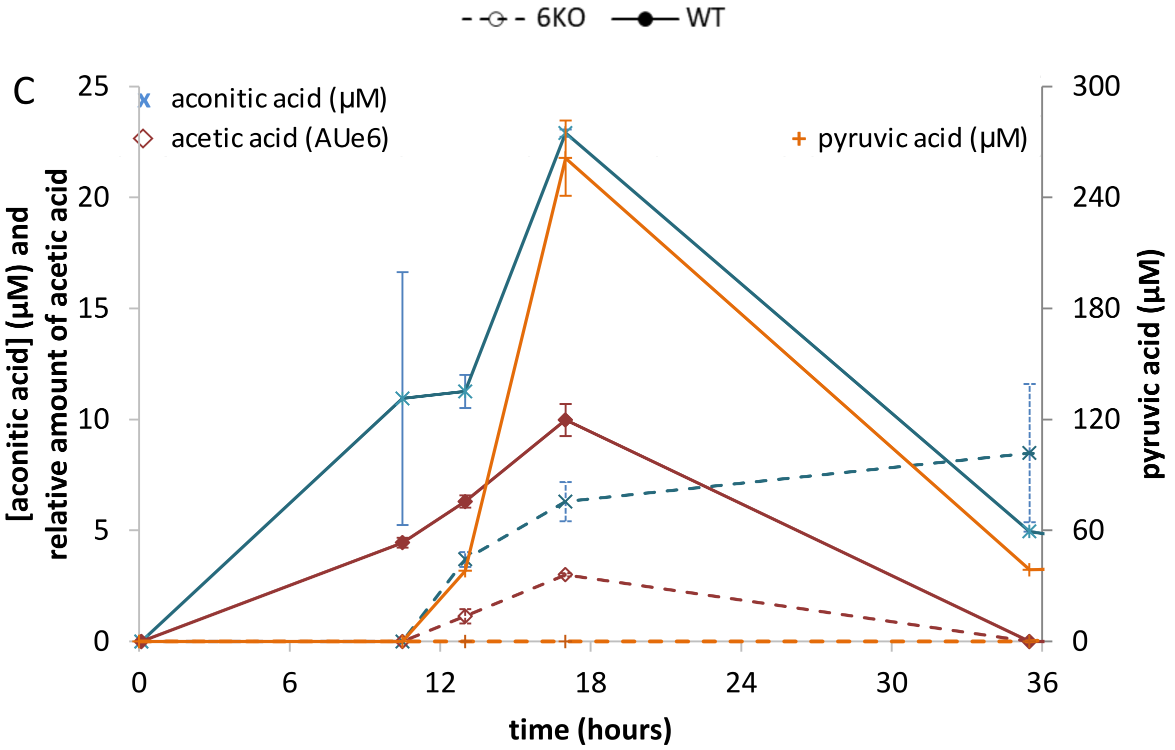
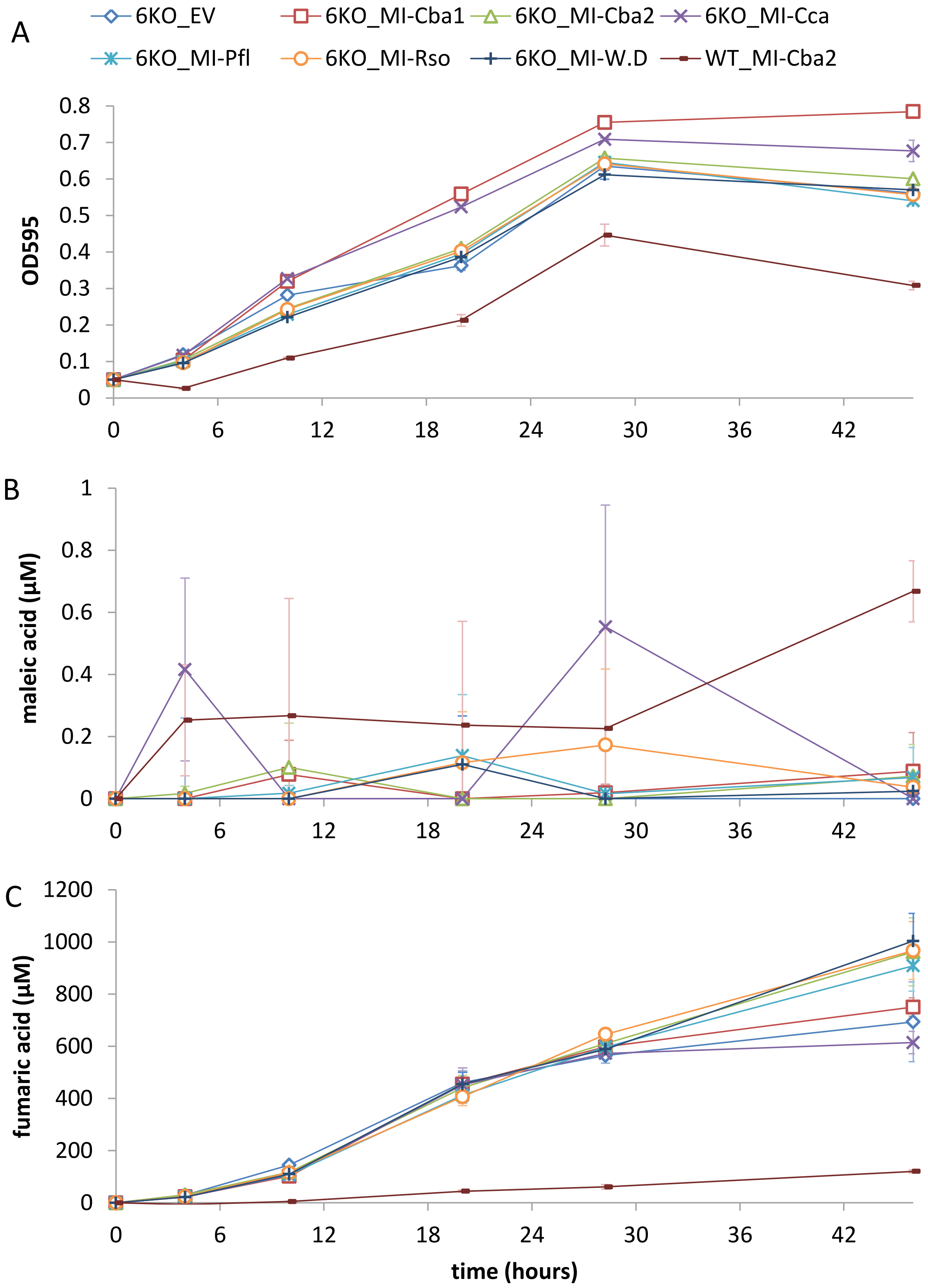


Figure 11: Growth and metabolite analysis of strains without MIs expressed.

Select metabolites surrounding the TCA cycle were measured from culture supernatant in a time course assay for the 6KO and WT strains (dashed and solid lines, respectively). Strains were inoculated at OD_{595} = 0.05 in triplicate into 50mL minimal M9 media in shake flasks, and then incubated at 37°C, 200rpm. (A) Growth was measured by spectrophotometry, glucose concentrations were measured by HPLC-RI, and HPLC-UV was used to quantitate the metabolites shown in B: fumaric acid (green squares) and α-ketoglutaric acid (purple triangles), and in C: aconitic acid (blue Xs), acetic acid (maroon squares), and pyruvic acid (orange pluses). Error bars indicate the standard deviation between replicates.


A
6KO_EV
6KO_MI-Cba1
6KO_MI-Cba2
6KO_MI-Cca
6KO_MI-Pfl
6KO_MI-Rso
6KO_MI-W.D
WT_MI-Cba2
OD595
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
6
12
18
24
30
36
42
B
maleic acid (μM)
1
0.8
0.6
0.4
0.2
0
0
6
12
18
24
30
36
42
C
fumaric acid (μM)
1200
1000
800
600
400
200
0
0
6
12
18
24
30
36
42
time (hours)

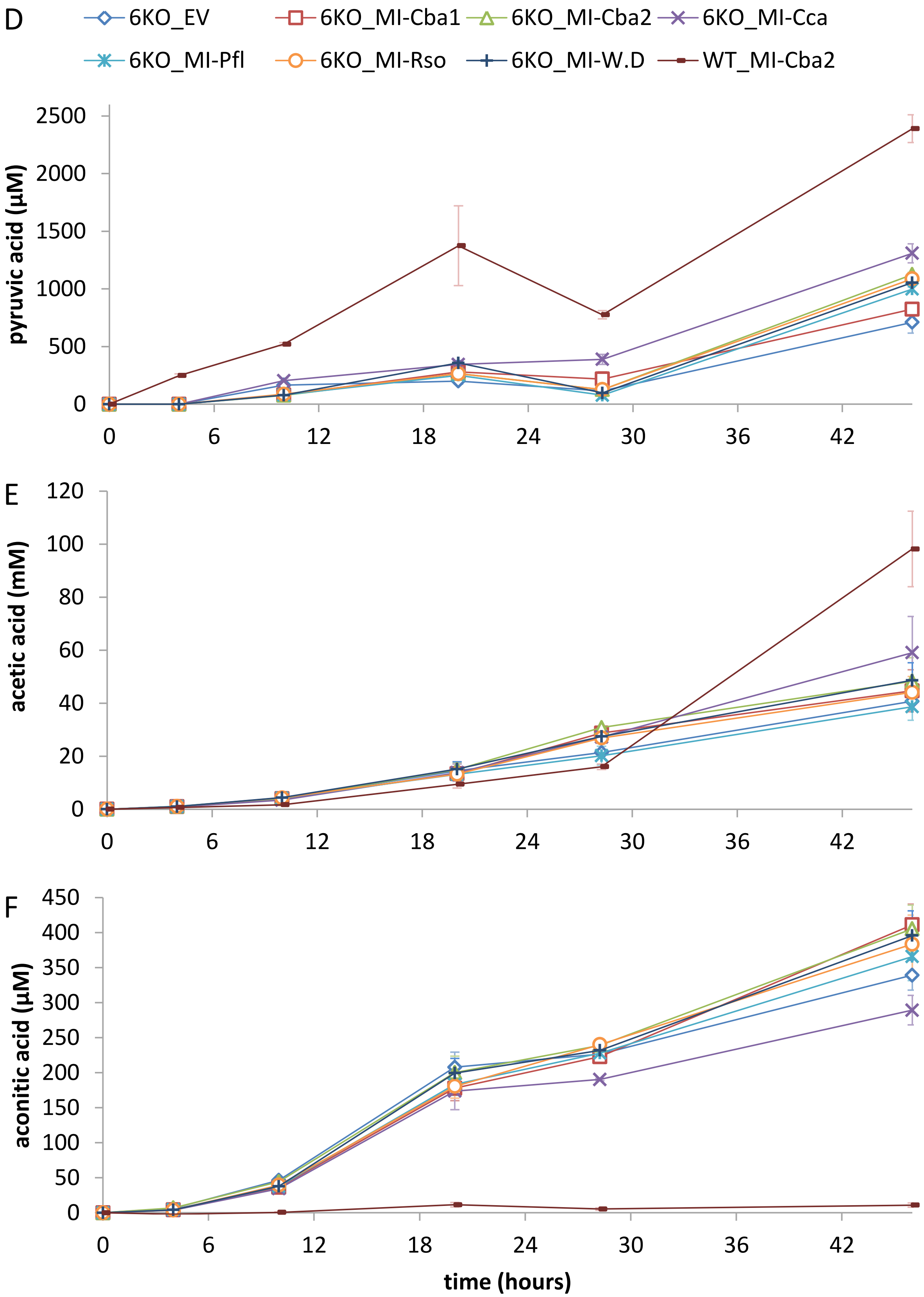

D
6KO_EV
6KO_MI-Cba1
6KO_MI-Cba2
6KO_MI-Cca
6KO_MI-Pfl
6KO_MI-Rso
6KO_MI-W.D
WT_MI-Cba2
pyruvic acid (µM)
2500
2000
1500
1000
500
0
0
6
12
18
24
30
36
42
E
acetic acid (mM)
120
100
80
60
40
20
0
0
6
12
18
24
30
36
42
F
aconitic acid (µM)
450
400
350
300
250
200
150
100
50
0
0
6
12
18
24
30
36
42
time (hours)

Figure 12: Growth and metabolite analysis of strains with MIs expressed.

Select metabolites surrounding the TCA cycle were measured from culture supernatant in a time course assay for the 6KO and WT strains with or without MI variants expressed from the pTRCMI- vectors. Strains were inoculated at OD_{595} = 0.05 in triplicate into 1mL minimal M9 media with ampicillin in 2mL deep-well 96-well plates, and then incubated at 37°C, 90% humidity, 300rpm. Growth (A) was measured by spectrophotometry, maleic acid concentrations (B) were measured by LC-MS QToF, and HPLC-UV was used to quantitate the remaining metabolites: fumaric acid, pyruvic acid, acetic acid, and aconitic acid (C-F). Error bars indicate the standard deviation between replicates. Samples were taken from cultures of 6KO with: empty pTRC99a vector (6KO_EV, blue diamonds); pTRCMI-Cba1 (red squares); pTRCMI-Cba2 (green triangles); pTRCMI-Cca (purple Xs); pTRCMI-Pfl (cyan asterisks); pTRCMI-Rso (orange circles); pTRCMI-W.D (navy pluses), and of WT with pTRCMI-Cba2 (maroon dashes).

DISCUSSION

The production of maleic acid using a sustainable platform would help alleviate the dependence on petroleum-based derivations for a number of products which can be produced using maleic acid as a starting point. This study showed that the production of maleic acid in a synthetically engineered *E. coli* strain is possible. This was achieved by first screening the expression of a heterologous MI for production of maleic acid from fumaric acid, which is the enzyme's reverse, unfavourable reaction. Promising candidate MIs were expressed in an *E. coli* strain metabolically engineered to produce the maleic acid precursor, fumaric acid. The strategy involved a variety of techniques including a bioinformatics mining approach, the expression and activity assessment of candidate enzymes, the design and construction of the host production strain, and the metabolic profiling of the resulting strains. The subsequent variation in results and any observed trends are discussed herein. As well, modifications for future maleic acid-producing strains are discussed.

The presence of a nic2 *cluster shows a potential trend with activity*

The MIs required for maleic acid production have not been extensively characterized in the past. This prompted the generation of a library of putative MIs for activity screening. The MI library selection included an analysis of the genomic environment in which putative MIs were found. This was due to a *nic*2 cluster described for MI-PpuS containing 4 genes in the pyrrolidine (nicotine degradation) pathway [14]. The *nic*2 cluster was later more thoroughly described as being comprised of 13 genes [44], although we chose to focus on the 4 original members of the *nic*2 cluster which are the most proximal to the MI in *P. putida* S16. Other bacterial species which contained the *nic*2 cluster were not known to have nicotine degradation abilities [14] (also inferred from MI activities observed in this work).

Evidence suggests that the *nic*2 cluster is most likely contained in the region surrounding 4 putative MIs from the 55-MI library (Appendix Table 4). Many of the regions that were investigated are vaguely annotated possibly due to the fact that the % identity between orthologs is often low [45]. The vague or missing annotations resulted in an inability to classify the genetic environments surrounding the MIs as certainly not containing the *nic*2 cluster. In addition to the % identity of hits dropping off quickly, sometimes the only hits using basic local alignment searches were ORFs similarly declared as “hypothetical proteins”. It is apparent however that as the number of potential *nic*2 cluster members increases in the ±4-ORF space, the likelihood of the putative MI having any observed activity increased (Figure 3).

More specifically, trends were observed for the presence of specific types of ORFs in the ±4-ORF space (Appendix Figure 5). The putative presence of three *nic*2 members (HSP, NFO, and HPO, Appendix Figure 5A-C) corresponded to higher fluorescence levels with MI-GFP in *S. cerevisiae* cells, higher forward kinetic values, and to the presence of reversible activity. Furthermore, the number of transcription factors and transporter protein coding sequences found in the ±4-ORF space also influenced these parameters. Fewer of both transcription factors and transport proteins appeared to correlate to more active MI variants (Appendix Figure 5D-E). This could be due to the fact that if more ORFs surrounding the MI are not specifically a part of the *nic*2 cluster, the cluster becomes more spread out and is less likely to be subject to horizontal gene transfer which is suggested to be selected for by “selfish operons” [46]. Together, these results show that a survey of the genomic environment of a protein may provide clues about its predicted activity.

Expression of maleate isomerases

In order to screen the activity of the MI library, MIs were overexpressed from IPTG-inducible plasmids in an *E. coli* expression strain. Forty-five MI variants exhibited soluble expression with the particular conditions tested. The expression of 2 MIs is unknown and 8 were confirmed to be insoluble. The insoluble variants were spread throughout the phylogenetic tree (Figure 6). MI-Afa was found to have insoluble expression in this study although it has been previously purified directly from its native host, *A. faecalis* [17]. There are no published reports of attempts to express and purify the other 7 MIs that were insoluble in this study. Four of these were isolated from water bodies (MI-Pha, MI-Cbae, MI-S.S, and MI-meB) and three were from soil samples (MI-Ach, MI-Rop and MI-Car). *P. haloterans* B2 (MI-Pha) is a marine halotolerant species [47], *R. opacus* PD630 (MI-Rop) is a chemoheterotroph [48], and *C. baekdonensis* B30 (MI-Cbae) is found in the Arctic ocean (unpublished, NCBI biosample accession: SAMN02471168). The environments in which these species are generally found may represent specialized conditions required for optimally active MIs. For example, it has been documented that there are trypsin homologs which can be either warm- or cold-active, a feature which is determined by the environment in which the enzyme will be needed [49]. The MI from *S. solfataricus* P2 (MI-Sso), which is a sulfur-metabolizing thermophile [50], displayed soluble expression however, the expression level was low and neither the crude nor purified extracts exhibited activity. This further supports the hypothesis that a specialized environment may be required for optimal activity of these MI variants. However, there are MIs that were soluble and active which also originated from species from distinct environments as well, namely, MI-Cba1 and MI-Cba2 (which were active in both the forward and reverse directions) come from *C. basilensis* isolated from heavy-metal contaminated ground water [23], and MI-W.D (which showed reverse activity only) from *W. sp.* D3 which was found in the Antarctic [51].

MIs are typically native to soil-dwelling bacteria as a part of the nicotinate/nicotinamide degradation pathway [15]. Native *P. putida* S16 gene expression has been tested in response to growth on nicotine as a sole carbon source, however the MI in particular was not included in that analysis [44]. The changes observed in stress-response pathways may have overshadowed any changes in MI expression, which may be low to begin with. In other words, it can be hypothesized that because the MI is responsible for the last degradation step before the TCA cycle is reached, MI expression may not be affected as much as the initial metabolic steps in nicotine degradation. It would be interesting to investigate native MI expression levels in comparison to MI expression in our *in vivo* assays. Considering native MI expression is a part of a specialized pathway, the species in which MIs are found may require an environmental cue for expression, such as the presence of nicotine for degradation. Furthering the idea that specific conditions are required for expression, and ultimately activity, is the fact that there were 12 MIs which were indeed soluble but did not show any forward activity, although this may have been due to possible interference by the His-tag. It is also possible, although no other concrete examples could be found, that the inactive putative MIs may have no longer been required in certain organisms and the lack of selection for MI activity may have resulted in an accumulation of deleterious mutations. However, the inactive putative MIs may simply serve another purpose altogether.

Activity of maleate isomerases

MI variants which were more active tended to be located in clusters I and II of the phylogenetic tree and were completely absent from cluster VII (Figure 8). A wide range of kinetic values were measured with the best and worst performers appearing throughout the tree. The class and phylum from which each MI originated (Appendix Table 5) showed some correlation to activity but not more than what was apparent in the phylogenetic tree. In the search for other trends regarding the origin species of the MIs and their activities, we investigated the isolation location of the sample and the type of species (whether symbiont, or pathogenic) the MI was isolated from. In this regard, we could only conclude that those from symbiont organisms had relatively low kinetics values overall (Appendix Figure 5F) and those from contaminated sites had lower k_{cat}/K_M values (ranging from $8.7x10^2s^{-1},M^{-1}$ to $1.6x10^4s^{-1},M^{-1}$) compared to the entire set. These were not very informative relationships considering the variation in activity that was observed.

Concerning those MIs with kinetic values published, our assay results agreed in terms of magnitude for MI-Nfa, and MI-Pfl but not for MI-Rjo (Table 3). For MI-Nfa, the V_{max} values disagree because of differing amounts of total enzyme used (E_t). Four buffer types (20mM potassium phosphate, 20mM sodium phosphate, 50mM HEPES, and 50mM Tris-HCl) at varying pH (between 6.4 and 9.5) were used to assess optimal buffer conditions with the V_o of MI-Pfl as an indicator. Variation of MI-Pfl activity between buffer types was about 10-fold (data not shown) meaning that the discrepancy between the kinetics values measured in this study and those values that have been previously published is possibly due to different buffer conditions. Despite these discrepancies, the values reported here are consistent based on the conditions used. The reactions were performed in triplicate and V_o calculations included at least 12 substrate concentrations ranging 3 orders of magnitude to include concentrations below predicted K_M and above saturating.

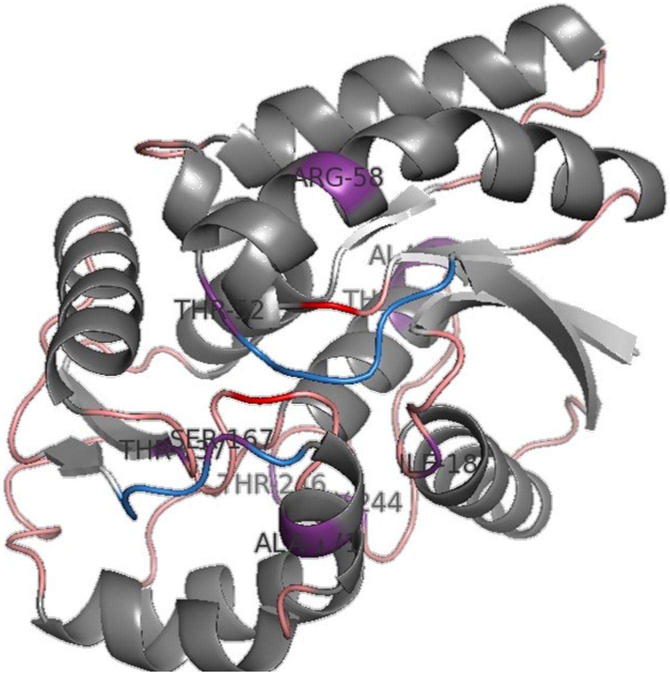
Table 3: Published kinetic values compared to those obtained in this study.

MI	E_t	V_{max}	K_M	k_{cat}	k_{cat}/K_M	Assay	Ref.
	(μg)	(μM/min)	(μM)	(s^{-1})	(s^{-1}, M^{-1})	conditions	
MI-Nfa	.0028	0.08	4.6 ± 0.5	2.8	$6.1x10^5$	50mM HEPES, pH 7.5	[30]
	1	115 ± 12.5	16.7 ± 20	10.6	$6.4x10^5$	20mM KPhos, pH 7.6	this study
MI-Pfl	NA	NA	300	30	$1.0 x10^5$	50mM Tris-HCl, pH8.4, 5mM α-thioglycerol	[24]
	10	304 ± 15	601 ± 67	2.9	$4.8 x10^3$	50mM Tris-HCl pH 8.0	this study
	10	143 ± 16	293 ± 97	1.4	$4.6 x10^3$	20mM KPhos, pH 7.6	this study
MI-Rjo	10	1973	149.6 ± 9.2	96.7	$6.5x10^5$	20mM KPhos, pH 8.0, 5mM β-mercaptoethanol	[29]
	10	216 ± 222	2364 ± 2828	2.1	$8.7 x10^2$	*50mM Tris-HCl pH 8.0	this study

*Standard errors for values from this study were obtained using the Michaelis-Menten non-linear regression fit curves calculated by GraphPad. Assay conditions in this study all included 5mM β-mercaptoethanol. *Activity for MI-Rjo was measured prior to the buffer optimization test to match the K_M value of MI-Pfl (293μM) to literature (300μM).*

For the forward maleic acid → fumaric acid reaction, the best performers in terms of K_M and k_{cat}/K_M are MI-P.B and MI-Nfa, however the similarities between these two enzymes does not go further than their high activities and the fact that neither was isolated from a contaminated site. *N. farcinica* IFM is infectious and was isolated from a mammal in Japan [52], while *P. sp.* BAY1663 in not infectious and was isolated from soil in Hungary (unpublished, NCBI BioProject accession PRJNA232351). There is nothing remarkably similar throughout their sequences and they are not close on the phylogenetic tree. MI-P.B has 4 unique residues that are not present throughout the rest of the MIs for which activity was tested, while MI-Nfa has none (Appendix Table 6). The reverse reaction was only tested for MI-P.B *in vitro* but despite its performance in the forward direction, it converted only about half the amount (about 1%) that the best performers did (almost 2%).

Many of the enzymes showing reverse activity were indeed active in the forward direction as well. Although there was no observable correlation between the degrees of activity in the two directions, this validates that the initial screen for forward activity was a good indicator of possible functional reverse activity. The exception of 3 MIs (MI-W.D, MI-H.C, and MI-Pas) out of the 22 screened for reverse activity did not have any detectable forward activity but also were not amongst the top maleic acid-producing MIs. The reverse reaction was best performed *in vitro* by MI-Cba1 (1.99%), MI-Sma (1.95%), MI-Rso (1.87%), MI-Cba2 (1.73%), and MI-Aph (1.53%) (Figure 8). Except for the two from *C. basilensis*, these MIs are spread throughout the phylogenetic tree. *C. basilensis* is non-pathogenic and is from contaminated North American groundwater [23], the other 3 are from soil [25, 53, 54]. *A. phenanthrenivorans* Sphe3 is non-pathogenic and was isolated from contaminated soil in Greece [53], while both *S. marcesens* and *R. solanacearum* FQY_4 are pathogenic and were isolated from soil in Japan [25] and China [54] respectively. In other words, 2 of the 4 origin species yielding reversely active MIs are from contaminated sites. Intriguingly, MI-Aph has 9 unique residues, MI-Rso has 2, and MI-Cba1 has 1. *In vivo* assays suggest that MI-Cca and MI-Cba2 produced the absolute most maleic acid (553nM, Figure 12B and 901nM, Appendix Figure 1B, respectively). It is noteworthy that two of the best performing MIs in the reverse direction come from a single species isolated from a contaminated site. Perhaps the presence of gene duplication events allowed the organism to adapt paralogs to environmental stresses while maintaining the still necessary original function of the enzyme, a phenomenon that has been suggested with mannose-binding proteins in the bacteria, *Thermotoga maritima* [55]. Together these results indicate that MI-Cba2 may be the best candidate for future enzyme and maleic acid-production strain modifications. The sequence of MI-Cba2 has 10 residue replacements with different types of residues compared to MI-PpuS (Figure 13) which could account for its greater reverse activity.



MI-PpuS residue	MI-Cba2 replace-ment
I18	T
T52	V
A58	R
T137	A
S167	P
A171	E
T233	K
A234	Q
G244	A
T246	A

Figure 13: Amino acid replacement in MI-Cba2

A monomer of the MI_PpuS structure (PDB: 4FQ7 [13]) is shown with 10 residues (in violet) which are replaced with different types of residues in MI-Cba2. Structure colors are as described in Figure 5. In the list, hydrophobic residues (I, V, A) are filled in green, uncharged amino acids (T, S, Q) are filled in purple, blue fill denotes charged amino acids, with red or black text for positively (R, K) or negatively (E) charged respectively, and special cases of glycine and proline are filled in yellow.

Conserved sequences and activity trends

Chen *et al*. [13] determined the crystal structure of MI-PpuS (PDB: 4FQ7) which revealed that the substrate maleic acid is recognized by asparagine residues 17 and 169 which form H-bonds with one carboxylate group of maleic acid, while tyrosine 139 forms an H-bond to the other. Proline 14 and valine 84 form Van der Waals interactions with the C2 and C3 of maleic acid. Furthermore, valine 51 controls substrate entry into the active pocket of the enzyme. All six of these important residues plus two catalytic cysteines were shown to be conserved throughout 4 MIs: MI-PpuS, MI-Nfa, MI-Sma, and MI-Bst [13] and are actually seen to be conserved throughout the library of putative MIs studied here with the exception of those from marine metagenomes. We found that an additional 8 residues are conserved throughout the entire library: E21, R45, L56, M59, P138, S198, L232, and G245. These residues are positioned largely around the active pocket according to the published MI-PpuS structure. Residues close to the active site could be involved in the catalytic mechanism which would be maintained over an evolutionary time period. The conserved charged residues (E21 and R45) may aid in the positioning of the substrate or in the formation of the product. They may also have a role in the enzyme-substrate conformational changes which may assure substrate specificity. The hydrophobic residues (L56, M59, and L232) may help maintain appropriate folding of the enzyme if their side chains are positioned inwardly. The conserved proline (P138) may also have a role in the structure since the amine group can disrupt secondary structures. The fact that S198 is essentially next to the catalytic C200 may provide a clue as to why it has been conserved. Interestingly, position 199 is occupied by small residues: predominantly by alanine (in 33 of 55 cases), by another cysteine (17 cases), by valine in one case, or as a gap space for the 4 MI variants in cluster VII of the phylogenetic tree (Figure 4). These inferences are based on the residue characteristics, more information concerning these 8 newly found conserved residues could be established with a more in depth structural analysis.

MI-PpuS was also found to exhibit a breathing motion of two loop domains (β2-α2 and β6-α7) which could be a mechanism for the enzyme to control the specificity of the reactions it performs [13]. It is currently unknown what the maleic acid to fumaric acid isomerization intermediate is [18], but it is proposed that the engulfment of the substrate can prevent reactions between the intermediate and H_2O present around the MI [13]. Serine 167 is within the β6-α7 loop and a MI-PpuS S167W mutant lost almost half of its specific activity due to steric hindrance of tryptophan in the loop obstructing the entrance of maleic acid to the active pocket [13]. Interestingly, MI-Cba2, one of the most active MIs in both the forward and reverse directions, has a proline in the S167 position. Proline can be considered a smaller amino acid [56], but depending on its positioning, it may be considered obstructive. This may contradict the hypothesis presented by Chen *et al.* that larger residues would block the substrate from entering the MI. Thirteen MIs include a proline in position 167 of the breathing loops, 5 of which definitely exhibited activity in both directions (MI-Cba1, MI-Cba2, MI-Pfl, MI-Rso, and MI-Sma). The 4 MIs investigated by Chen *et al*. showed variation in the breathing loop sequences with 6 of the 11 total residues being conserved [13]. Excluding MIs in cluster VII of the phylogenetic tree, the breathing loop consensus sequences for the library of MIs are:

β2-α2 (47-52): [RGQ] M [KHQMAS] [ζA] V [ζV] and
β6-α7 (164-168): L [EG] [VI] [πζ] [DNG]

The peptide motif nomenclature [56] is used: residues with no notation are conserved, residues in square brackets represent the choices for that position, ζ represents hydrophilic residues: N Q S T E D K R H, and π represents small chain residues: P G A S. The breathing loop sequence alignments for all 55 MIs are shown in Appendix Table 7.

If we further reduce the set to only those with observed activity (in either direction), the breathing loop consensus sequences are more refined:

β2-α2 (47-52): [RGQ] M [KHQM] ζ V [ζV] and
β6-α7 (164-168): L [EG] [VI] [πQ] D

There appears to be an increased importance in the presence of hydrophilic residues in the β2-α2 loop, while the β6-α7 loop shows a new conservation of aspartic acid and an increased preference for small residues at position 167. In fact, Q167 only appears in MI-P.U (from *P. sp.* UMTAT08) which was only found to have unquantifiable activity in the forward direction.

An effort was made to find other conserved regions according to the activity trends of the MIs. Although there were some other residues that commonly appeared in active or inactive groups of MIs, it was difficult to discern any absolute trends using the available information. Based on the MI activity trends and observed sequence conservation, one possible way to improve reverse MI activity might be to introduce or remove portions of the loop(s) which may allow the enzyme to more specifically accept fumaric acid as a substrate to be converted into maleic acid. Efforts to optimize superior maleic acid production may also involve a directed evolution strategy of MI, either focusing on the loops or on the enzyme as a whole in order to improve the enzyme's ability to convert fumaric acid. Based on successes of enzyme catalysis parameter (e.g. V_{max} or K_M) enhancements using directed evolution [57-60], this could be a promising strategy for increasing maleic acid production with MIs.

Synthetic model design

Initially we intended on building a maleic acid-producing *S. cerevisiae* strain because it is a robust host suitable for industrial-scale productions [28] and has been previously utilized in the production of industrially useful carboxylic acids such as succinic acid [61], pyruvic acid [62], and *cis-cis* muconic acid [63]. Because of the multiple cellular compartments in *S. cerevisiae*, the inclusion of additional transport molecules may have been necessary to ensure an efficient metabolic flux was met. This approach was not pursued because of foreseen complications with transport and because our collaborators could not design a reliable *in silico* model (early pathway ideas for a maleic acid-producing *S. cerevisiae* model are described in Appendix II and Appendix Figure 3).

Another industrially relevant host organism is *E. coli*. The metabolic models used for engineering *E. coli* are generally accepted as necessary and reliable predictors of carbon flux [10, 64]. The minimal cut set metabolic model used to design the 6KO *E. coli* strain (communications with Lu and Mahadevan, 2015) was geared toward producing an appropriate amount of substrate for the MI while not substantially compromising the fitness of the strain. The resulting 6KO strain did acquire a reduction in fitness as seen in the lag in growth (Figure 11, Appendix Figure 4). The knocked out genes are generally involved in the TCA cycle; the *frd*A gene is contained in the TCA cycle, the *gdh*A, *sfc*A, and *mae*B genes act on metabolites within the TCA cycle, and *pyk*F and *pyk*A act on pyruvate/phosphoenolpyruvate which are each one reaction away from the TCA cycle. Considering that the TCA cycle and its surrounding metabolites are necessary for several cellular functions [43], reduced fitness is expected to a degree when genes with a role in central metabolism are tinkered with [65]. The growth of the 6KO in rich media is not greatly affected as it still manages to reach the maximum OD levels that the WT strain attains in only about 1 extra hour (Appendix Figure 4A). However, in minimal media reduced fitness of the 6KO strain was more apparent as it struggled to reach the WT's maximum OD level (Appendix Figure 4B). In an attempt to learn what was affected by the six KOs, we followed the levels of some key metabolites in culture supernatants over time.

Metabolite analysis

Metabolites in and around the TCA cycle were investigated for the 6KO and WT strains however not all metabolites provided useful profiles. Oxaloacetic acid is known to decompose spontaneously into pyruvic acid and CO_2 [43] meaning that it could not be quantified reliably. Also, the calculated amount of pyruvic acid is therefore inflated to an unknown degree due to oxaloacetic acid decomposition. Nevertheless, the areas of peaks corresponding to pyruvic acid were different between strains. Citric and isocitric acids coeluted with non-degraded oxaloacetic acid in the standards and therefore could not be distinguished from one another. However, the peak corresponding to these metabolites did not exhibit any differences between the experiments and controls. Also, the metabolic profiles of glyoxylic, lactic, shikimic (as a potential indicator a reason for the phenotypic color change of cultures), and succinic acids were investigated but no change was noted in the experiments.

The decreased fitness of the 6KO is hypothesized to be caused by the increased levels of fumaric acid, which the cell must now cope with, and also due to the observed increase in α-ketoglutaric acid and decreases in pyruvic acid, aconitic acid, and acetic acid. The intracellular pH of *E. coli* is maintained around 7.5, which is higher than the pK_a of organic acids meaning that an increase in organic acid production can disrupt the cytosolic pH and thus reduce the cell's fitness [66]. A study which developed a model for carboxylic acid toxicity on bacterial growth found that smaller carboxylic acids have a more severe effect on cell fitness [67]. In this study, fumaric acid was measured up to 1mM in 96-well plate assays and 160μM in shake-flasks. An *E. coli* strain (ABCDIA-RAC) producing about 11mM fumaric acid in shake flasks has been described [68] which also suffered a decrease in fitness but to a lesser degree than the 6KO strain. That strategy included knockout of the fumarate reductase and fumarase genes and the redirection of carbon with the induction of the urea cycle. Similar to the strain engineered in this study, there was a decrease in pyruvic acid (normalized to culture densities). The ABCDIA-RAC strain produced about 10% less acetate compared to WT JM109(DE3) while the

6KO strain produced only about a third of the acetate produced by WT MG1655. Another study increased fumaric acid production from 1.45g/l (12mM) in shake flasks to 28.2g/l in fed-batch cultures [69] and more recently fumaric acid production has reached up to 41.5g/l in glycerol fed-batch cultures [70]. Based on these comparisons, the increase in fumaric acid alone cannot account for the decreased 6KO fitness since the levels were not as high as those demonstrated elsewhere.

Interestingly, α-ketoglutaric acid is an inhibitor of at least 7 *E. coli* enzymes, 2 of which, citrate synthase (*glt*A) and phosphoglycerate dehydrogenase (*ser*A), catalyze essential reactions [11]. No growth was observed for *glt*A or *ser*A knockout strains on minimal media [71]. An abundance of α-ketoglutaric acid was not predicted by the modelling algorithms and thus could reasonably be a source of the 6KO's reduced fitness. The inhibition of the *glt*A and *ser*A reactions is not severe enough to completely eliminate viability, but could be at a rate that the cell's fitness is reduced.

There is little indication that the remaining affected metabolites contribute greatly to the strain's fitness. Acetic acid is necessary for the production of acetyl-CoA and is produced in the cell by a number of different reactions performed by lyases, oxidases, and deacetylases [11]. Pyruvic acid is one of the substrates used by pyruvate oxidase to create acetic acid, therefore a decrease in pyruvic acid should result in a decrease in acetic acid as was seen in the 6KO strain. However, because acetic acid is produced through numerous reactions, it would seem that the cell should be able to adapt to a decrease in a single reaction with little difficulty. Pyruvic acid is also produced by many cellular reactions however, the 6KO strain included 4 deletions which are known to produce pyruvic acid: *pyk*F, *pyk*A, *sfc*A (also known as *mae*A), and *mae*B [11]. This could account for a significant proportion of pyruvate production which would mean that the cell is under greater stress to keep up with pyruvate demands. Finally, no evidence could be found that a decrease in aconitic acid would result in reduced fitness although the decrease may correlate with a decrease in citrate/isocitrate which could lead to more drastic effects.

Realistically, the 6KO's reduced fitness is likely to be caused by the negative epistasis of the genes knocked out. This shows that metabolic modelling algorithms cannot necessarily predict the outcomes of metabolic engineering. This inference was also stated in other studies for both *S. cerevisiae* [72] and *E. coli* [73] metabolic modelling. Actual *in vivo* tests can be used to 'teach' the algorithms to make better predictions thus providing better production strains for industrially important molecules.

Maleic acid production and improvement strategies

Regardless of the fitness of the 6KO strain, the heterologous expression of MIs effectively resulted in the production of maleic acid. The maleic acid levels produced here (less than 1µM) will require substantial improvement before this strategy is ready for industrial-scale productions. It should be noted that only maleic acid monomers were quantified from culture supernatant. Because polymer species were observed in the LC-MS spectra, the actual levels of maleic acid are higher than those reported here. The method of separation and detection used was optimized for low micromolar concentrations of maleic acid monomers but could perhaps be further optimized for reduction of polymer species. Preliminary method optimization resulted in observed differences in the proportions of polymers (data not shown), which is evidence that this is possible although similar published methods were not found.

In addition to the polymer species being unaccounted for, we do not know if any maleic acid remains inside the cell. In fact, if maleic acid levels are below a toxic threshold of about 40mM (reported for extracellular maleic acid [74]), the cell may not be obliged to expel maleic acid. However, it is unclear whether maleic acid efflux is due simply to abundance or to its possible effects on fitness. Future investigations may include quantification of intracellular maleic acid production via removal of culture supernatant and lysis of cells in buffer. Despite these inaccuracies in total maleic acid quantification, maleic acid levels are still likely to be less than ideal based simply on the fact that MIs appear to highly favour the forward fumaric acid to maleic acid reaction.

It was noticed that upon induction of MI expression, maleic acid levels dropped (Appendix Figure 1B). This suggests that the enzyme to fumaric acid substrate ratio may require balancing for optimal maleic acid production. Different levels of both MI expression and substrate should be tested. This could be achieved with the 6KO + pTRCMI strains using different amounts of IPTG induction or by feeding the substrate fumaric acid to MI-expressing WT strains. Furthermore, maleic acid levels appeared to be higher in shake-flasks compared to 96-well plates which could be due to differences in aeration of the cultures, although error bars were large in the plate dataset. Culture oxygen level was not a feature that was examined in these experiments but may provide insights to the mechanisms of the engineered strain if features of aerobic or anaerobic growth were affected. An improvement might also be seen with chromosomal integration of the MI so that the maintenance of a plasmid is not necessary. However, chromosomal integration would remove the ability to tinker with MI expression levels unless another induction system were introduced, for example, inducible promoters. Another intuitive route to improve maleic acid production would be the engineering of MI to achieve a more favourable reverse reaction via the principles of the Haldane relationship if possible [27]. While the equilibrium constant, K_{eq}, of maleic acid and fumaric cannot be altered, the kinetic values of a given forward or reverse reaction can [59, 75, 76]. Directed evolution may focus on the catalytic pocket or on the breathing loops of MIs.

In addition to the improvement of MI itself, an interesting aspect to ponder concerning maleic acid-production is how the maleic acid ends up in the culture supernatant. Dicarboxylate transport proteins (such as dctA [77] and dcuA [78]) are responsible for the uptake and efflux of a variety of molecules (such as succinate, fumarate, and orotate) demonstrating their promiscuous transport ability. The cell might exploit one of these transport proteins to rid itself of maleic acid being produced. A series of transport protein knockouts in a 6KO + pTRCMI strain might reveal which protein is being used. If a specific transport protein could be pinpointed, perhaps it could be engineered to prefer maleic acid secretion.

To summarize, the production of maleic acid in *E. coli* using MIs is possible but not yet optimal. A number of options are available for the amelioration of this maleic acid-producing strain based on the data obtained in this work.

CONCLUSIONS

The purpose of this research was to screen the activity of a library of enzyme candidates for the expression of an optimal enzyme candidate in a metabolically engineered microbial strain to produce an industrially important molecule. We successfully provide a proof-of-concept design for the production of maleic acid using a heterologous MI in an engineered *E. coli* strain. The outcome is far from ideal and would require many degrees of improvement before being suitable for industrial-scale applications and replacement of current maleic acid sources. However, this is the first time that any microbial strain has been engineered to produce maleic acid using MIs. Furthermore, this work expands on available information concerning MIs, which could be useful in future endeavours.

REFERENCES

1. Abatemarco, J., A. Hill, and H.S. Alper, *Expanding the metabolic engineering toolbox with directed evolution.* Biotechnology Journal, 2013. **8**(12): p. 1397-1410.
2. Rodriguez, G.M. and S. Atsumi, *Isobutyraldehyde production from Escherichia coli by removing aldehyde reductase activity.* Microbial Cell Factories, 2012. **11**: p. 11.
3. Bailey, J.E., et al., *Strategies and challenges in metabolic engineering.* Annals of the New York Academy of Sciences, 1990. **589**(1): p. 1-15.
4. Arnold, F.H., *Design by directed evolution.* Accounts of Chemical Research, 1998. **31**(3): p. 125-131.
5. Silver, P.A., et al., *Engineering explored.* Nature, 2014. **509**(7499): p. 166-167.
6. Lohbeck, K., et al., *Maleic and fumaric acids*, in *Ullmann's Encyclopedia of Industrial Chemistry*, B. Elvers, Editor. 2000, Wiley-VCH Verlag GmbH & Co. KGaA: Online.
7. Inc., G.V.R., *Maleic anhydride market analysis by application (unsaturated polyester resins (UPR), 1,4-butanediol (BDO), additives, copolymers) and segment forecasts to 2024.* 2016, Grand View Research Inc.: California. p. 1-148.
8. Davies, P., *High feedstock costs challenge maleic anhydride margins*. 2017, Tecnon OrbiChem: Surrey, England.
9. Yim, H., et al., *Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol.* Nature Chemical Biology, 2011. **7**(7): p. 445-52.
10. Weaver, D.S., et al., *A genome-scale metabolic flux model of Escherichia coli K-12 derived from the EcoCyc database.* BMC Systems Biology, 2014. **8**.
11. Keseler, I.M., et al., *The EcoCyc database: reflecting new knowledge about Escherichia coli K-12.* Nucleic Acids Research, 2017. **45**(D1): p. D543-D550.
12. Hasty, J., D. McMillen, and J.J. Collins, *Engineered gene circuits.* Nature, 2002. **420**(6912): p. 224-230.
13. Chen, D., et al., *Structural and computational studies of the maleate isomerase from Pseudomonas putida S16 reveal a breathing motion wrapping the substrate inside.* Molecular Microbiology, 2013. **87**(6): p. 1237-1244.
14. Tang, H., et al., *Genomic analysis of Pseudomonas putida: genes in a genome island are crucial for nicotine degradation.* Scientific Reports, 2012. **2**.
15. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.
16. Webb, J.L., *Enzyme and metabolic inhibitors*. Vol. 3. 1966, New York: Academic Press. 1060.
17. Hatakeyama, K., et al., *Gene cloning and characterization of maleate cis-trans isomerase from Alcaligenes faecalis.* Biochemical and Biophysical Research Communications, 1997. **239**(1): p. 74-79.
18. Dokainish, H.M., B.F. Ion, and J.W. Gauld, *Computational investigations on the catalytic mechanism of maleate isomerase: the role of the active site cysteine residues.* Physical Chemistry Chemical Physics, 2014. **16**(24): p. 12462-12474.
19. Placzek, S., et al., *BRENDA in 2017: new perspectives and new tools in BRENDA.* Nucleic Acids Research, 2017. **45**(D1): p. D380-D388.
20. Sacks, W. and C.O. Jensen, *Malease, a hydrase from corn kernels.* Journal of Biological Chemistry, 1951. **192**(1): p. 231-6.
21. Drevland, R.M., A. Waheed, and D.E. Graham, *Enzymology and evolution of the pyruvate pathway to 2-oxobutyrate in Methanocaldococcus jannaschii.* Journal of Bacteriology, 2007. **189**(12): p. 4391-4400.

22. Finn, R.D., et al., *The Pfam protein families database: towards a more sustainable future.* Nucleic Acids Research, 2016. **44**(D1): p. D279-D285.

23. Ray, J., et al., *Complete genome sequence of Cupriavidus basilensis 4G11, isolated from the Oak Ridge Field Research Center site.* Genome Announcements, 2015. **3**(3): p. e00322-15.

24. Scher, W. and W.B. Jakoby, *Maleate isomerase.* Journal of Biological Chemistry, 1969. **244**(7): p. 1878-1882.

25. Hatakeyama, K., et al., *Analysis of oxidation sensitivity of maleate cis-trans isomerase from Serratia marcescens.* Bioscience, Biotechnology, and Biochemistry, 2000. **64**(7): p. 1477-1485.

26. Hatakeyama, K., et al., *Molecular analysis of maleate cis-trans isomerase from thermophilic bacteria.* Bioscience, Biotechnology, and Biochemistry, 2000. **64**(3): p. 569-576.

27. Mellors, A., *The Haldane relationship: enzymes & equilibrium.* Biochemical Education, 1976. **4**(4): p. 1.

28. Ostergaard, S., L. Olsson, and J. Nielsen, *Metabolic engineering of Saccharomyces cerevisiae.* Microbiology and Molecular Biology Reviews, 2000. **64**(1): p. 34-+.

29. Liu, X., et al., *N-terminal truncation of a maleate cis-trans isomerase from Rhodococcus jostii RHA1 results in a highly active enzyme for the biocatalytic production of fumaric acid.* Journal of Molecular Catalysis B: Enzymatic, 2013. **93**: p. 44-50.

30. Fisch, F., et al., *A covalent succinylcysteine-like intermediate in the enzyme-catalyzed transformation of maleate to fumarate by maleate isomerase.* Journal of the American Chemical Society, 2010. **132**(33): p. 11455-11457.

31. Takamura, Y., et al., *Studies on the induced synthesis of maleate cis-trans isomerase by malonate.* Agricultural and Biological Chemistry, 1969. **33**(5): p. 718-728.

32. Sambrook, J. and D.W. Russell, *Molecular Cloning: A laboratory manual.* 3 ed. 2001, New York: CSHL Press.

33. Thomason, L.C., N. Costantino, and D.L. Court, *Escherichia coli genome manipulation by P1 transduction.* Current Protocols in Molecular Biology, 2007. **Chapter 1**: p. Unit 1.17-Unit 1.17.

34. Baba, T., et al., *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.* Molecular Systems Biology, 2006. **2**: p. 2006 0008.

35. Pyne, M.E., et al., *Reconstituting plant secondary metabolism in Saccharomyces cerevisiae for production of high-value benzylisoquinoline alkaloids.* Methods in Enzymology, 2016. **575**: p. 195-224.

36. Bergmans, H.E.N., I.M. Vandie, and W.P.M. Hoekstra, *TRANSFORMATION IN ESCHERICHIA-COLI - STAGES IN THE PROCESS.* Journal of Bacteriology, 1981. **146**(2): p. 564-570.

37. Sambrook, J. and D.W. Russell, *Transformation of Escherichia coli by electroporation.* CSH Protocols, 2006. **2006**(1).

38. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.

39. Wilde, C., et al., *Expression of a library of fungal beta-glucosidases in Saccharomyces cerevisiae for the development of a biomass fermenting strain.* Applied Microbiology and Biotechnology, 2012. **95**(3): p. 647-59.

40. Narcross, L., et al., *Mining enzyme diversity of transcriptome libraries through DNA synthesis for benzylisoquinoline alkaloid pathway optimization in yeast.* ACS Synthetic Biology, 2016. **5**(12): p. 1505-1518.

41. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0.* Molecular Biology and Evolution, 2013. **30**(12): p. 2725-2729.

42. Qiagen, *The QIAexpressionist™ A handbook for high-level expression and purification of 6xHis-tagged proteins*. 2003.

43. Kubota, K., et al., *Development of an HPLC-fluorescence determination method for carboxylic acids related to the tricarboxylic acid cycle as a metabolome tool.* Biomedical Chromatography, 2005. **19**(10): p. 788-95.

44. Tang, H., et al., *Systematic unraveling of the unsolved pathway of nicotine degradation in Pseudomonas.* PLoS Genetics, 2013. **9**(10).

45. Jimenez, J.I., et al., *Deciphering the genetic determinants for aerobic nicotinic acid degradation: the nic cluster from Pseudomonas putida KT2440.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(32): p. 11329-34.

46. Lawrence, J., *Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes.* Current Opinion in Genetics & Development, 1999. **9**(6): p. 642-648.

47. Huo, Y.Y., et al., *Complete genome sequence of Pelagibacterium halotolerans B2(T).* Journal of Bacteriology, 2012. **194**(1): p. 197-198.

48. Holder, J.W., et al., *Comparative and functional genomics of Rhodococcus opacus PD630 for biofuels development.* PLoS Genetics, 2011. **7**(9).

49. Isaksen, G.V., J. Aqvist, and B.O. Brandsdal, *Protein surface softness is the origin of enzyme cold-adaptation of trypsin.* PLoS Computational Biology, 2014. **10**(8).

50. She, Q., et al., *The complete genome of the crenarchaeon Sulfolobus solfataricus P2.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(14): p. 7835-7840.

51. Guerrero, L.D., et al., *Draft genome sequence of Williamsia sp. strain D3, isolated from the Darwin Mountains, Antarctica.* Genome Announcements, 2014. **2**(1).

52. Ishikawa, J., et al., *The complete genomic sequence of Nocardia farcinica IFM 10152.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(41): p. 14925-14930.

53. Kallimanis, A., et al., *Complete genome sequence of Arthrobacter phenanthrenivorans type strain (Sphe3).* Standards in Genomic Sciences, 2011. **4**(2): p. 123-130.

54. Cao, Y., et al., *Genome Sequencing of Ralstonia solanacearum FQY_4, Isolated from a Bacterial Wilt Nursery Used for Breeding Crop Resistance.* Genome Announcements, 2013. **1**(3): p. e00125-13.

55. Ghimire-Rijal, S., et al., *Duplication of genes in an ATP-binding cassette transport system increases dynamic range while maintaining ligand specificity.* Journal of Biological Chemistry, 2014. **289**(43): p. 30090-30100.

56. Aasland, R., et al., *Normalization of nomenclature for peptide motifs as ligands of modular protein domains.* FEBS Letters, 2002. **513**(1): p. 141-144.

57. You, L. and F.H. Arnold, *Directed evolution of subtilisin E in Bacillus subtilis to enhance total activity in aqueous dimethylformamide.* Protein Engineering, 1996. **9**(1): p. 77-83.

58. McIsaac, R.S., et al., *Directed evolution of a far-red fluorescent rhodopsin.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(36): p. 13034-13039.

59. Larue, K., M. Melgar, and V.J.J. Martin, *Directed evolution of a fungal beta-glucosidase in Saccharomyces cerevisiae.* Biotechnology for Biofuels, 2016. **9**.

60. Owens, A.E., et al., *Two-tier screening platform for directed evolution of aminoacyl-tRNA synthetases with enhanced stop codon suppression efficiency.* ChemBioChem, 2017. **18**(12): p. 1109-1116.

61. Otero, J.M., et al., *Industrial systems biology of Saccharomyces cerevisiae enables novel succinic acid cell factory.* PLoS One, 2013. **8**(1): p. e54144.

62. Wang, Z., et al., *Production of pyruvate in Saccharomyces cerevisiae through adaptive evolution and rational cofactor metabolic engineering.* Biochemical Engineering Journal, 2012. **67**: p. 126-131.

63. Weber, C., et al., *Biosynthesis of cis,cis-muconic acid and its aromatic precursors, catechol and protocatechuic acid, from renewable feedstocks by Saccharomyces cerevisiae.* Applied and Environmental Microbiology, 2012. **78**(23): p. 8421-8430.
64. Yilmaz, L.S. and A.J.M. Walhout, *Metabolic network modeling with model organisms.* Current Opinion in Chemical Biology, 2017. **36**: p. 32-39.
65. Pandit, A.V., S. Srinivasan, and R. Mahadevan, *Redesigning metabolism based on orthogonality principles.* Nature Communications, 2017. **8**: p. 1-11.
66. Warnecke, T. and R.T. Gill, *Organic acid toxicity, tolerance, and production in Escherichia coli biorefining applications.* Microbial Cell Factories, 2005. **4**: p. 25.
67. Vazquez, J.A., et al., *Evaluation of toxic effects of several carboxylic acids on bacterial growth by toxicodynamic modelling.* Microbial Cell Factories, 2011. **10**.
68. Zhang, T., et al., *Pull-in urea cycle for the production of fumaric acid in Escherichia coli.* Applied Microbiology and Biotechnology, 2015. **99**(12): p. 5033-5044.
69. Song, C.W., et al., *Metabolic engineering of Escherichia coli for the production of fumaric acid.* Biotechnology and Bioengineering, 2013. **110**(7): p. 2025-2034.
70. Li, N., et al., *Engineering Escherichia coli for fumaric acid production from glycerol.* Bioresource Technology, 2014. **174**: p. 81-87.
71. Joyce, A.R., et al., *Experimental and computational assessment of conditionally essential genes in Escherichia coli.* Journal of Bacteriology, 2006. **188**(23): p. 8259-71.
72. Gold, N.D., et al., *Metabolic engineering of a tyrosine-overproducing yeast platform using targeted metabolomics.* Microbial Cell Factories, 2015. **14**(1): p. 73.
73. Teleki, A., et al., *Robust identification of metabolic control for microbial L-methionine production following an easy-to-use puristic approach.* Metabolic Engineering, 2017. **41**: p. 159-172.
74. Nicolle, J. and Y. Joyeux, *Action des acides fumarique et maléique sur la croissance de certaines espèces bactériennes.* Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, 1950. **231**: p. 3.
75. Giger, L., et al., *Evolution of a designed retro-aldolase leads to complete active site remodeling.* Nature Chemical Biology, 2013. **9**(8): p. 494-U49.
76. Prier, C.K., et al., *Enantioselective, intermolecular benzylic C-H amination catalysed by an engineered iron-haem enzyme.* Nature Chemistry, 2017. **9**(7): p. 629-634.
77. Baker, K.E., et al., *Utilization of orotate as a pyrimidine source by Salmonella typhimurium and Escherichia coli requires the dicarboxylate transport protein encoded by dctA.* Journal of Bacteriology, 1996. **178**(24): p. 7099-7105.
78. Six, S., et al., *Escherichia coli possesses two homologous anaerobic C4-dicarboxylate membrane transporters (DcuA and DcuB) distinct from the aerobic dicarboxylate transport system (Dct).* Journal of Bacteriology, 1994. **176**(21): p. 6470-8.
79. Xu, G.Q., et al., *Fumaric acid production in Saccharomyces cerevisiae by simultaneous use of oxidative and reductive routes.* Bioresource Technology, 2013. **148**: p. 91-96.
80. Zelle, R.M., et al., *Malic acid production by Saccharomyces cerevisiae: engineering of pyruvate carboxylation, oxaloacetate reduction, and malate export.* Applied and Environmental Microbiology, 2008. **74**(9): p. 2766-77.

APPENDICES

Appendix I: Results of preliminary activity screen of MIs in crude lysate.

The catalytic activity of the putative MIs was investigated in both predicted directions of the isomerase reaction using maleic acid and fumaric acid as substrates. The ability of MIs to perform the catalytically favoured forward reaction (maleic acid to fumaric acid) was assessed by an observed increase in the optical density measured at 290nm signifying the formation of fumaric acid in solution. Preliminary activity assays were performed with crude whole cell lysate preparations of induced MI-expressing BL21(DE3) cells. The lysates of 27 cultures with different MI variants were assessed to verify the suitability of the spectrophotometric assay. Detectable levels of forward activity were observed with 13 of the crude lysate MI preparations (Figure 8). Of the remaining 14 lysates not showing any forward activity, 3 turned out to be expressing insoluble MIs and the solubility of another 1 was not confirmed, meaning that 10 were inactive with the crude assay conditions. These results assured us that some of the MIs were active in the forward direction and that this activity could be measured using a spectrophotometer. Because the MI protein concentrations could not be quantified in these crude lysate assays, the results could be interpreted qualitatively only and were not informative enough to fully characterize the kinetic values.

Appendix II: Potential model for maleic acid production in S. cerevisiae

We explored the idea of building a maleic acid-producing *S. cerevisiae* strain, however multiple complications led us to pursue maleic acid-production in *E. coli.* The literature was mined for examples of previous successful overproduction of suitable maleic acid precursor molecules (fumaric acid and malate) in *S. cerevisiae.* An engineered *S. cerevisiae* strain that can produce 5.64g/l fumaric acid was reported in 2013 [79], however, this fumaric acid-producing *S. cerevisiae* strain had not been discovered at the start of this project (early 2015). Two models were found which could have been used in this study with some modifications. The first model overproduced succinate, which precedes fumaric acid in the TCA cycle [61]. In order to convert the succinate-producing model into a fumaric acid-producing one, the KO of succinate dehydrogenase, *SDH1*, would be excluded and instead the KO of the downstream fumarase, *FUM1*, was theorized to result in the strain producing fumaric acid. In order to produce maleic acid, the expression of an MI would also be included. (Appendix Figure 3: Option I). Another possible strategy would be based on a strain capable of overproducing malate which includes the overexpression of pyruvate carboxylase, *PYC2*, the removal of the peroxisomal targeting sequence from malate dehydrogenase, *MDH3*, so that it expresses in the cytosol, and the heterologous expression of malate transporter gene from *Schizosaccharomyces pombe*, Sp*MAE1* [80]. The expression of a non-native protein with malate hydratase activity would be included to produce maleic acid (Appendix Figure 3: Option II). Both of these strategies present a possible issue of transport through cellular compartments, and the second approach also included the problem that there is very little evidence of the existence of maleate hydratases with favourable activity versus the existence of the desired reaction simply as a side reaction [21]. A heterologous promiscuous enzyme might have also had activity on other metabolites in the cells thus further complicating the metabolic engineering as has been seen with cytochrome P450s [40].

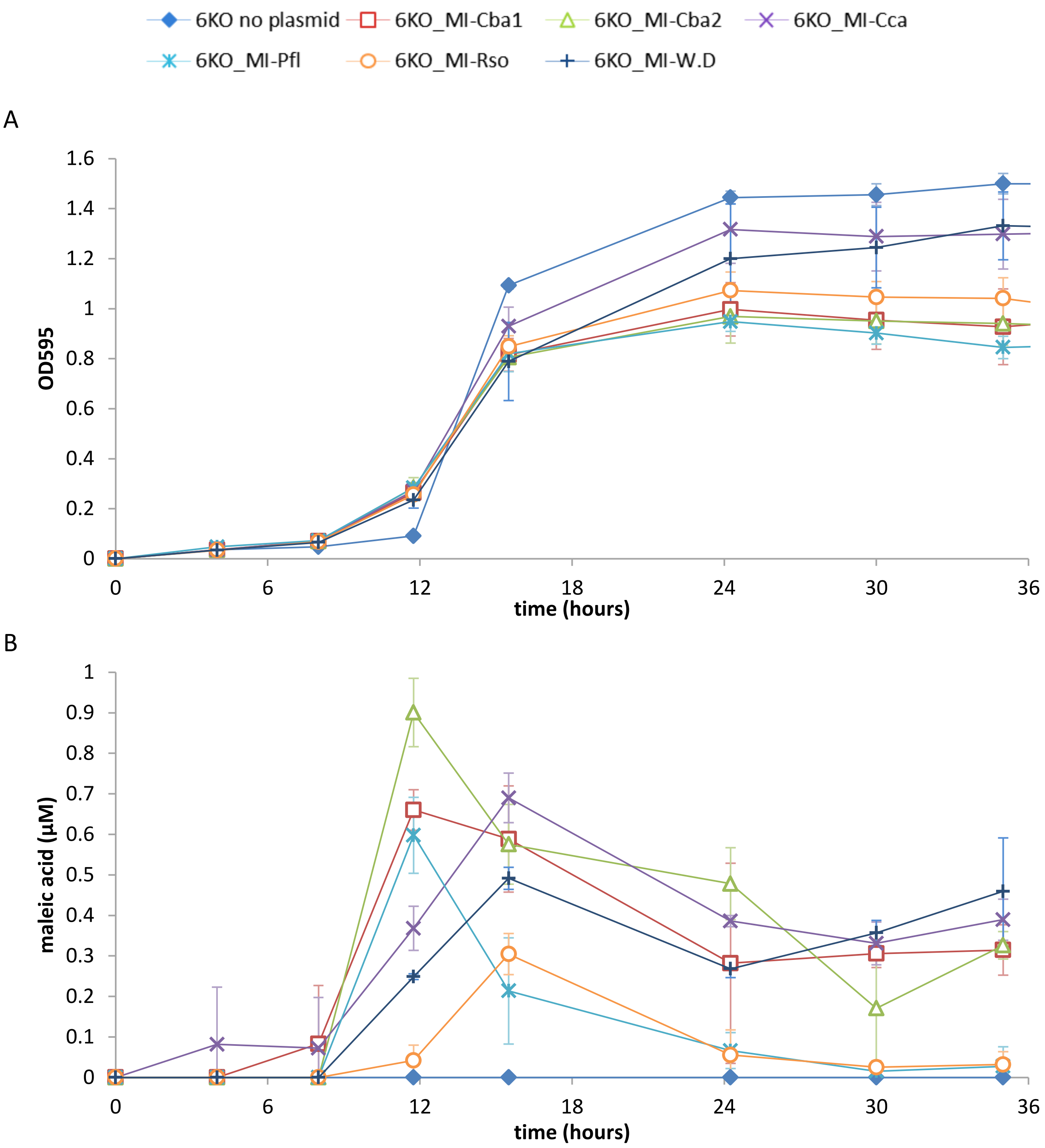


Figure A1: Growth and maleic acid quantification of 6KO with MIs expressed and induced.

A preliminary time course assay in shake flasks was set up in triplicate. The pTRCMI- vectors were induced with 1mM IPTG at 11.75hrs at OD_{595} of approximately 0.25. The induction appears to have caused a decline in the amount of maleic acid produced. The cases are as described in Figure 12 except for 6KO with no plasmid, which was not a part of the uninduced MI experiments, and is represented by filled blue diamonds. Error bars indicate the standard deviation between replicates.

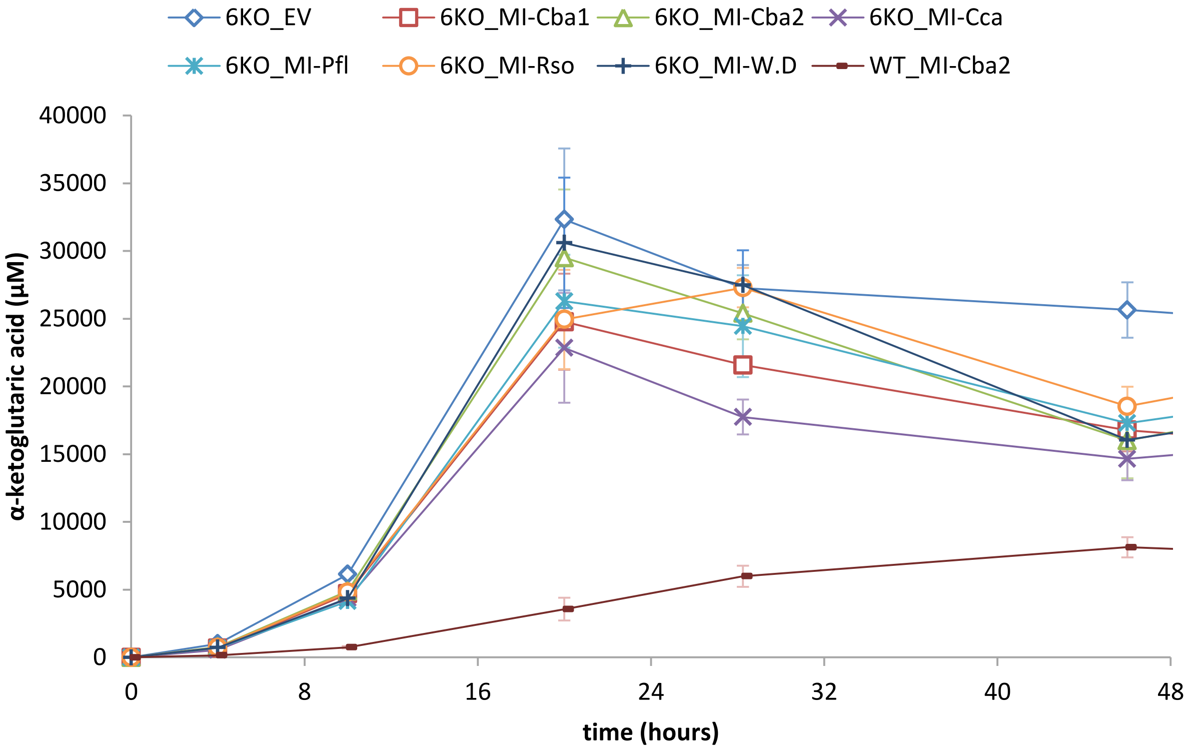


Figure A2: Profile of α-ketoglutaric acid in strains with MIs expressed.

Experiment conditions and graph are as described in Figure 12.

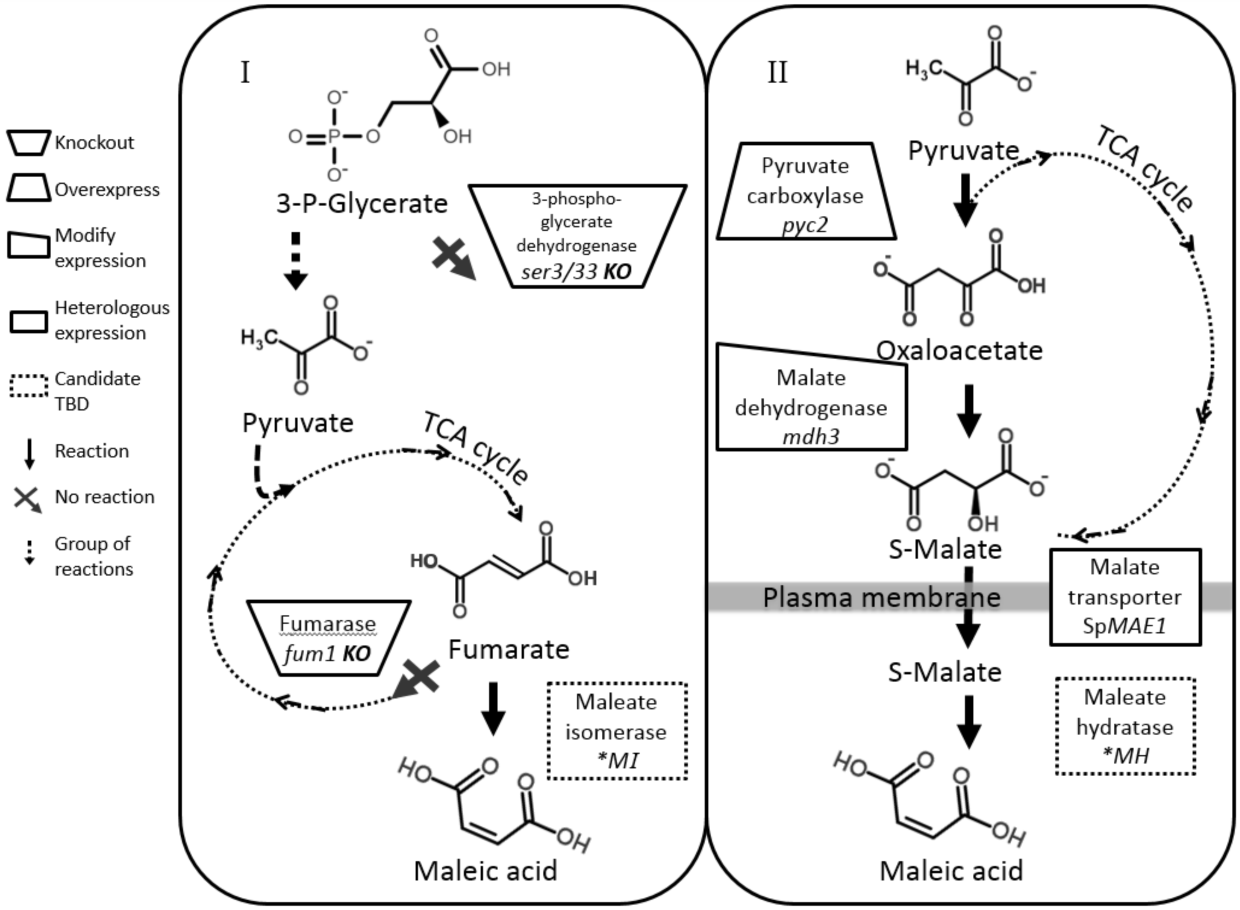


Figure A3: Proposed synthetic *S. cerevisiae* pathways for maleic acid production.

Building on previously established synthetic *S. cerevisiae* strains, it is hypothetically possible to produce maleic acid. Based on a strain designed to overproduce succinate [61], Option I includes the KOs of *ser3/33* and *fum1*, and the expression of a non-native MI. Based on a strain capable of overproducing malate [80], Option II includes the overexpression of *pyc2*, the removal of the peroxisomal targeting sequence from *mdh3* so that it expresses in the cytosol, the heterologous expression of Sp*MAE1*, and the expression of a non-native protein with maleate hydratase activity.

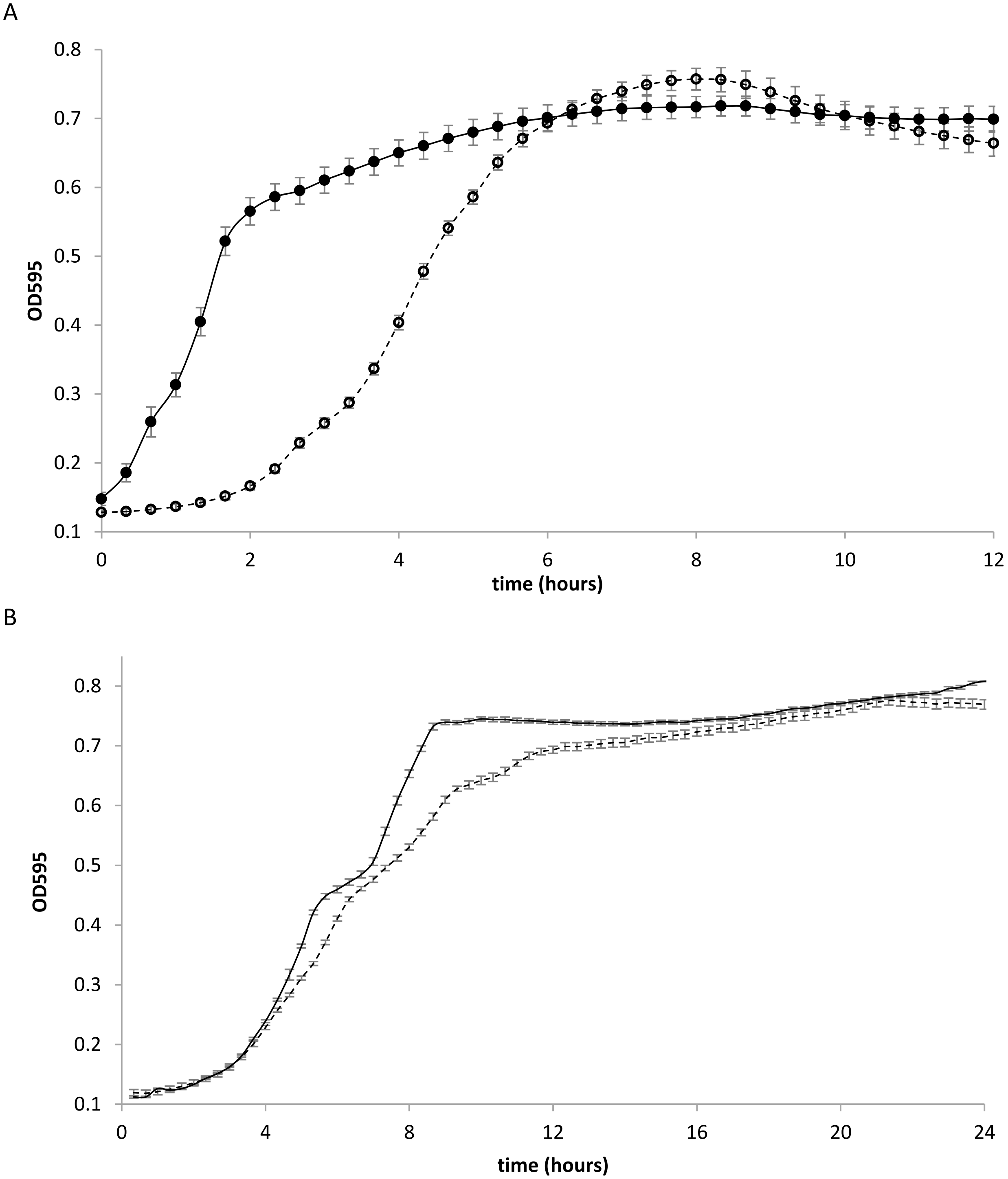
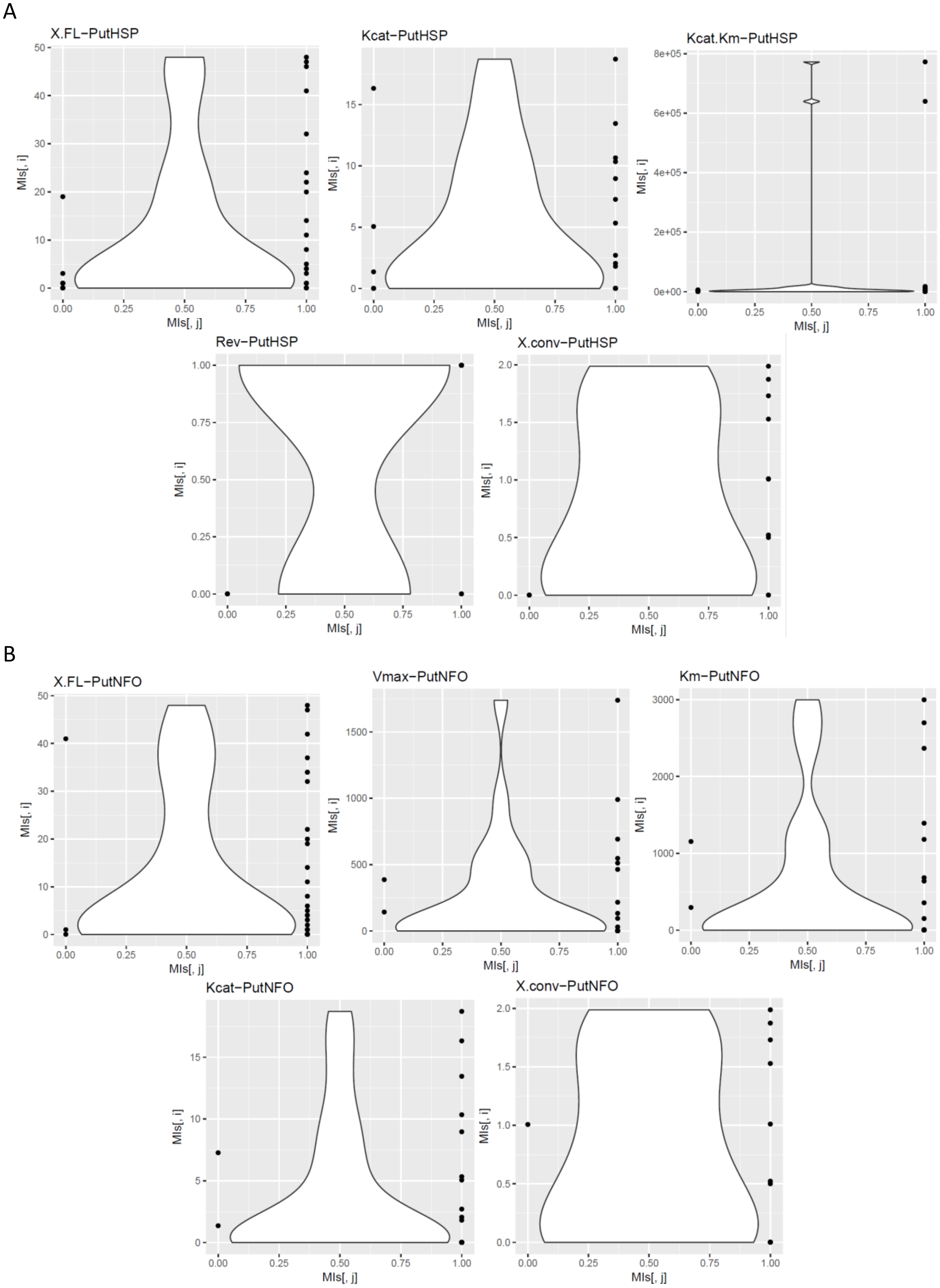
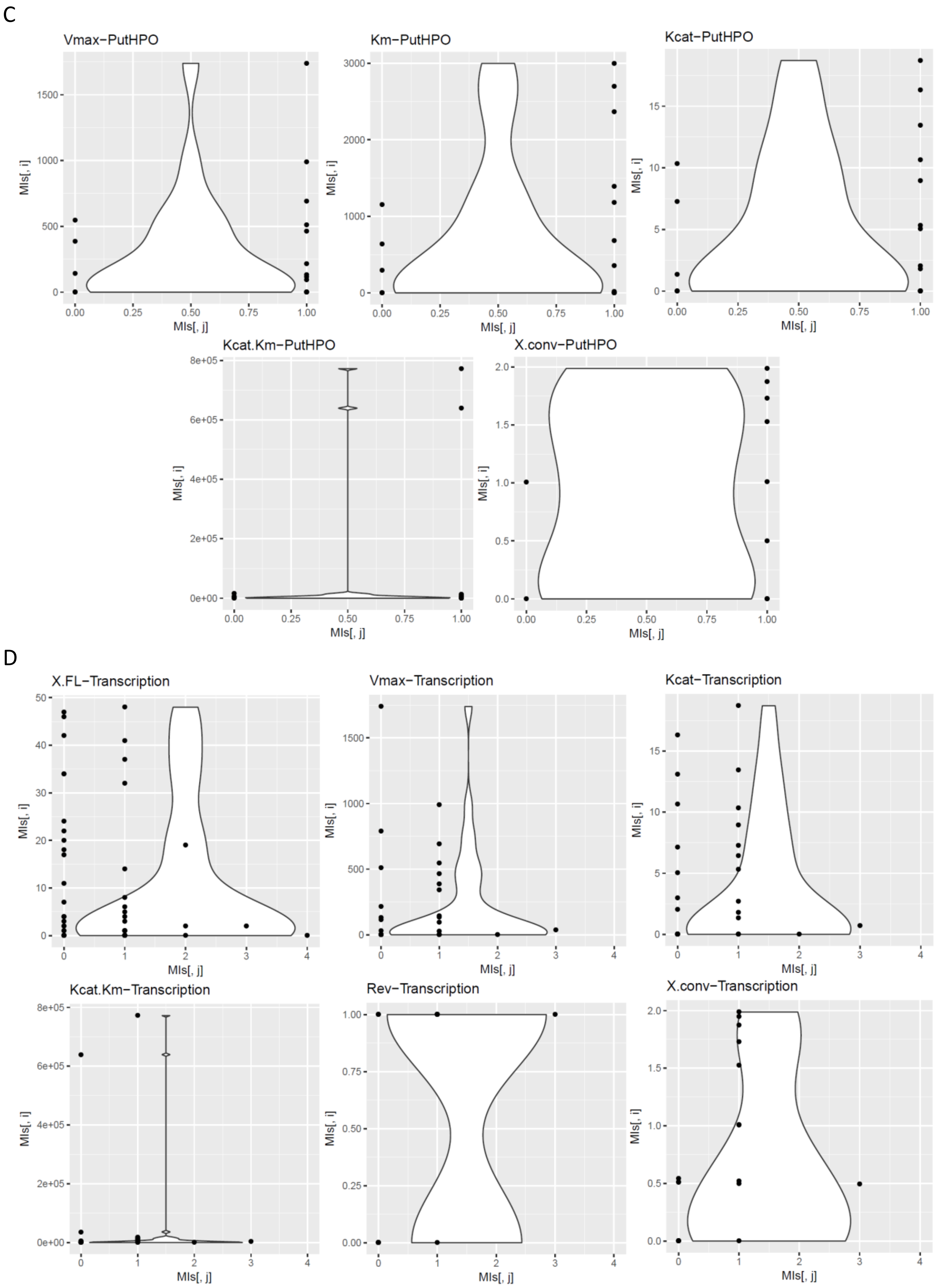


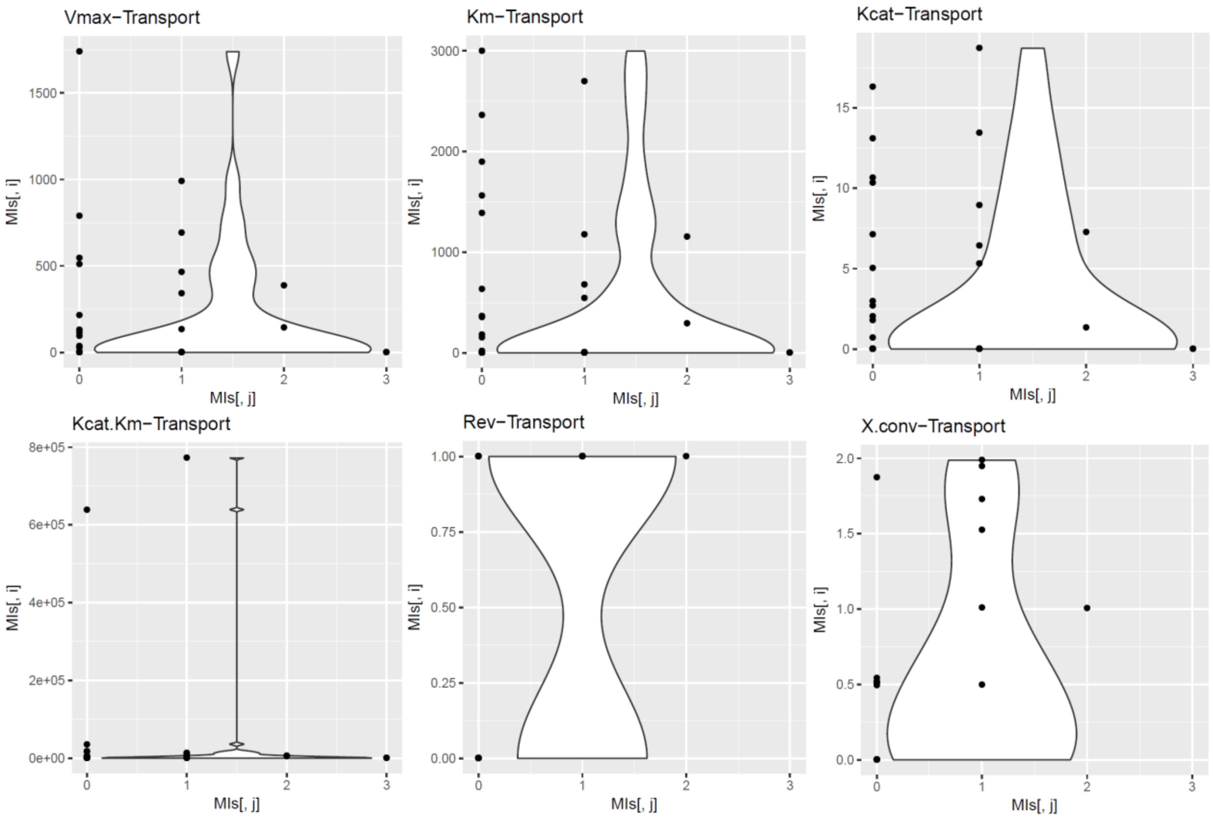
Figure A4: Growth of WT MG1655 and 6KO in LB and M9

WT (solid line) and 6KO (dashed line) strains were grown in triplicate in 180µl rich LB media (A), or minimal M9 media (B) in 96-well plates. Absorbance readings were taken at 595nm with the Sunrise® absorbance microplate reader (Tecan) every 20 min. Error bars represent standard deviation of triplicates.


A
X.FL−PutHSP
Kcat−PutHSP
Kcat.Km−PutHSP
Rev−PutHSP
X.conv−PutHSP
B
X.FL−PutNFO
Vmax−PutNFO
Km−PutNFO
Kcat−PutNFO
X.conv−PutNFO
MIs[, i]
MIs[, j]
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50
5
15
0e+00
2e+05
4e+05
6e+05
8e+05
0.5
1.0
1.5
2.0
500
1000
1500
2000
3000


C
Vmax−PutHPO
0
500
1000
1500
MIs[, i]
0.00
0.25
0.50
0.75
1.00
MIs[, j]
Km−PutHPO
0
1000
2000
3000
MIs[, i]
0.00
0.25
0.50
0.75
1.00
MIs[, j]
Kcat−PutHPO
0
5
10
15
MIs[, i]
0.00
0.25
0.50
0.75
1.00
MIs[, j]
Kcat.Km−PutHPO
0e+00
2e+05
4e+05
6e+05
8e+05
MIs[, i]
0.00
0.25
0.50
0.75
1.00
MIs[, j]
X.conv−PutHPO
0.0
0.5
1.0
1.5
2.0
MIs[, i]
0.00
0.25
0.50
0.75
1.00
MIs[, j]
D
X.FL−Transcription
0
10
20
30
40
50
MIs[, i]
0
1
2
3
4
MIs[, j]
Vmax−Transcription
0
500
1000
1500
MIs[, i]
0
1
2
3
4
MIs[, j]
Kcat−Transcription
0
5
10
15
MIs[, i]
0
1
2
3
4
MIs[, j]
Kcat.Km−Transcription
0e+00
2e+05
4e+05
6e+05
8e+05
MIs[, i]
0
1
2
3
4
MIs[, j]
Rev−Transcription
0.00
0.25
0.50
0.75
1.00
MIs[, i]
0
1
2
3
4
MIs[, j]
X.conv−Transcription
0.0
0.5
1.0
1.5
2.0
MIs[, i]
0
1
2
3
4
MIs[, j]

E


Vmax−Transport
MIs[, i]
0
500
1000
1500
MIs[, j]
0
1
2
3
Km−Transport
MIs[, i]
0
1000
2000
3000
MIs[, j]
0
1
2
3
Kcat−Transport
MIs[, i]
0
5
10
15
MIs[, j]
0
1
2
3
Kcat.Km−Transport
MIs[, i]
0e+00
2e+05
4e+05
6e+05
8e+05
MIs[, j]
0
1
2
3
Rev−Transport
MIs[, i]
0.00
0.25
0.50
0.75
1.00
MIs[, j]
0
1
2
3
X.conv−Transport
MIs[, i]
0.0
0.5
1.0
1.5
2.0
MIs[, j]
0
1
2
3

F

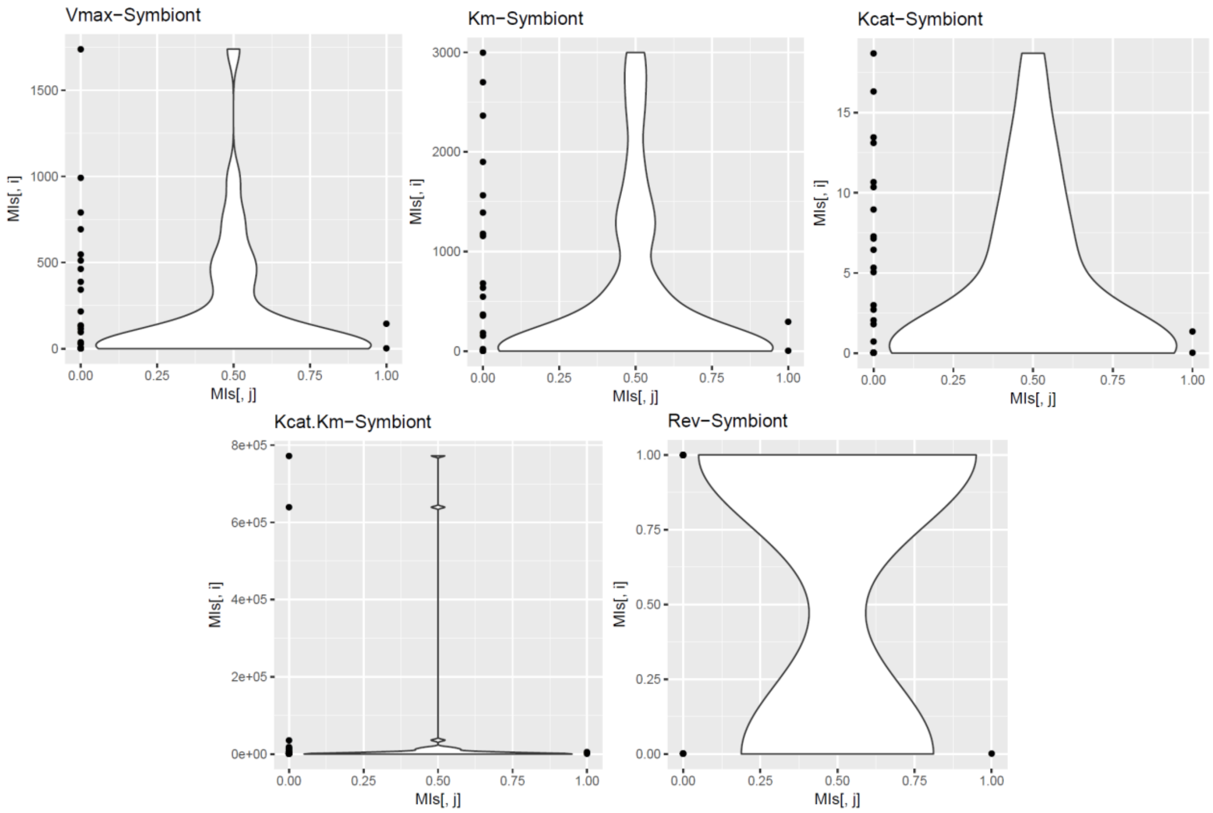

Vmax−Symbiont
MIs[, i]
0
500
1000
1500
MIs[, j]
0.00
0.25
0.50
0.75
1.00
Km−Symbiont
MIs[, i]
0
1000
2000
3000
MIs[, j]
0.00
0.25
0.50
0.75
1.00
Kcat−Symbiont
MIs[, i]
0
5
10
15
MIs[, j]
0.00
0.25
0.50
0.75
1.00
Kcat.Km−Symbiont
MIs[, i]
0e+00
2e+05
4e+05
6e+05
8e+05
MIs[, j]
0.00
0.25
0.50
0.75
1.00
Rev−Symbiont
MIs[, i]
0.00
0.25
0.50
0.75
1.00
MIs[, j]
0.00
0.25
0.50
0.75
1.00

Figure A5: Trends in MI activity vs. MI qualities.

Different parameters were graphed against one another using R software and violin plots of the complete dataset for the MI library. The above plots were chosen as they displayed some visual correlation between the parameters which are listed as y-values vs. x-values in the plot titles. The relationships showing trends include parameters plotted against the traits: A) the presence of a putative HSP from *nic*2 cluster, B) the presence of a putative NFO from *nic*2 cluster, C) the presence of a putative HPO from *nic*2 cluster, D) the number of transcription factors in the ±4-ORF genomic space, E) the number of transport genes in the ±4-ORF space, and F) symbiont status (1 if yes, 0 if no). The parameters that appear to be influenced by the traits are: % fluorescence in *S. cerevisiae* cells (in A,B,D), V_{max} (B-F), K_M (in B,C,E,F), k_{cat} (all traits), k_{cat}/K_M (A,C-F), presence (1) or absence (0) of reverse activity (A,D-F), and % reverse conversion *in vitro* (A-E).

Table A1: Strains built in this study.

Strain ID	# of KOs	Description
MM090	1	ΔpykF::FRT
MM292	2	ΔpykF::FRT, ΔgdhA::kan
MM293	2	ΔpykF::FRT, ΔpykA::kan
MM294	2	ΔpykF::FRT, ΔmaeB::kan
MM295	2	ΔpykF::FRT, ΔfrdA::kan
MM296	2	ΔpykF::FRT, ΔsfcA::kan
MM297	2	ΔpykF::FRT, ΔgdhA::FRT
MM298	2	ΔpykF::FRT, ΔpykA::FRT
MM299	2	ΔpykF::FRT, ΔmaeB::FRT
MM300	2	ΔpykF::FRT, ΔfrdA::FRT
MM301	2	ΔpykF::FRT, ΔsfcA::FRT
MM302	3	ΔpykF::FRT, ΔgdhA::FRT, ΔpykA::kan
MM305	3	ΔpykF::FRT, ΔgdhA::FRT, ΔsfcA::kan
MM312	3	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::kan
MM160	3	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT
MM217	4	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔgdhA::kan
MM218	4	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔpykA::kan
MM219	4	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔsfcA::kan
MM310	4	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔpykA::FRT
MM314	4	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔsfcA::FRT
MM315	5	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔpykA::FRT, ΔsfcA::kan
MM318	5	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔsfcA::FRT, ΔgdhA::kan
MM319	5	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔsfcA::FRT, ΔpykA::kan
MM322	5	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔpykA::FRT, ΔsfcA::FRT
MM325	6	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔpykA::FRT, ΔsfcA::FRT, ΔgdhA::kan
MM326	6	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔpykA::FRT, ΔsfcA::FRT, ΔgdhA::FRT +pCP20
MM336	6	ΔpykF::FRT, ΔmaeB::FRT, ΔfrdA::FRT, ΔpykA::FRT, ΔsfcA::FRT, ΔgdhA::FRT
MM381	6	6KO + pTRCMI-PpuS
MM382	6	6KO + pTRCMI-Pfl
MM383	6	6KO + pTRCMI-Rso
MM384	6	6KO + pTRCMI-R.G
MM385	6	6KO + pTRCMI-Cba1
MM386	6	6KO + pTRCMI-Cba2
MM387	6	6KO + pTRCMI-P.U
MM388	6	6KO + pTRCMI-Cbae
MM389	6	6KO + pTRCMI-S.N2
MM390	6	6KO + pTRCMI-Aor
MM391	6	6KO + pTRCMI-Grh
MM392	6	6KO + pTRCMI-Gar
MM393	6	6KO + pTRCMI-W.D
MM394	6	6KO + pTRCMI-B.R
MM395	6	6KO + pTRCMI-P.B
MM396	6	6KO + pTRCMI-Cca
MM397	6	6KO + pTRCMI-Pas
MM398	0	WT + pTRCMI-Cba2

All strains were built from the E. coli *strain MG1655 (WT).*

Table A2: Primers used for plasmid construction with MI ORFs.

Library #:	MI ID	Primer: Internal check	Primer: 5´ addition of *Nde*I (and *Eco*RI) or *Ase*I RE cut site	Primer: 3´ addition of *Xho*I RE cut site
MI01	MI-PpuS	MM071 GGTGACTACAGAGCTTTAGAAATCTC	MM141 ggaattccatATGGAAGCCATGAAAGTCGTTC	MM201 cgatcctcgagATAAGCACCGGATAACAAAGTACC
MI02	MI-Rjo	MM072 TAACGCCGAAGTCGGTTGTA	MM291 ggaattccatATGAAGGGTCACAACATGGGT	MM202 cgatcctcgagTTGAACAGTGGTAGCTGGAGC
MI03	MI-Nfa	MM073 CGACAACACTGAAGTCGGTTG	MM143 ggaattccatATGGGTATTCGTCGTATTGGTTTG	MM203 cgatcctcgagGGAAGCGGTGACGGC
MI04	MI-Bst	MM074 CGGTATCGAAGTCACCGATTC	MM144 ggaattccatATGGCTAAGCACTTTCGTATTGG	MM204 cgatcctcgagGTATTTACCAGACAACAAGGAACCAG
MI05	MI-Sma	MM075 TTAGACGTTGGTCGTCACGA	MM145 ggaattccatATGTCTAACCATTATCGTATTGGTCAAATC	MM205 cgatcctcgagATAAGCACCGGACAACAAGG
MI06	MI-Afa	MM076 CCGATAACTTGGAAGTCGGTT	MM146 ggaattccatATGAAGACTTACCGTATCGGTCA	MM206 cgatcctcgagAGCTTGTGGCTTAACACCAG
MI07	MI-Pfl	MM077 GAAGTTGCTCGTCATGATCCA	MM147 ggaattccatATGCAAAAGCCATACCGTATCG	MM207 cgatcctcgagATAAGCACCAGATAACAAAGCACC
MI08	MI-Rso	MM078 TTGGTCGTCAAGACCCTAGAG	MM148 ggaattccatATGTCCACTGCTCGTGCT	MM208 cgatcctcgagGTAAGCACCAGACAACAAAGAAC
MI09	MI-R.G	MM079 ACAGAGCCTTGGAAATCCCA	MM149 ggaattccatATGCAAAAGGTTTTCCGTATTGGT	MM209 cgatcctcgagGTAAGCACCGGACAACAAAGC
MI10	MI-Sso	MM080 CTGCCGGTATGGGTATCAGA	MM150 ggaattccatATGGAATGGATTAAGGTCGGTGT	MM210 cgatcctcgagAGTGATTTCGAACAAGTGACCC
MI11	MI-Cne	MM081 CCAGACAATTTGGAAGTCGGTC	MM151 ggaattccatATGACCGAATCTACTGTTCAAAAAGT	MM211 cgatcctcgagATAGGCACCAGACAATAAAGCAC
MI12	MI-Cba1	MM082 CGTGCTTTGGAAATTCCAGAC	MM152 ggaattccatATGACTGAGTCCACCGTCC	MM212 cgatcctcgagGTAAGCACCAGACAACAAAGCA
MI13	MI-Cba2	MM083 GTGCCTTGGAAATCCCAGATA	MM153 ggaattccatATGCAAAAGACTTTTCGTATTGGTC	MM213 cgatcctcgagGTAGGCACCGGACAACAAG
MI14	MI-Pha	MM084 AATAACTTGGACGTTGCTGCT	MM154 ggaattccatATGGAAACCCGTATGAAGACTTACA	MM214 cgatcctcgagATAGGCACCGGACAACAAAG
MI15	MI-P.U	MM085 CGAAGTTGTCACTTGGAGAGC	MM155 ggaattccatATGACCCCAATCCGTGTCG	MM215 cgatcctcgagATAAGCACCAGACAACAAAGTACC
MI16	MI-Cbae	MM086 GCTCACGACCCAATGAACTT	MM156 ggaattccatATGGAAACTGAAATCCCAGCC	MM216 cgatcctcgagGTAGGCACCGGATAACAAGG
MI17	MI-Cin	MM087 AGACAACTTAGATGTTGCCGC	MM157 ggaattccatATGGGTATTATGAAATCTTTCCGTATCG	MM217 cgatcctcgagATAAGCACCAGACAATAAGGTACC
MI18	MI-P.C	MM088 AGTTGTTGACTGGAGAGCCTT	MM158 ggaattccatATGATCAAGCCTTACCGTATCGG	MM218 cgatcctcgagCTTGTAAGCACCAGACAACAAAG
MI19	MI-PpuN	MM089 TGGAGAGCTTTGGAAATCCCA	MM159 ggaattccatATGACTAAATTGTACCGTATTGGTCAAA	MM219 cgatcctcgagGTAAGCACCGGACAACAAAGC
MI20	MI-Aac	MM090 GAACTTGGTCGACATCGTCG	MM160 ggttccattaATGGCTAAAAACTACCGTATCGG	MM220 cgatcctcgagGTAAGCACCGGACAATAAACG
MI21	MI-Msm	MM091 TAAGGTCGATTTACGTGACGC	MM161 ggaattccatATGACTGATTACCACGTCGGT	MM221 cgatcctcgagAGTTTCAACTTCACGAGAGACAGA
MI22	MI-S.N1	MM092 CTGATAACACCGAAGTCGGTTG	MM162 ggaattccatATGACCTCCACTTCCCGTC	MM222 cgatcctcgagAGCTTGAACCTTGGCATCAG
MI23	MI-S.N2	MM093 AAGGTTTTGAAATCGCCGGTT	MM163 ggaattccatATGTCTACTCATCGTATTGGTTTAGTTG	MM223 cgatcctcgagAACAGTAGCTCTAGCACGCA
MI24	MI-S.S	MM094 CGCTGAAGTTGGTAGAATTCCT	MM164 ggttccattaATGCCTGCTCCAACTCGT	MM224 cgatcctcgagGTTAGTGTGCAACAAGGAACCG
MI25	MI-Pdi	MM095 GCTGCTTTGGAAGTCGATGA	MM165 ggttccattaATGACTCACCGTATTGGTTTAGTC	MM225 cgatcctcgagACGAGAACGAGCCAACAAAG
MI26	MI-Abe	MM096 CCGATAACACTGAAGTCGGTTG	MM166 ggaattccatATGACTTCCAATTCCCGTCGT	MM226 cgatcctcgagAGCTTGGACTTTAGCGTCG
MI27	MI-Ach	MM097 GGTGATAACACCGAAGTTGGT	MM167 ggaattccatATGATCGTCCCTTCTTCTAACGT	MM227 cgatcctcgagACCGGCGTGGGCA
MI28	MI-Aor	MM098 TTGGAAGTCTCTGACAACAACG	MM168 ggaattccatATGTCCACTCATCGTATTGGTTTG	MM228 cgatcctcgagAGCAGTAGCACCAGCACG
MI29	MI-Grh	MM099 TGTCATGGCTGCTGCTAGAT	MM169 ggaattccatATGACCGAGCAAGGTTCCG	MM229 cgatcctcgagAGCTGGAACAGATTCAGTGGC
MI30	MI-Gar	MM072 TAACGCCGAAGTCGGTTGTA	MM170 ggaattccatATGACTGTTCCACACCGTATTG	MM230 cgatcctcgagAGCTGGGGTAACATTGACATCA
MI31	MI-Gkr	MM101 GCTTTGGAAGTCGCTGATAACA	MM171 ggaattccatATGTCCAACTCCCAAGTCGC	MM231 cgatcctcgagAGCGGTGACTGGAGCG
MI32	MI-Nno	MM102 TTGGTTGTGTCCCAGGTGAA	MM172 ggaattccatATGTATGTTAATTTGGGTGAATCTTCCT	MM232 cgatcctcgagGTGAGGGGCAACGGC
MI33	MI-Nbr	MM103 CTGAAGTTGGTTGTATCCCAGG	MM173 ggaattccatATGGGTATTCACCGTATCGGT	MM233 cgatcctcgagAACAGAAGTAGCAGCGGTGT
MI34	MI-W.D	MM104 CGTGCTTTAGAGGTTGCTGAT	MM174 ggaattccatATGTCTGTTCATCGTATTGGTTTGG	MM234 cgatcctcgagAGCTGGAACGGAAGCGA
MI35	MI-Rop	MM105 GGTTGTATCCCAGGTGACAGA	MM175 ggaattccatATGACTAACCGTCGTTGGAGA	MM235 cgatcctcgagTGGCAAGGCTGGAGCT
MI36	MI-Rfa	MM106 GTTGGTTGTATCCCAGGTGAAC	MM176 ggaattccatATGTCTATTCACCGTGTCGGT	MM236 cgatcctcgagGGAGGTAACGACGGCATCA
MI37	MI-Rwr	MM107 GTATCCCAGGTGACCGTGT	MM177 ggaattccatATGGGTATCCACCGTATCGG	MM237 cgatcctcgagTTGCAAAGCAGCAGCTGG
MI38	MI-Car	MM108 GGTTGTATCCCAGGTGATCGT	MM178 ggaattccatATGAAGCACGTCCGTGTTG	MM238 cgatcctcgagAGCTGGAACTGGAACATCAGC
MI39	MI-T.1	MM109 GGTTGTATTCCAGGTGATCGTG	MM179 ggaattccatATGACCAAGCGTATTGGTTTAGTT	MM239 cgatcctcgagTGGGGCTTCTAAAGCAACATC
MI40	MI-Rpi	MM110 CAACTTGGCCGTTGGTAGATT	MM180 ggaattccatATGAAGAACTTCCGTATCGGTCA	MM240 cgatcctcgagAAAAGCACCGGACAACAAATAACC
MI41	MI-Axy	MM111 CTTGAAGGTTGGTGCTCAAGAT	MM181 ggaattccatATGACCCAATCTCCTCGTGT	MM241 cgatcctcgagGTATTTACCGGTCAACAACTCACC
MI42	MI-Agu	MM112 CATTGCTGTCGCTCAACAC	MM182 ggaattccatATGCAACGTATTTACAAGATTGGTCA	MM242 cgatcctcgagGGCGTAGGCACCAGACA
MI43	MI-H.C	MM113 CTTAGAAGTTGGTCGTCACGAC	MM183 ggaattccatATGTTCTCCACTGGTAATGCTAATC	MM243 cgatcctcgagGTAAGCACCAGACAACAAAGCA
MI44	MI-B.R	MM114 CTTGGATGTCGCTCGTCATG	MM184 ggaattccatATGTCTAAGTCTTTCCGTATCGGT	MM244 cgatcctcgagGTAAGCACCGGACAACAAAGC
MI45	MI-P.B	MM115 CAACTTGGCTGTCGGTCG	MM292 ggaattccatATGGTTTCTTTATTGCGTACTTTGGAAG	MM245 cgatcctcgagAAAACGACCAGATAACAAGTAACCAG
MI46	MI-Dac	MM116 CGATAACTTGCAAGTTGGTGC	MM186 ggaattccatATGACCGCTAAAAACCATTTTCG	MM246 cgatcctcgagGTACTTACCGGATAACAAAGCACC
MI47	MI-Bbr	MM117 CTTGGAAGTCGGTCGTCATG	MM187 ggaattccatATGCAAAAGACTTACCGTATCGG	MM247 cgatcctcgagATAGGCACCAGACAATAAGGCA
MI48	MI-Cca	MM118 CTTGGCTGTCGCTAGACAC	MM188 ggaattccatATGACTCGTATCTACAGAATCGGC	MM248 cgatcctcgagCCAAGCACCAGACAATAAGGC
MI49	MI-Aph	MM119 CTTGCCAGCTTTGGCTAGATC	MM189 ggaattccatATGTCTACTCCAGCCATCACTT	MM249 cgatcctcgagAGCTTTGGCTGGGGCTAA
MI50	MI-Ama1	MM120 GTGACCGTGTCATGGCTG	MM190 ggaattccatATGCGTACCTACCGTATTGGT	MM250 cgatcctcgagTCTGACACGGTGACCTGG
MI51	MI-Ama2	MM121 GGTTGTATCCCAGGTGATAGAGT	MM191 ggaattccatATGCATCACCGTATTGGTTTGAT	MM251 cgatcctcgagGACGACAGCTCTACGAGAACC
MI52	MI-Pas	MM122 GTTGTATTCCAGGTGACAGAGTC	MM192 ggaattccatATGACCGGTGCTCCACG	MM252 cgatcctcgagAGCACCAACTAAAGCGTCAC
MI53	MI-meB	MM123 CATTGCCTCCGACTTGGAC	MM193 ggaattccatATGAAATTAACCAAAATTGAACCAAAATATGTTAAG	MM253 cgatcctcgagAGAATTGAACAATTGACCGAAACCT
MI54	MI-meC	MM124 GAAGCCGATTACGATATTGGTAAGG	MM194 ggaattccatATGAACTCTACTAAGATTGACCCAAAGT	MM254 cgatcctcgagATTTTCGTTGAACAACTTACCAAAACC
MI55	MI-meD	MM125 GGTTTCGAAATCACTTCCAACTCC	MM195 ggaattccatATGGAAACTAAGAAGATTCCACCAAAG	MM255 cgatcctcgagGTTGGAGAACAACTTACCGAAACC

Lower case bases denote the sequence with RE site added up- or down-stream of the MI ORFs.
Internal check primers anneal approximately 225bp from the 3´ end of the ORF.
Table is continued on next page.

Table A2 (cont'd): Primers used for plasmid construction with MI ORFs.

Plasmid	Primer	Anneal location	Sequence
pET41b	MM067	GST-tag internal	CCAGCAAGTATATAGCATGGCC
	GC1873	KanR internal	GCAGTTTCATTTGATGCTCGATGAG
	T7 (Eurofins standard sequencing primer)	upstream inserted ORF	TAATACGACTCACTATAGGG
	T7term (Eurofins standard sequencing primer)	downstream inserted ORF	CTAGTTATTGCTCAGCGGT
pTRC99a	MM293	AmpR internal	GAGATCCAGTTCGATGTAACCC
	MM294	trc-promoter	GCCGACATCATAACGGTTCT

Table A3: Primers used to confirm KOs in strain construction.

Primer	Locus	Sequence
MM275	frdA 500bp upstream	CGTGTCTCAAACGGGACC
MM276	frdA 500bp downstream	GGACCGATGAACTCTGGGT
MM277	gdhA 500bp upstream	CCAGACCAGCGGATTAGTG
MM278	gdhA 500bp downstream	CTCAGACTATATCTCTGATTGCGC
MM279	maeB 500bp upstream	CCTGTTGTTCCTGGGTCTTG
MM280	maeB 500bp downstream	CCGCTGAATCGGTCGAG
MM281	pykA 500bp upstream	GCAACGCATGAGTTGTATGAATTG
MM282	pykA 500bp downstream	CCATTCATCCAGTCGGTACG
MM283	pykF 500bp upstream	GCCAACTATCAGCATATATCAATCTAACC
MM284	pykF 500bp downstream	CCATGTTGTCCAGACGCTG
MM285	sfcA 500bp upstream	GGTTGAAGCTTCTGATAACGACA
MM286	sfcA 500bp downstream	CCACGATGCACTCTTCCG
MM287	KanR Internal, 300bp from 3´ end	GAAGCCGGTCTTGTCGATC

Table A4: *Nic*2 cluster survey of coding sequences within ±4-ORF positions of putative MIs.

MI ID	Annotation	-4	-3	-2	-1	+1	+2	+3	+4	Information source
MI-PpuS	Maleate Isomerase	NA	orf4	Hydroxy-succinoylpyridine hydroxylase	hyp orf	N-formylmaleamate deformylase	Dihydroxy pyridine dioxygenase	maleamate amidase	cytochrome c-type biogenesis protein CcmC	[14]
MI-Rjo	possible maleate cis-trans isomerase	monooxygenase/aromatic ringhydroxylase	probable 2,3-dihydroxybenzoate-AMP ligase	hydantoin utilization protein B	hydantoin utilization protein A	possible glyoxalase	hyp orf	hyp orf	possible phthalate 4,5-dioxygenase	CP000431.1
MI-Nfa	maleate cis-trans isomerase	putative phosphoesterase	hyp orf	hyp orf	hyp orf	putative hydantoinase	putative hydantoinase	putative acyl-CoA synthetase	putative oxidoreductase	AP006618.1
MI-Bst	maleate cis-trans isomerase	hydroxyglutarate oxidase	hyp orf	acetamidase	GntR family transcriptional regulator	hydroxymethylglutaryl-CoA lyase	carnitine dehydratase	MFS transporter	hyp orf	
MI-Sma	maleate cis-trans isomerase	MarR family transcriptional regulator	2,5-dihydroxypyridine 5,6-dioxygenase	6-hydroxynicotinate 3-monooxygenase	alpha/beta hydrolase	isochorismatase	MFS transporter	hyp orf (Blast result: only hyp prots)	dipeptidyl carboxypeptidase II	NZ_JPQY01000040.1
MI-Afa	maleate cis-trans isomerase	nitrate ABC transporter permease	ABC transporter	4-hydroxy-2-oxo-heptane-1,7-dioate aldolase	LysR family transcriptional regulator	UDP-glucose 4-epimerase	gamma carbonic anhydrase family protein	ribosomal protein S12 methylthiotransferase	16S rRNA (cytosine(967)-C(5))-methyltransferase	NZ_JQCV01000003.1
MI-Pfl	maleate cis-trans isomerase	putative MarR-family transcriptional regulator	conserved hypothetical protein (99% ident, 100% cov to 2,5-dihydroxypyridine 5,6-dioxygenase [Pseudomonas sp. OV546])	putative monooxygenase	putative alpha/beta hydrolase	putative isochorismatase	putative integral membrane protein	putative aminotransferase	putative molybdenum-pterin binding protein II	AM181176.4
MI-Rso	maleate cis-trans isomerase	bifunctional hydroxylase/ oxidoreductase	NADPH dehydrogenase	Fumarylacetoacetate hydrolase	aliphatic nitrilase	transcription regulator protein	Asp/Glu/hydantoin racemase	sugar-binding protein	methionine--tRNA ligase	NC_020799.1
MI-R.G	maleate cis-trans isomerase	CO dehydrogenase	amidase-like nicotinamidase	M29 family peptidase	alpha/beta fold family hydrolase/ acetyltransferase	amidohydrolasefamily protein	MFS transporter	ABC transporter substrate-binding protein	ABC transporter ATP binding protein	NZ_KB944462.1
MI-Sso	maleate cis-trans isomerase	5-oxoprolinase	5-oxoprolinase	hyp orf	MFS transporter	MBL fold metallo-hydrolase	2-hydroxyhepta-2,4-diene-1,7-dioate isomerase	aldehyde oxidase	transposase	NC_002754.1
MI-Cne	maleate cis-trans isomerase	metabolite transport protein	alcohol dehydrogenase cytochrome c subunit	isoquinoline 1-oxidoreductase subunit alpha	6-hydroxynicotinate 3-monooxygenase	hydrolase or acyltransferase alpha/beta hydrolase superfamily	hyp orf	N-carbamoylsarcosine amidase	transcriptional regulator TetR/AcrR family	CP002877.1
MI-Cba1	Maleate cis-trans isomerase	Nitrate/nitrite transporter	Putative diheme cytochrome c-553	Isoquinoline 1-oxidoreductase alpha subunit	Salicylate hydroxylase	Hydrolase, alpha/beta fold familyfunctionally coupled to Phosphoribulokinase	Leucyl aminopeptidase	N-carbamoylsarcosine amidase	Transcriptional regulator, TetR family	CP010537.1
MI-Cba2	Maleate cis-trans isomerase	Branched-chain amino acid transport ATP-binding potein	Leu-, isoleu-, val-, thr-, and ala-binding protein	Permeases of the major facilitator superfamily	2-amino-3-carboxymuconate-6-semialdehyde decarboxylase	Hydrolase, alpha/beta fold familyfunctionally coupled to Phosphoribulokinase	Leucyl aminopeptidase	N-carbamoylsarcosine amidase	CO dehydrogenase large chain paralog usually without motifs	CP010537.1
MI-Pha	Maleate cis-trans isomerase	ABC-type nitrate/ sulfonate/ bicarbonate transport system, permease component	ABC-transporter, ATP-binding component	ABC transporter protein	Salicylate hydroxylase	hydrolase, alpha/beta fold family	Leucyl aminopeptidase	N-carbamoylsarcosine amidase	electron transfer flavoprotein, beta subunit	CP003075.1
MI-P.U	Maleate isomerase	Salicylate hydroxylase	Isochorismatase family hydrolase	Leucyl aminopeptidase	Hydrolase, alpha/beta fold family protein	TetR family transcriptional regulator	Integral membrane protein	Tripartite tricarboxylate transporter TctB family protein	Extracytoplasmic binding receptor	JSUQ01000002.1
MI-Cbae	maleate cis-trans isomerase	Aerobic-type carbon monoxide dehydrogenase, large subunit CoxL/CutL homologs	aerobic carbon monoxide dehydrogenase	transcriptional regulator	TetR family transcriptional regulator	hydrolase, alpha/beta fold family protein	Leucyl aminopeptidase	isochorismatase family hydrolase	feruloyl esterase	AMRK01000010.1
MI-Cin	maleate cis-trans isomerase	Conserved protein/domain typically associated with flavoprotein oxygenases, DIM6/NTAB family	Isochorismatase family hydrolase	leucyl aminopeptidase	hydrolase, alpha/beta fold family protein	TetR family transcriptional regulator	methyl-accepting chemotaxis protein	acetate--CoA ligase	solute:sodium symporter, small subunit domain containing protein	CP004393.1
MI-P.C	maleate cis-trans isomerase	MarR family transcriptional regulator	hyp orf	monooxygenase	alpha/beta hydrolase fold protein	isochorismatase	MFS transporter	hyp orf	shikimate 5-dehydrogenase	NZ_ATLR01000025.1
MI-PpuN	maleate cis-trans isomerase	putative MarR family transcriptional regulator	hyp orf	putative monooxygenase	putative hydrolase	isochorismatase family protein	putative major facilitator superfamilytransporter	putative porin	hyp orf	AP013070.1
MI-Aac	hypothetical protein	hyp orf	hyp orf	hyp orf	hyp orf	hyp orf	hyp orf	hyp orf	hyp orf	AURB01000151.1

MI ID	Annotation	-4	-3	-2	-1	+1	+2	+3	+4	Information source
MI-Msm	maleate cis-trans isomerase	hyp orf (99% ident 100% cov to carbamoylsarcosine amidase [Mycobacterium smegmatis])	putative salicylate 1-monooxygenase	marR-family transcriptional regulator	peptidase, M29 family	oligopeptide transporter, OPT family	hyp orf (73% ident, 100%cov to Protein of uncharacterised function (DUF2510) [Mycobacterium abscessus subsp. abscessus])	alkaline phosphatase	transporter, small conductance mechanosensitive ion channel family protein	AOCJ01000026.1
MI-S.N1	maleate cis-trans isomerase	MFS transporter	alpha/beta hydrolase family protein	hyp orf (63% ident, 94%cov to LOW QUALITY PROTEIN: glyoxalase/ bleomycin resistance protein/ dioxygenase [Rhodococcus opacus PD630])	glyoxalase	hyp orf (47% ident, 97%cov to oxidoreductase [Amycolatopsis sp. ATCC 39116])	hyp orf (57% ident 99%cov to 2,3-dihydroxybenzoate-AMP ligase [Pseudonocardia ammonioxydans])	end contig	end contig	NZ_JOIQ01000010.1
MI-S.N2	maleate cis-trans isomerase	oxidoreductase	acyl-CoA synthetase	hydantoinase	hydantoinase	glyoxalase	hyp orf	alpha/beta hydrolase family protein	hyp orf	NZ_JOIQ01000011.1
MI-S.S	maleate cis-trans isomerase	NAD(P)-dependent alcohol dehydrogenase	hyp orf	hyp orf	TetR family transcriptional regulator	glyoxalase	hyp orf	hyp orf	glycoside hydrolase family 2	NZ_JALM01000100.1
MI-Pdi	maleate cis-trans isomerase	5-oxoprolinase (ATP-hydrolyzing)	5-oxoprolinase (ATP-hydrolyzing)	monooxygenase FAD-binding protein	(2,3-dihydroxybenzoyl) adenylate synthase	Glyoxalase/bleomycin resistance protein/ dioxgenase	hyp orf	hyp orf	Integrase catalytic region	CP002593.1
MI-Abe	hypothetical protein	alpha/beta hydrolase	TetR family transcriptional regulator	hyp orf	hyp orf	hyp orf	hyp orf	alpha/beta hydrolase	glyoxalase	NZ_KB912942.1
MI-Ach	maleate cis-trans isomerase	hydantoinase	hydantoinase	hyp orf	hyp orf	hyp orf	hyp orf	hyp orf	hyp orf	NZ_KB903220.1
MI-Aor	maleate cis-trans isomerase	hyp orf (33%ident 90%cov to High-affinity carbon uptake protein Hat/HatR [Labilithrix luteola])	hyp orf (73%ident, 99%cov to alpha/beta hydrolase family protein [Rhodococcus opacus])	hyp orf (68% ident, 97%cov to LOW QUALITY PROTEIN: glyoxalase/ bleomycin resistance protein/ dioxygenase [Rhodococcus opacus PD630])	glyoxalase	hydantoinase	hydantoinase	acyl-CoA synthetase	oxidoreductase	NZ_ASXG01000196.1
MI-Grh	maleate cis-trans isomerase	putative oxidoreductase	putative 2,3-dihydroxybenzoate-AMP ligase	hydantoinase B	hydantoinase A	glyoxylase	hyp orf (70% ident, 98%cov to LOW QUALITY PROTEIN: glyoxalase/ bleomycin resistance protein/ dioxygenase [Rhodococcus opacus PD630])	hyp orf (77% ident 100%cov to alpha/beta hydrolase [Rhodococcus opacus])	sporulation protein	BAHC01000115.1
MI-Gar	maleate cis-trans isomerase	2-oxobutyrate oxidase	hyp orf (79%ident, 99%cov to alpha/beta hydrolase family protein [Tsukamurella pseudospumae])	hyp orf (62% ident, 97%cov to LOW QUALITY PROTEIN: glyoxalase/ bleomycin resistance protein/ dioxygenase [Rhodococcus opacus PD630])	glyoxalase	hydantoinase	hydantoinase	acyl-CoA synthetase	oxidoreductase	NZ_BAEE01000078.1
MI-Gkr	maleate cis-trans isomerase	hyp orf	TetR family transcriptional regulator	hyp orf	short-chain dehydrogenase	hyp orf	hyp orf	alpha/beta hydrolase family protein	hyp orf	NZ_AQYG01000049.1
MI-Nno	maleate isomerase	putative monooxygenase	putative AMP-binding ligase	putative hydantoin utilization protein B	putative hydantoin utilization protein A	putative glyoxalase	hyp orf	putative hydrolase, alpha/beta hydrolase family	gentisate 1 2-dioxy genase-like protein	CP006850.1
MI-Nbr	maleate cis-trans isomerase	end contig	end contig	end contig	end contig	hydantoinase	end contig	end contig	end contig	NZ_BAFU01000047.1
MI-W.D	maleate cis-trans isomerase	hyp orf	hyp orf	TetR family transcriptional regulator	oxidoreductase	glyoxalase	hyp orf (73% ident, 98%cov to LOW QUALITY PROTEIN: glyoxalase/ bleomycin resistance protein/ dioxygenase [Rhodococcus opacus PD630])	hyp orf (82% ident, 97% cov to alpha/beta hydrolase [Gordonia sp. IITR100])	ATPase	AYTE01000027.1
MI-Rop	ectoine utilization protein EutA	phthalate 4,5-dioxygenase	aromatic acid:H symporter	hyp orf	Glyoxalase/bleomycin resistance protein/ dioxgenase	hydantoin utilization protein A	hydantoin utilization protein B	2,3-dihydroxybenzoate-AMP ligase	monooxygenase/aromatic ring hydroxylase	JH377109.1
MI-Rfa	maleate cis-trans isomerase	oxidoreductase	acyl-CoA synthetase	hydantoinase	hydantoinase	glyoxalase	hyp orf	alpha/beta hydrolase family protein	MFS transporter	NZ_JMEZ01000015.1
MI-Rwr	maleate cis-trans isomerase	end contig	hyp orf	hyp orf	glyoxalase	hydantoin utilization protein A	hydantoin utilization protein B	2,3-dihydroxybenzoate-AMP ligase	monooxygenase/aromatic ring hydroxylase	ANIU01000508.1

MI ID	Annotation	-4	-3	-2	-1	+1	+2	+3	+4	Information source
MI-Car	maleate cis-trans isomerase	Zn-dependent oxidoreductase, NADPH:quinone reductase	N-methylhydantoinase B/acetone carboxylase, alpha subunit	N-methylhydantoinase A/acetone carboxylase, beta subunit	transcriptional regulator	lactoylglutathione lyase-like lyase	hyp orf	Alpha/beta hydrolase family	acyl dehydratase	KK073874.1
MI-T.1	maleate cis-trans isomerase	hyp orf	hyp orf	hyp orf	TetR family transcriptional regulator	glyoxalase	hyp orf	hyp orf	hyp orf	NZ_HE997626.1
MI-Rpi	maleate cis-trans isomerase	hyp orf	hyp orf	transcriptional regulator	arabinose efflux permease family protein	hyp orf	hyp orf	major Facilitator Superfamily protein	cytochrome bo3 quinol oxidase subunit 2	APMQ01000001.1
MI-Axy	maleate cis-trans isomerase	3-oxoadipate CoA-transferase subunit B	3-oxoadipate CoA-transferase subunit A	hyp orf	LysR family transcriptional regulator	transcriptional regulator, LysR family	LysR family transcriptional regulator	Thioredoxin reductase	thioredoxin reductase	NC_021285.1
MI-Agu	maleate cis-trans isomerase	putative nitroreductase	hyp orf	putative major facilitator superfamily transporter	isochorismatase family protein	putative hydrolase	putative monooxygenase	hyp orf	putative MarR family transcriptional regulator	AP014630.1
MI-H.C	maleate cis-trans isomerase	aerobic-type carbon monoxide dehydrogenase large subunit CoxL/CutL-like protein	nicotinamidase-like amidase	hyp orf	putative hydrolaseor acyltransferase of alpha/ beta superfamily	2-polyprenyl-6-methoxyphenol hydroxylase-like oxidoreductase	aerobic-type carbon monoxide dehydrogenase small subunit CoxS/CutS-like protein	aerobic-type carbon monoxide dehydrogenase large subunit CoxL/CutL-like protein	PAS domain S-box	AKJW01000018.1
MI-B.R	maleate cis-trans isomerase	putative MarR-family transcriptional regulator	gluconate 2-dehydrogenase	isoquinoline 1-oxidoreductase alpha subunit	monooxygenase FAD-binding	alpha/beta hydrolase fold protein	hyp orf	putative isochorismatase	MFS transporter	NC_021289.1
MI-P.B	maleate cis-trans isomerase	LysR family transcriptional regulator	putative lipoprotein	AraC family transcriptional regulator	kinase inhibitor	LysR family transcriptional regulator	Surfeit locus 4-related protein	glutathione S-transferase domain-containing protein	HxlR family transcriptional regulator	AZSV01000017.1
MI-Dac	maleate isomerase	hyp orf	hyp orf	hyp orf	hyp orf	peptidase M23B	type III pantothenate kinase	type II secretion system protein E	general secretion pathway protein F	NZ_KE150371.1
MI-Bbr	maleate cis-trans isomerase	marR family transcriptional regulator	hyp orf	hyp orf	hyp orf	alpha/beta hydrolase	hyp orf	isochorismatase	ring hydroxylating protein subunit beta	NC_019382.1
MI-Cca	PREDICTED: maleate isomerase-like	HTH-type transcriptional repressor NicR-like	2,5-dihydroxypyridine 5,6-dioxygenase-like	6-hydroxynicotinate 3-monooxygenase-like	N-formylmaleamate deformylase-like	maleamate amidohydrolase-like	putative metabolite transport protein NicT-like	uncharacterized oxidoreductase YfjR-like	vibriobactin receptor-like	NW_004523891.1
MI-Aph	maleate cis-trans isomerase	uncharacterized conserved protein	amidase, Asp-tRNAAsn/ Glu-tRNAGln amidotransferase A subunit	arabinose efflux permease family protein	transcriptional regulator	transporter, UIT1 family	conserved protein of DIM6/NTAB family	2-haloalkanoic acid dehalogenase, type II	transcriptional activator of acetoin/glycerol metabolism	CP002379.1
MI-Ama1	Maleate cis-trans isomerase	transcriptional regulator	hyp orf	Cytochrome P450	TetR family transcriptional regulator	glyoxalase	hyp orf	hyp orf	hydantoinase	NZ_AWOO02000009.1
MI-Ama2	Maleate cis-trans isomerase	hyp orf	hyp orf	hyp orf	hyp orf	glyoxalase	hyp orf	hyp orf	hyp orf	NZ_AWOO02000009.1
MI-Pas	Maleate cis-trans isomerase	hyp orf (73%ident, 99%cov 2,3-dihydroxybenzoate-AMP ligase [Pseudonocardia ammonioxydans])	oxidoreductase	glyoxalase	hyp orf (66% ident, 90%cov to LOW QUALITY PROTEIN: glyoxalase/ bleomycin resistance protein/ dioxygenase [Rhodococcus opacus PD630])	hyp orf (52% ident, 95%cov to alpha/beta hydrolase [Mycobacterium indicus pranii])	transposase	site-specific integrase	hyp orf (38% ident, 69%cov to IrrE-like protein [Mycobacterium phage Sbash] IrrE is a novel transcriptional regulator)	NZ_AUII01000013.1
MI-meB	hypothetical protein	end contig	end contig	end contig	end contig	hyp orf	end contig	end contig	end contig	EP526545.1
MI-meC	hypothetical protein	end contig	end contig	end contig	end contig	hyp orf	hyp orf	end contig	end contig	EM554172.1
MI-meD	hypothetical protein	end contig	end contig	end contig	hyp orf	hyp orf	end contig	end contig	end contig	EP608503.1

Abbreviations: hyp orf: hypothetical ORF, ident: identity, cov: coverage.

The 4 nic2 cluster members of PpuS were assigned colors: HSP: magenta, NFO: blue, HPO: purple, and AMI: aqua. The same colors appear throughout the table for annotations which either do or may match the 4 cluster genes. Broad enzyme family annotated coding sequences were considered to be potential cluster members if the activities described were similar,

Coding sequences annotated as 'hyp orf' were used as BLASTp queries for further clarification, hits are included in the table if there were any with annotations. Information was verified in April 2017.

Table A5: Characteristics of the species from which 55 putative MIs originate.

ID	Family	Order	Class	Phylum	Isolation	Location	Environment Note	Other Note(s)	S	P	C	Reference(s)
MI-PpuS	Pseudomonadaceae	Pseudomonadales	G	P	soil	China	Isolated from soil samples obtained from a field under continuous tobacco cropping in Shandong, China.	Nicotine-degrading	0	0	1	PMID:21914868
MI-Rjo	Nocardiaceae	Corynebacteriales	Ac	Ac	soil	Canada	Isolated from lindane contaminated soil in BC.	Methylotroph, Potent polychlorinated biphenyl(PCB)-degrading soil Actinomycete that catabolizes a wide range of compounds, high GC Gram+	0	0	1	PMID:17030794
MI-Nfa	Nocardiaceae	Corynebacteriales	Ac	Ac	mammal	Japan	Isolated from the bronchus of a 68-year-old male Japanese patient.	high GC Gram+	0	1	0	PMID:15466710
MI-Bst	Bacillaceae	Bacillales	Ba	F	soil	Japan	Soil samples collected from nature	Gram+, rod, aerobic, mobile	0	0	0	https://www.google.ch/patents/US5783428
MI-Sma	Yersiniaceae	Enterobacterales	G	P	soil	Japan	unspecified soil sample	rod, Gram-, HAIs, motile, 5-40°, pH5-9	0	1	0	[25]
MI-Afa	Alcaligenaceae	Burkholderiales	B	P	NA	NA	nbrc13111: no source listed	rod, Gram-, soil, water, humans	0	0	0	http://www.nbrc.nite.go.jp/NBRC2/NBRCCatalogueDetailServlet?ID=NBRC&CAT=00013111
MI-Pfl	Pseudomonadaceae	Pseudomonadales	G	P	leaf	UK	Isolated in 1989 from the leaf surface of a sugar beet plant grown at the University Farm, Wytham, Oxford, UK.	*P. fluorescens* are common soil bacteria that can improve plant health through nutrient cycling, pathogen antagonism and induction of plant defenses.	1	1	0	PMID:19432983
MI-Rso	Burkholderiaceae	Burkholderiales	B	P	soil	China	Isolated from a bacterial wilt nursery built for breeding crop resistance in 1997 in the Guizhou province in southwest China		0	1	0	PMID:23661471
MI-R.G	Burkholderiaceae	Burkholderiales	B	P	soil	Australia	isolated from suburban soil in Canberra, Australia.	hexachlorocyclohexane (HCH)-degrading bacterial strain	0	0	1	PUBMED 23833131
MI-Sso	Sulfolobaceae	Sulfolobales	T	C	NA	NA	Grows optimally at 80 deg C and pH 2- to 4-metabolizing sulfur.	Carbon fixation. Aerobic crenarchaeon. ARCHEA	0	0	0	PMID:11427726
MI-Cne	Burkholderiaceae	Burkholderiales	B	P	soil	USA	isolated from soil in the vicinity of University Park, PA, USA	Copper-resistant bacterium. Gram-	0	0	1	PMID:21742890 http://www.uniprot.org/Proteo-mes/UP000006798
MI-Cba1	Burkholderiaceae	Burkholderiales	B	P	ground water	USA	Isolated from groundwater at the Oak Ridge Field Research Center site.	Diverse metabolic capabilities, heavy-metal resistance, biodegradation capabilities	0	0	1	PMID:25977418
MI-Cba2	Burkholderiaceae	Burkholderiales	B	P	ground water	USA	Isolated from groundwater at the Oak Ridge Field Research Center site.	Diverse metabolic capabilities, heavy-metal resistance, biodegradation capabilities	0	0	1	PMID:25977418
MI-Pha	Hyphomicrobiaceae	Rhizobiales	A	P	seawater	China	Isolated from seawater sample (depth of 70m) in East China Sea.	marine halotolerant bacterium	0	0	0	PMID:22156395
MI-P.U	Rhodobacteraceae	Rhodobacterales	A	P	algae	Malaysia	isolated from clonal culture of toxic dinoflagellate *Alexandrium tamiyavanichii*, Malaysia: Sebatu, Malacca	rod, Gram-, non-motile	0	0	1	PMID: 19661508
MI-Cbae	Rhodobacteraceae	Rhodobacterales	A	P	ocean	Arctic	Arctic ocean, deep-sea sediment		0	0	0	SAMN02471168
MI-Cin	Rhodobacteraceae	Rhodobacterales	A	P	ocean	India	Isolated from deep-sea sediment from the Indian Ocean.	polycyclic aromatichydrocarbon-degrading bacterium, fluoranthene degradation pathway	0	0	1	PMID:25582347
MI-P.C	Pseudomonadaceae	Pseudomonadales	G	P	river	USA	isolated from the hyporheic zone of the highly contaminated Clark Fork river in Montana for studies of tolerance to cadmium exposure	cadmium exposure	0	0	1	PRJNA59591
MI-PpuN	Pseudomonadaceae	Pseudomonadales	G	P	NA	NA	Cultivate 30°C	*P.putida* has attracted much interest for its environmental, industrial, biotechnological, and clinical importance	0	0	0	http://www.nbrc.nite.go.jp/NBRC2/NBRCCatalogueDetailServlet?ID=NBRC&CAT=00014164
MI-Aac	Alicyclobacillaceae	Bacillales	Ba	F	soil	Germany	nonpathogenic bacterium which contaminates commercial pasteurized fruit juices	spore-forming Gram+, Thermoacidophilic	0	0	0	PMID: 24009113
MI-Msm	Mycobacteriaceae	Corynebacteriales	Ac	Ac	free	Japan	free living, Japan	Spontaneous streptomycin-resistant subclone of *M. smegmatis* mc2874, acts as recipient during conjugation with *M. smegmatis* mc(2)155,	0	0	0	PMID: 23618714
MI-S.N1	Streptomycetaceae	Streptomycetales	Ac	Ac	NA	NA	unknown source	high GC Gram+	0	0	0	SAMN02645511
MI-S.N2	Streptomycetaceae	Streptomycetales	Ac	Ac	NA	NA	unknown source	high GC Gram+	0	1	0	SAMN02645512
MI-S.S	Pseudonocardiaceae	Pseudonocardiales	Ac	Ac	sediment	Malaysia		high GC Gram+ , antibacterial marine Actinobacteria with Polyketide Synthase and Peptide Synthase Genes	0	0	0	PRJNA233984
MI-Pdi	Pseudonocardiaceae	Pseudonocardiales	Ac	Ac	sludge	USA	industrial sludge contaminated with 1,4-dioxane	high GC Gram+	0	0	1	PMID:21725009
MI-Abe	Pseudonocardiaceae	Pseudonocardiales	Ac	Ac	mammal	NA	clinical isolate from submandibular mycetoma tissue	high GC Gram+, degrades aromatic compounds such as m-hydroxybenzoate, does not produce antibiotics	0	1	0	PMID: 16403887
MI-Ach	Pseudonocardiaceae	Pseudonocardiales	Ac	Ac	soil	Thailand	isolated from soil of a tropical rainforest in N Thailand,	high GC Gram+	0	0	0	PMID: 18218940
MI-Aor	Pseudonocardiaceae	Pseudonocardiales	Ac	Ac	NA	NA		high GC Gram+, vancomycin-producing	0	0	0	PMID:24884615

ID	Family	Order	Class	Phylum	Isolation	Location	Environment Note	Other Note(s)	S	P	C	Reference(s)
MI-Grh	Gordoniaceae	Corynebacteriales	Ac	Ac	root	Japan		high GC Gram+ , industrial relevance	0	0	0	PRJDB4
MI-Gar	Gordoniaceae	Corynebacteriales	Ac	Ac	mammal	Japan	isolated from a clinical specimen (human sputum)	high GC Gram+	0	1	0	PMID 16902014
MI-Gkr	Gordoniaceae	Corynebacteriales	Ac	Ac	stream	SKorea	isolated from a polluted stream	high GC Gram+ , phenol-degrading Actinomycete	0	0	1	PMID 19567568
MI-Nno	Nocardiaceae	Corynebacteriales	Ac	Ac	root	Brazil	isolated from a root of Couma macrocarpa in Brazil	high GC Gram+, degradation of rubber and gutta-percha	0	1	0	PRJNA225667/PMID: 24747905
MI-Nbr	Nocardiaceae	Corynebacteriales	Ac	Ac	mammal	NA	isolated from human sputum	high GC Gram+	0	1	0	https://www.atcc.org/products/all/15333.aspx
MI-W.D	Williamsiaceae	Corynebacteriales	Ac	Ac	soil	Antarctic	soil, Antarctica: Darwin Mountains	high GC Gram+, psychrotolerant	0	0	0	PMID: 24459282
MI-Rop	Nocardiaceae	Corynebacteriales	Ac	Ac	soil	NA		high GC Gram+, chemo-heterotrophic. Biofuels interest: model of prokaryotic oleaginy	0	0	1	PMID: 21931557
MI-Rfa	Nocardiaceae	Corynebacteriales	Ac	Ac	plant	USA	colonization of tobacco tissues	phytopathogen, high GC Gram+	0	1	1	PMID: 11332724
MI-Rwr	Nocardiaceae	Corynebacteriales	Ac	Ac	NA	France	isolated from a microbial consortium that degraded 15 petroleum compounds or additives	high GC Gram+ , degradation of a mixture of hydrocarbons, gasoline, and diesel oil additives	0	0	1	PRJNA167644
MI-Car	Cryptosporangiaceae	Frankiales	Ac	Ac	soil	Japan	Cultivated soil from a vegetable field in Japan	high GC Gram+	0	0	0	EXG81230.1
MI-T.1	Tsukamurellaceae	Corynebacteriales	Ac	Ac	mammal	NA	isolated from a human sputum specimen	high GC Gram+	0	1	0	PMCID: PMC3457207
MI-Rpi	Burkholderiaceae	Burkholderiales	B	P	sediment	USA	isolated from subsurface sediments in Oak Ridge, TN.	A member of this genus is a potential bioremediation agent. Strain OR214 is tolerant to various heavy metals, such as uranium, nickel, cobalt, and cadmium.	0	0	1	PMID: 23792748
MI-Axy	Alcaligenaceae	Burkholderiales	B	P	mammal	Denmark	Isolated from sputum from a cystic fibrosis (CF) patient (first time for this patient) of the Copenhagen CF Centre	b-Proteobacteria, environmental opportunistic pathogen	0	1	0	PMID:23894309
MI-Agu	Moraxellaceae	Pseudomonadales	G	P	soil	Japan		g-Proteobacteria, utilize DMS as a sole sulfur source and degrade chloroethylenes	0	0	1	PMID: 25323718
MI-H.C	Oxalobacteraceae	Burkholderiales	B	P	root	USA	Isolated from Rhizosphere and Endosphere of *Populus deltoides*, TN	b-Proteobacteria	0	0	0	PMID: 23045501
MI-B.R	Burkholderiaceae	Burkholderiales	B	P	insect	Japan	Symbiont of the Bean Bug *Riptortus pedestris*	b-Proteobacteria, use to develop pest control strategies	1	0	0	PMID: 23833137
MI-P.B	Pseudomonadaceae	Pseudomonadales	G	P	soil	Hungary		g-Proteobacteria , sequencing for use in groundwater during bioremediation	0	0	0	PRJNA232351
MI-Dac	Comamonadaceae	Burkholderiales	B	P	mammal	NA	Human Microbiome Project	Gram-, aerobe, non spore forming, ability to produce gold nuggets with *Cupriavidus metallidurans*	0	0	0	PRJNA52169
MI-Bbr	Alcaligenaceae	Burkholderiales	B	P	mammal	USA	Isolated from dog in the USA	Gram-negative respiratory pathogen that infects a wide range of hosts	0	1	0	PMCID: PMC2493278
MI-Cca	Enterobacteriaceae	Enterobacterales	G	P	insect	Italy	DNA provided by the laboratory of Dr. G. Gasperi of Insect Molecular Biology, University of Pavia, Italy	Entero-bacteriaceae Genome sequencing and assembly	0	0	0	PRJNA285085/PMID:25502672/SAMN02953830
MI-Aph	Micrococcaceae	Micrococcales	Ac	Ac	soil	Greece	Isolated from Perivleptos, a creosote polluted site in Epirus, Greece where a wood preserving industry was operating for > 30yrs	high GC Gram+	0	0	1	PMID:21677849
MI-Ama1	Thermomonosporaceae	Streptosporangiales	Ac	Ac	mammal	Mexico		high GC Gram+ , non-motile, chemo-organotrophic, aerobic Actinomycete can infect immunosuupressant patients	0	1	0	PMCID: PMC265203
MI-Ama2	Thermomonosporaceae	Streptosporangiales	Ac	Ac	mammal	Mexico		high GC Gram+ , non-motile, chemo-organotrophic, aerobic Actinomycete can infect immunosuppressant patients	0	1	0	PMCID: PMC265203
MI-Pas	Pseudonocardiaceae	Pseudonocardiales	Ac	Ac	animal	NA		high GC Gram+ , dimethyl disulfide-degrading Actinomycete	0	0	1	PMID 9731282
MI-meB	Pelagibacteraceae	Pelagibacterales	A	P	marine	Pacific	part of large dataset isolated from surface water samples collected from the Eastern N American coast to the Eastern Pacific Ocean		0	0	0	PRJNA13694
MI-meC	Pelagibacteraceae	Pelagibacterales	A	P	marine	Pacific	same as previous sample		0	0	0	PRJNA13695
MI-meD	Pelagibacteraceae	Pelagibacterales	A	P	marine	Pacific	same as previous sample		0	0	0	PRJNA13696

Abbreviations: Headers: S: symbiont, P: pathogens, C: contaminated environmental sites. 1/0: yes/no. Class/Phylum: A: Alphaproteobacteria, Ac: Actinobacteria, B: Betaproteobacteria, Ba: Bacilli, C: Crenarcheaota, F: Firmicutes, G: Gammaproteobacteria, P: Proteobacteria, T: Thermoprotei.

Table A6: Unique residue replacements in the MI library

Residue #	Domain	Common residue	Unique residue	Appears in MI variant
-	β1	N	M	Aph
-	β1	N	I	Rjo
8	β1	R/K	H	Msm
16	β1-α1	S	T	Gar
28	α1	L	F	Aph
30	α1	-/A	R	Aph
30	α1	-/A	S	Bst
30	α1	-/A	L	Cba1
30	α1	-/A	W	Aac
31	α1	-/R	Q	Aph
-	α1	-	E	Msm
-	α1	-	R	P.B
39	β2	F	Y	Aph
41	β2	F	Y	Cin
46	β2	M	A	Aph
47	β2-α2 breathing loop	R	G	Aph
47	β2-α2 breathing loop	R	Q	Pas
65	α2	R	D	Aph
69	α2	E	A	Aph
80	β3	Y	S	Msm
92	α3-α4	G	K	Msm
140	β5-α6	A	F	P.B
155	α6-β6	G	A	P.B
155	α6-β6	G	D	W.D
213	α9	E	Q	Bst
217	α9-α10	G	E	Rwr
217	α9-α10	G	R	PpuS
217	α9-α10	G	A	P.B
220	α9-α10	V	I	Gkr
223	α10	A	S	Rso
224	α10	A	S	Rso
239	α10	L	M	Agu
235	NA	L	H	Aph

Unique residue replacements are listed for the MI library with members of phylogenetic cluster VII excluded. Residue numbering and domains are as described for PpuS16 [13]. Residue types are denoted by colors as described in Figure 13. The cases in which there was no residue type change are shaded in grey.

Table A7: Sequence alignment of breathing loops for MI library.

					B2-A2 LOOP							B6-A7 LOOP				
Cluster	Library #:	ID	Forward active?	Reverse active?	47	48	49	50	-	51	52	164	165	166	167	168
I-1	MI13	MI-Cba2	active	active	R	M	K	K	-	V	V	L	E	I	P	D
I-2	MI43	MI-H.C	no	active	R	M	K	R	-	V	E	L	E	I	P	D
I-3	MI12	MI-Cba1	active	active	R	M	K	K	-	V	V	L	E	I	P	D
I-4	MI09	MI-R.G	active	no	R	M	K	K	-	V	V	L	E	I	P	D
I-5	MI11	MI-Cne	not tested	not tested	R	M	K	K	-	V	V	L	E	I	P	D
I-6	MI47	MI-Bbr	active	no	R	M	K	K	-	V	V	L	E	I	A	D
I-7	MI44	MI-B.R	active	no	R	M	K	K	-	V	V	L	E	I	P	D
II-1	MI42	MI-Agu	active	active	R	M	K	H	-	V	K	L	E	I	A	D
II-2	MI19	MI-PpuN	no	not tested	R	M	K	Q	-	V	K	L	E	I	P	D
II-3	MI07	MI-Pfl	active	not tested	R	M	K	Q	-	V	R	L	E	I	P	D
II-4	MI18	MI-P.C	not tested	not tested	R	M	K	Q	-	V	R	L	E	I	P	D
II-5	MI48	MI-Cca	active	active	R	M	K	H	-	V	N	L	E	I	A	D
II-6	MI05	MI-Sma	active	active	R	M	K	H	-	V	N	L	E	I	P	D
III-1	MI15	MI-P.U	active	not tested	R	M	K	H	-	V	T	L	E	I	Q	D
III-2	MI14	MI-Pha	not tested	not tested	R	M	K	H	-	V	T	L	E	I	Q	N
III-3	MI01	MI-PpuS	active	not tested	R	M	K	H	-	V	T	L	E	I	S	D
III-4	MI16	MI-Cbae	not tested	not tested	R	M	K	H	-	V	T	L	E	I	Q	D
III-5	MI17	MI-Cin	no	not tested	R	M	K	H	-	V	T	L	E	I	Q	D
III-6	MI08	MI-Rso	active	active	R	M	K	H	-	V	T	L	E	I	P	D
III-7	MI46	MI-Dac	not tested	not tested	R	M	K	H	-	V	T	L	E	I	A	D
III-8	MI41	MI-Axy	not tested	not tested	R	M	K	Q	-	V	T	L	E	I	A	D
IV-1	MI04	MI-Bst	no	not tested	R	M	M	H	-	V	T	L	E	I	P	D
IV-2	MI20	MI-Aac	no	not tested	R	M	M	H	-	V	T	L	E	I	P	D
IV-3	MI45	MI-P.B	active	active	R	M	M	H	-	V	T	L	E	V	S	D
IV-4	MI06	MI-Afa	not tested	not tested	R	M	M	H	-	V	N	L	E	V	S	D
IV-5	MI40	MI-Rpi	not tested	not tested	R	M	M	H	-	V	T	L	E	V	S	D
V-01	MI21	MI-Msm	active	active	R	M	K	H	-	V	T	L	E	V	S	D
V-02	MI49	MI-Aph	active	active	G	M	K	K	-	V	T	L	E	V	A	D
V-03	MI25	MI-Pdi	not tested	not tested	R	M	H	T	-	V	S	L	E	V	D	D
V-04	MI27	MI-Ach	not tested	not tested	R	M	H	K	-	V	T	L	E	V	G	D
V-05	MI52	MI-Pas	no	active	Q	M	H	T	-	V	S	L	E	V	A	D
V-06	MI22	MI-S.N1	active	active	R	M	H	A	-	V	T	L	G	V	A	D
V-07	MI26	MI-Abe	active	not tested	R	M	H	A	-	V	T	L	G	V	A	D
V-08	MI31	MI-Gkr	no	no	R	M	Q	A	-	V	T	L	E	V	A	D
V-09	MI39	MI-T.1	not tested	not tested	R	M	Q	K	-	V	S	L	E	V	S	D
V-10	MI02	MI-Rjo	active	not tested	R	M	Q	S	-	V	S	L	E	V	A	D
V-11	MI35	MI-Rop	not tested	not tested	R	M	Q	S	-	V	S	L	E	V	A	D
V-12	MI37	MI-Rwr	active	not tested	R	M	Q	A	-	V	S	L	E	V	A	D
V-13	MI36	MI-Rfa	not tested	not tested	R	M	Q	A	-	V	S	L	E	V	S	D
V-14	MI24	MI-S.S	not tested	not tested	R	M	H	T	-	V	S	L	E	V	A	D
V-15	MI50	MI-Ama1	not tested	not tested	R	M	Q	T	-	V	S	L	E	V	A	G
V-16	MI38	MI-Car	not tested	not tested	R	M	A	R	-	V	S	L	E	V	E	D
V-17	MI51	MI-Ama2	not tested	not tested	R	M	S	A	-	V	T	L	G	V	E	D
VI-1	MI29	MI-Grh	active	no	R	M	H	T	-	V	S	L	E	V	A	D
VI-2	MI30	MI-Gar	active	no	R	M	K	T	-	V	S	L	E	V	A	D
VI-3	MI34	MI-W.D	no	active	R	M	H	R	-	V	S	L	E	V	A	D
VI-4	MI32	MI-Nno	not tested	not tested	R	M	H	T	-	V	S	L	E	V	A	D
VI-5	MI03	MI-Nfa	active	not tested	R	M	H	T	-	V	S	L	E	V	A	D
VI-6	MI33	MI-Nbr	not tested	not tested	R	M	H	K	-	V	S	L	E	V	E	D
VI-7	MI23	MI-S.N2	no	not tested	R	M	H	T	-	V	S	L	E	V	S	D
VI-8	MI28	MI-Aor	active	no	R	M	H	T	-	V	S	L	E	V	S	D
VII-1	MI10	MI-Sso	no	not tested	K	L	R	N	-	V	T	M	G	I	R	E
VII-2	MI53	MI-meB	not tested	not tested	K	C	Y	N	P	L	T	F	D	I	A	S
VII-3	MI54	MI-meC	no	no	E	C	Y	N	P	L	T	F	D	I	E	A
VII-4	MI55	MI-meD	no	no	E	T	F	N	P	L	T	F	D	I	A	S

Residue numbering and domains are as described for PpuS16 [13].