FUNCTIONAL AND PHYSIOLOGICAL DISCOVERY IN THE MANNONATE DEHYDRATASE SUBGROUP OF THE ENOLASE SUPERFAMILY

BY

DANIEL JOSEPH WICHELECKI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biochemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor John Gerlt, Chair
Professor John Cronan
Professor Scott Silverman
Professor Wilfred van der Donk

ABSTRACT

In the current post-genomic world, the exponential amassing of protein sequences is overwhelming the scientific community's ability to experimentally assign each protein's function. The use of automated, homology-based annotations has allowed a reprieve from this efflux of data, but has led to widespread misannotation and nonannotation in protein sequence databases. This dissertation details the functional and physiological characterization of the mannonate dehydratase subgroup (ManD) of the enolase superfamily (ENS). The outcome affirms the dangers of homology-based annotations while discovering novel metabolic pathways. Furthermore, the experimental verification of these pathways (*in vitro* and *in vivo*) has provided a platform to test the general strategies for improved functional and metabolic characterization being developed by the Enzyme Function Initiative (EFI).

Prior to this study, one member of the ManD subgroup had been characterized and was shown to dehydrate D-mannonate to 2-keto-3-deoxy-D-gluconate. Forty-two additional members of the ManD, selected from across the sequence space of the subgroup, were screened for activity and kinetic constants were determined. The members of the once isofunctional subgroup were found to differ in both catalytic efficiency and substrate specificity: 1) high efficiency ($k_{cat}/K_M = 10^3$ to 10^4 $M^{-1}s^{-1}$) dehydration of D-mannonate, 2) low efficiency ($k_{cat}/K_M = 10^1$ to 10^2 $M^{-1}s^{-1}$) dehydration of D-mannonate and/or D-gluconate, and 3) no-activity with either D-mannonate or D-gluconate (or any other acid sugar tested). The novel D-gluconate activity in this subgroup was investigated, and the mechanism of its enzymatic action was discovered.

Physiologically, D-mannonate dehydration is essential to D-glucuronate metabolism. The D-mannonate dehydratase, UxuA, is not a member of the ENS. No *uxu*A genes are found in the

genome of organisms with high efficiency ManDs. Through *in vitro* characterization and *in vivo* verification, a high efficiency ManD was discovered in *Caulobacter crescentus* CB15 that fulfills the same physiological role as UxuA and is an example of convergent evolution.

The genomes of organisms with low efficiency members of the ManD subgroup generally have the *uxu*A gene. Therefore, they likely fulfill a different physiological role than the high efficiency ManDs. Their *in vitro* characterization and *in vivo* functional verification lead to the discovery of a novel L-gulonate metabolic pathway in *Chromohalobacter salexigens* DMS3043 where L-gulonate is converted to D-mannonate by a dehydrogenase and a reductase. While the low efficiency ManD found in *C. salexigens* is not metabolically essential to this pathway, its presence lead to the discovery of the pathway. Similar methods in *Salmonella enterica subsp. enterica serovar Enteritidis* str. P125109 lead to the discovery a novel L-idonate pathway where L-idonate is converted to D-gluconate by two dehydrogenases and then dehydrated (the traditional pathway phosphorylates D-gluconate). This pathway directly involves a low efficiency GlcD that is a member of the ManD subgroup and raises interesting questions about the physiological role of low efficiency enzymes and redundant pathways. As a whole, this dissertation displays how function diverges as sequence diverges while laying bare the dangers of annotation via homology; concurrently, it demonstrates how the continually advancing assignment strategies of the EFI can be used to discover new enzymatic functions and metabolic pathways.

ACKNOWLEDGEMENTS

I would like to thank my thesis committee (Dr. John Cronan, Dr. Scott Silverman, and Dr. Wilfred van der Donk) for taking the time to check my progress, give advice, and ask challenging questions; the Cellular and Molecular Biology Training Grant for financial support; Dr. John Rakus for discovering the first D-mannonate dehydratase in the enolase superfamily; former and present members of the Gerlt lab (Jason Bouvier, Bijoy Desai, Salehe Ghasempur, Fiona Poe, Dr. Tiit Lukk, Dr. Ayano Sakai*, Dr. Xinshuai Zhang, and Dr. Ben Warlick); and the former and present members of the Microbiology core of the Enzyme Function Initiative (EFI) (Dr. Kyuil Cho, Dr. Brad Evans, Dr. Amy Jones*, Dr. Ritesh Kumar, Dr. Indu Rupassara, Dr. Jose Solbiati, Dr. Brian San Francisco, and Dr. Bryant McKay Wood). I would like to thank the Protein Production Core for the monumental amount of work they have contributed to my thesis research, including protein purification, protein crystallization, and ThermoFluor screening (Dr. Steve Almo, Dr. Nawar Al-Obaidi, Dr. Alexander Federov, Dr. Brandan Hillerich, Dr. Yuriy Patskovsky, Dr. Rafael Toro, and Dr. Matt Vetting). My best to the many talented undergraduate researchers I have mentored over the years (Bryan Balthazar, Anthony Chau, Dylan Graff, and Jean Vendiola); they took years off my "to graduation" time. I thank everyone for putting up with my automaton-like work persona, and for getting to know me more outside of work (most prominently Jason Bouvier, Anthony Chau, Dr. Ayano Sakai, Dr. Brian San Francisco, and Dr. Ben Warlick).

I would also like to specifically thank Dr. Ayano Sakai, who trained me in almost every experimental technique I learned in this 5 year experience. Her patience, knowledge, and friendship will never be forgotten. Above all, I would like to thank my advisor and mentor, Dr.

John Gerlt for his support and guidance through my PhD experience. Who, while beset by the immense time burden of directing the EFI, always found time to discuss my ManDs when I needed advice. I never imagined how much I would grow as a scientist under his influence. I thank him most for being honest in criticism and free with praise.

Most importantly, I would like to thank my fiancée, Dr. Katie Whalen, whose companionship, patience, love, and writing skill has carried me safely over the many hurdles I've encountered in the PhD process.

Naturally, I thank my loving family members for all of there support through the years: Mom, Dad, Jana, Steven, and Cassidy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SCHEMES

CHAPTER 1: INTRODUCTION

1.1. Focus of Study

This dissertation outlines the development and execution of an experimental strategy whose purpose is to combat the problem of mis- and nonannotation in protein databases. Specifically, this strategy was employed to facilitate functional annotations of enolase superfamily members in the D-mannonate (ManD) subgroup and the L-fuconate dehydratase family in the mandelate racemase (MR) subgroup. In doing so, a novel activity was found in the ManD subgroup, a previously undiscovered example of convergent evolution was discovered, a novel metabolic pathway was established, and many physiological functions were correctly annotated. This work was completed as part of the Enzyme Function Initiative (EFI), and, as such, is of general interest to a broad host of disciplines including, but not limited to, bioinformatics, enzymology, systems biology, metabolic engineering, and microbiology.

1.2. Dilemma of the Post-Genomic Era

The genomic era began in 1973, when Fiers determined the sequence of the first gene, bacteriophage MS2 coat protein [1]. Technology advanced such that in ~20 years the sequence of the first bacterial genome was published [2]. In 2003, exactly 30 years after the sequence of the first gene was deciphered, the International Human Genome Sequencing Consortium (IHGSC) concluded the genomics era with the complete sequencing of the human genome [3, 4]. Thus began the dawn of the post-genomic era, and realization of the informational overload which would follow.

In the post-genomic era, sequencing technology has advanced to such a degree that it is routine to sequence an organism's genome. This has led to an exponential rise in the number of proteins in the Translated European Molecular Biology Laboratory Nucleotide Sequence Database (TrEMBL). The TrEMBL database automatically annotates its proteins via homology, since it is impossible to experimentally verify every protein discovered in sequencing projects. This fact is evident in the Swiss-Prot protein database, which manually curates its annotations via literature citations, taxonomy, and cross-referenced databases (www.uniprot.org). At the time of composition of this thesis, the TrEMBL database comprised of ~53M proteins whereas the Swiss-Prot database only comprised of ~550k entries (~1% overlap). Strikingly, TrEMBL database entries are growing exponentially and Swiss-Prot entries are leveling off (**Figure 1.1**). The realization that the vast majority of proteins are annotated by automated, homology-based

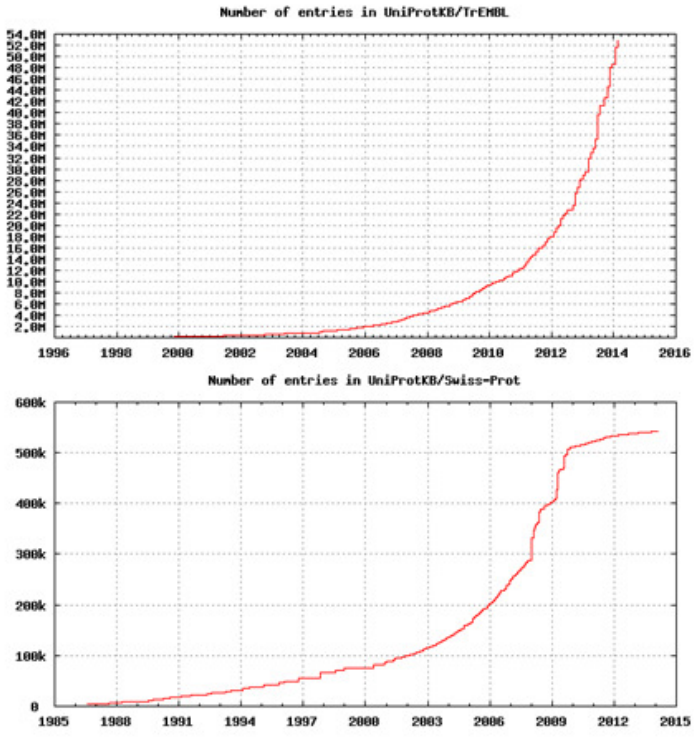


Figure 1.1. The number of TrEMBL and Swis-Prot database entries at the time of composition.

methods has raised concerns among the biological community of the accuracy of protein databases such as TrEMBL.

1.2.1. Mis- and Non-annotation in Protein Sequence Databases

As early as 1996, it was recognized that automated methods would be needed to sift through all of the genetic data that would be generated as more organisms had their genome's sequenced [5]. By 1998, apprehension of using homology-based annotations was already building. Karp called into question the accuracy of the process, and warned of the catastrophe rgr propagation of error could cause [6]. This warning was validated in 2002 when Gilks and co-workers studied the percolation of error in protein databases; their conclusion was that complete loss of information would ultimately arise if the quality of homology-based annotations was not improved [7]. In fact, by this time, error was already rampant in the databases: a 2001 study by Devos and Valencia claimed that erroneous annotations may account for as high as 30% of all annotations [8]. Presumably, this error percentage would be higher when only looking at homology-based annotations. As such, a 2007 study of the Gene Ontology database estimated that up to 49% of homology-based annotations may be misannotated [9]. As the number of proteins in the TrEMBL database (i.e. proteins annotated by homology) continue to grow, and dwarf the number of manually curated proteins; the warnings of Karp and Gilks become more corporeal. New methods are needed not only to keep misannotation in check, but to quickly and accurately aide in the annotation of newly discovered proteins from genome sequencing projects.

1.3 The Enzyme Function Initiative

The widespread non- and misannoation present in protein databases is an extensive problem. Any method which hopes to combat this must be capable of predicting functions in a high-throughput manner, but also enable the focused enzymatic characterization of a practical number of proteins. These newly discovered functions would be intelligently transferable to similar proteins and greatly improve the current number of protein annotations with experimental evidence (even if indirect). The Enzyme Function Initiative (EFI) was founded in May 2010 with this aim in mind [10]. The specific goal being "to develop and disseminate integrated sequence/structure-based strategies that incorporate a spectrum of essential expertise, including high-throughput bioinformatics, computation, and structural biology that will enable focused experimental enzymology, genetics, and metabolomics." (http://enzymefunction.org/about/overview).

1.3.1 Strategy of the EFI

The EFI follows a general strategy beginning with bioinformatics and culminating with the determination of a protein's physiological function through *in vivo* experimentation. An enzyme's "protein family" as classified by Pfam [11], InterPro [12], or *ad hoc* BLASTing can be used to generate a sequence similarity network (SSN, [13]) of all homologous proteins within that family where nodes represent protein sequences and connections are drawn between nodes that satisfy a user-defined sequence identity cutoff. Clusters formed within the SSN are evaluated for known/unknown functions. This network topology is instrumental in target determination and assignment of function to a protein of unknown function.

When a target is chosen, it is cloned from its source organism and expressed in *Escherichia coli*. If soluble, the protein is purified and crystallized, and subjected to screening for binding partners via ThermoFluor [14], and its *in vitro* enzymatic activity is characterized (high-throughput screening, testing of proposed substrates from docking/ThermoFluor, etc.). If the target is crystallized, or a structure for a closely-related homologue is available, metabolite docking [15-17] can be performed on the crystal structure (or homology model) to aid in substrate determination. ThermoFluor is a fluorescence-based method which monitors if a metabolite stabilizes a protein by determining the melting temperature of a target protein in the presence and absence of ligand. If a metabolite stabilizes the target protein, it is thought to bind the protein and is a potential ligand/substrate. Thus, substrate determination for i*n vitro* enzymatic characterization can be aided by metabolite docking hits, ThermoFluor hits, genome context, or high-throughput screening of compound libraries.

Once the substrate is determined, *in vivo* data must be obtained to verify that the reaction is physiologically relevant. Within the EFI, three main techniques exist to achieve this goal: transcriptomics, phenotypic knockout studies, and metabolomics. Reverse-transcriptase PCR [18] is the primary form of transcriptomics used in the EFI, but in some cases RNA-seq [19] is implemented. Phenotypic knockout studies aim to find growth/no growth phenotypes for mutants, but this can vary depending on the target of interest. While metabolomics has a large variety of uses [20-22], the EFI primarily utilizes it to find intermediates in specific metabolic pathways. It is particularly useful when combined with knockout studies, as specific intermediates will accumulate in the cellular cytoplasm as a result of impaired metabolism.

This is the idea of ‘new’ versus ‘old’ enzymology that has been introduced by the EFI. Old enzymology defines a protein’s function solely based on a chemical reaction observed *in*

vitro. New enzymology demands that *in vivo* evidence validate the *in vitro* reaction before a function is assigned. This ensures accurate annotation of the gene. The work in this thesis describes the implementation of the EFI strategy to functionally and physiologically characterize enolase superfamily members belonging to the D-mannonate dehydratase (ManD) subgroup in addition to L-fuconate dehydratases (FucD) in the MR subgroup.

1.4 The Enolase Superfamily

The enolase superfamily (ENS) was discovered when Babbitt and co-workers realized the structural and mechanistic similarities between enolase, muconate-lactonizing enzyme (MLE), and mandelate racemase (MR) [23, 24] (**Figure 1.2**). In the ENS, evolution modified the common active site architectures to allow various enolic intermediates to differ in their overall reactions. Specifically, ENS members share an (α + β) capping domain for substrate specificity and a modified $(\beta/\alpha)_8$-barrel domain ($(\beta/\alpha)_7\beta$-barrel) for acid/base chemistry [25, 26]. Mechanistic similarities stem from a common partial reaction: abstraction of a proton alpha to a carboxylate group to form an enedioate intermediate that is stabilized by a divalent metal cation (usually Mg^{2+}) [23] (**Figure 1.3**). Subgroups are differentiated based on the conserved metal-binding residues at the ends of the 3rd, 4th, and 5th β-strands, as well as conserved catalytic acid/base residues at the ends of the 2nd, 3rd, 6th, and 7th β-strands of the modified $(\beta/\alpha)_8$-barrel domain. The ENS is particularly interesting because the members share the same structural motifs and utilize common chemistry, but are functionally diverse (e.g. β-elimination and 1,1-proton transfer reactions) [10, 27] (**Figure 1.4**). This makes the superfamily an ideal model obtaining a better definition of the structure-function relationship. The following work has

implications not only in more refined functional annotation algorithms, but also enzyme evolution and protein engineering.

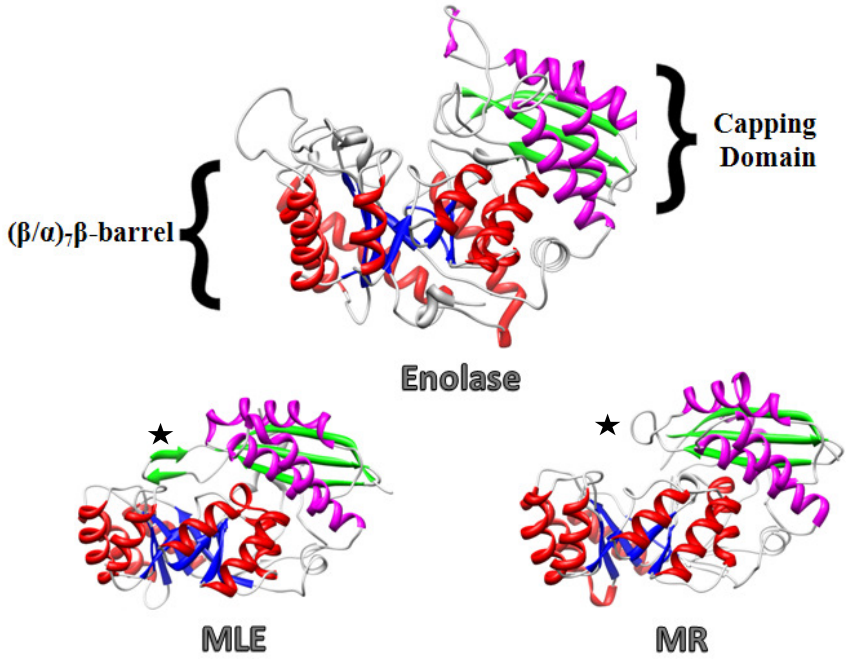


Figure 1.2. Crystal structures of an enolase (PDB 1PDY), MLE (PDB 1BKH), and MR (PDB 2MNR). The common structural motifs are labeled on 1PDY. The capping domain has green-labeled β-sheets and magenta-labeled α-helices. The modified TIM-barrel has blue-labeled β-sheets and red-labeled α-helices. In the MR and MLE subgroups, the "20s" loop (star) from capping domain extends over the active site.

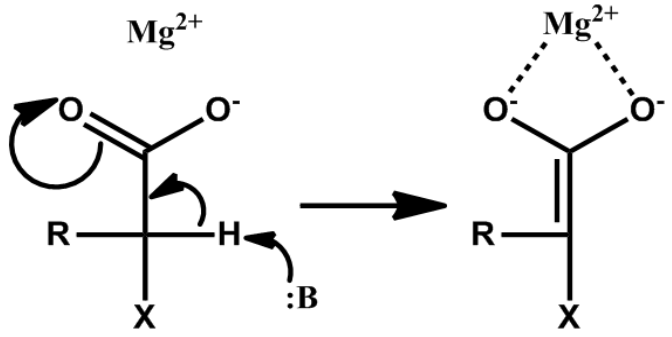


Figure 1.3. The common partial reaction shared by all ENS members: base-catalyzed abstraction of a proton alpha to a carboxylate group to form an enediolate intermediate. This intermediate is stabilized by the divalent metal cation, Mg^{2+}.

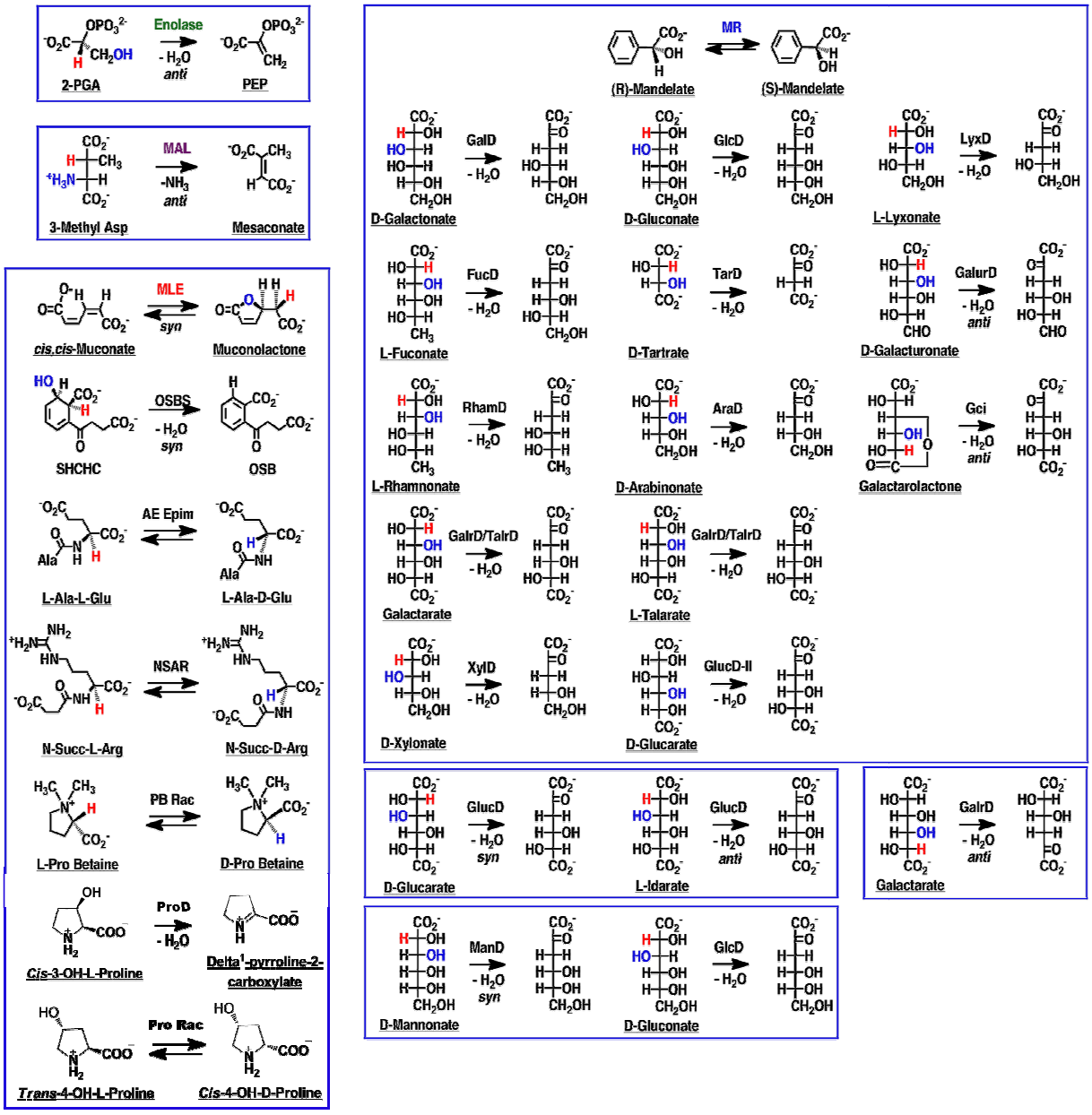


Figure 1.4. All known reactions within the ENS. Distinct subgroups are separated by blue boxes. Adapted from a power point slide provided by Dr. John Gerlt.

1.4.1 The Mannonate Dehydratase Subgroup

The D-mannonate dehydratase subgroup (ManD) of the ENS was discovered by Rakus and co-workers in 2007 [28]. A ManD from *Novosphingobium aromaticivorans* (*Na*ManD) was screened for dehydration activity using a library of acid sugars and found to be specific to the dehydration of D-mannonate to 2-keto-3-deoxy-D-mannonate: k_{cat} of 1.3 ± 0.1 s^{-1} and k_{cat}/K_M of 6.9 x 10^3 M^{-1} s^{-1} (**Figure 1.5**). The residues responsible for Mg^{2+} binding are Asp, Glu, and Glu at the end of the 3rd, 4th, and 5th β-strands [28].

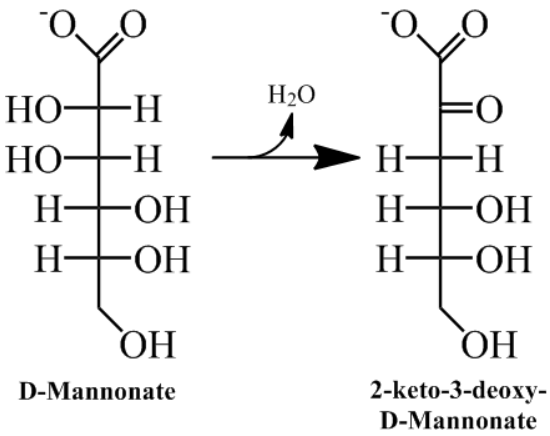


Figure 1.5. The dehydration of D-mannonate to 2-keto-3-deoxy-D-mannonate. *Na*ManD performs this reaction with a k_{cat} of 1.3 ± 0.1 s^{-1} and k_{cat}/K_M of 6.9 x 10^3 M^{-1} s^{-1}.

The catalytic base is a Tyr located at the start of the "150-180s" loop (between the 2nd and 3rd β-strands); the catalytic acid, His, is located at the end of the 3rd β-strand [28]. The "150-180s" loop is unique to the ManD subgroup. Other acid sugar dehydratases are found in the enolase superfamily, but they occlude the active site from solvent using a "20s" loop from the capping domain (seen in **Figure 1.2**). This sets ManDs apart in that their capping domain does not

interact with the active site directly like most other ENS members (**Figure 1.6**). Physiologically, D-mannonate dehydration is known to be involved in D-glucuronate metabolism [29], although, the enzyme characterized to dehydrate D-mannonate in this pathway, UxuA (**u**tilization of he**xu**ronate **A**), is not an ENS member. Therefore, it remains to be verified that ManDs are also involved in D-glucuronate metabolism. In this dissertation, it is determined how function diverges as sequence diverges in the ManD subgroup, and the physiological role of high and low efficiency ManDs is investigated.

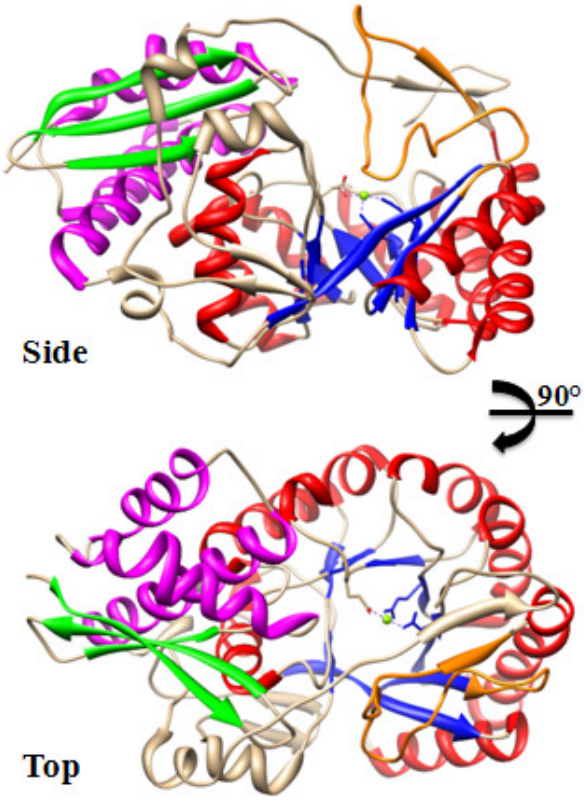


Figure 1.6. Crystal structure of *Na*ManD (PDB 2QJJ). The capping domain has green-labeled β-sheets and magenta-labeled α-helices. The modified TIM-barrel has blue-labeled β-sheets and red-labeled α-helices. The "150-180s" loop is colored orange. Mg^{2+} and the metal binding residues are shown to visualize the active site.

1.4.2 Fuconate Dehydratases in the Mandelate Racemase Subgroup

L-Fuconate dehydration activity was discovered in the MR subgroup of the ENS by Yew and co-workers [30]. The L-fuconate dehydratase (FucD) from *Xanthomonas campestris* (*Xc*FucD) was found to dehydrate L-fuconate with a k_{cat} of 15 ± 0.2 s^{-1} and k_{cat}/K_M of 4.5 x 10^4 M^{-1} s^{-1} [30] (**Figure 1.7**). Being of the MR subgroup, the Mg^{2+} binding residues are the Asp, Glu, and Glu at the end of the 3rd, 4th, and 5th β-strands respectively [24]. Additionally, all MR subgroup members have a conserved general base, KxK motif at the end of the 2nd β-strand, and general acid, His-Asp dyad located at the ends of the 6th and 7th β-strands, respectively [24]. Physiologically, L-fuconate dehydration is proposed to be involved in the metabolism of L-fucose though no *in vivo* experimentation has been performed [30]. Interestingly, other than the namesake enzyme, enolase, L-fuconate dehydratases are the only ENS members found in eukaryotes. In this dissertation, a human FucD, rTSγ, is characterized as an L-fuconate dehydratase and the role of an isoform, rTSβ, in chemotherapeutic resistance is investigated.

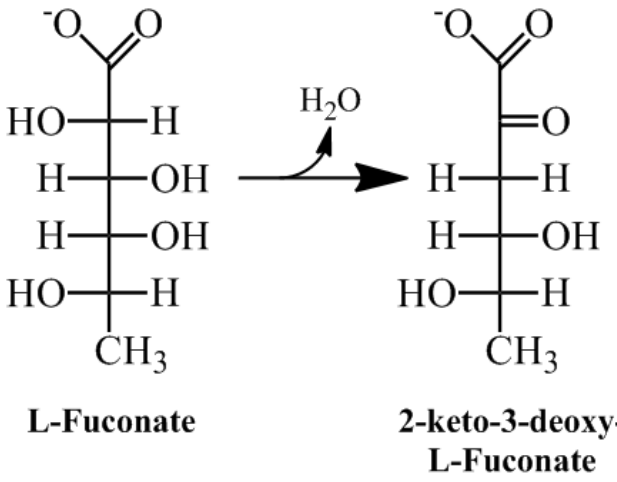


Figure 1.7. The dehydration of L-fuconate to 2-keto-3-deoxy-L-fuconate. *Xc*FucD performs this reaction with a k_{cat} of 15 ± 0.2 s^{-1} and k_{cat}/K_M of 4.5 x 10^4 M^{-1} s^{-1}.

CHAPTER 2: DISCOVERY OF FUNCTION IN THE ENOLASE SUPERFAMILY; D-MANNONATE AND D-GLUCONATE DEHYDRATASES IN THE D-MANNONATE DEHYDRATASE SUBGROUP

ABSTRACT: The continued increase in the size of the protein sequence databases as a result of advances in genome sequencing technology is overwhelming the ability to perform experimental characterization of function. Consequently, functions are assigned to the vast majority of proteins *via* automated, homology-based methods, with the result that as many as 50% are incorrectly annotated or unannotated (Schnoes *et al.*, 2009). This manuscript describes a study of the D-mannonate dehydratase (ManD) subgroup of the enolase superfamily (ENS) to investigate how function diverges as sequence diverges. Previously, one member of the subgroup had been experimentally characterized as ManD [dehydration of D-mannonate to 2-keto-3-deoxy-D-mannonate (equivalently, 2-keto-3-deoxy-D-gluconate)]. In this study, 42 additional members were characterized to sample sequence-function space in the ManD subgroup. These were found to differ in both catalytic efficiency and substrate specificity: 1) high efficiency (k_{cat}/K_M = 10^3 to 10^4 $M^{-1}s^{-1}$) for dehydration of D-mannonate, 2) low efficiency (k_{cat}/K_M = 10^1 to 10^2 $M^{-1}s^{-1}$) for dehydration of D-mannonate and/or D-gluconate, and 3) no-activity with either D-mannonate or D-gluconate (or any other acid sugar tested). Thus, the ManD subgroup is not isofunctional and includes D-gluconate dehydratases (GlcDs) that are divergent from the GlcDs that have been characterized in the mandelate racemase subgroup of the ENS (Lamble *et al.*, 2004) (Ahmed *et al.*, 2005). These observations signal caution for functional assignment based on sequence homology and lay the foundation for the studies of the physiological functions of the GlcDs and the promiscuous ManDs/GlcDs.

Reprinted (adapted) with permission from Wichelecki, D. J., Balthazor, B. M., Chau, A. A., Vetting, M. W., Fedorov, A. A., Fedorov, E. V., Lukk, T., Patskovsky, Y. V., Stead, M. B., Hillerich, B. S., Seidel, R. D., Almo, S. C., and Gerlt, J. A. Biochemistry. 2014 Apr 29;53(16):2722-31.

2.1. Introduction

The massive influx of sequence data since the first bacterial genome sequence was published in 1995 has necessitated a reliance on homology-based annotations of protein function [2, 5]. However, because this method assigns the function of the "closest" homologue, an estimated 30-50% of the functional annotations in the databases are incorrect [8, 9, 31], with the magnitude of the problem increasing as the incorrect annotations are propagated in assigning functions to proteins discovered in newly sequenced genomes. In a study of several functionally diverse superfamilies, Schnoes, Babbitt, and coworkers concluded that 85% of misannotations resulted from annotations which were more detailed than justified[3]. Automated methods often are able to achieve high degrees of accuracy in the transfer of the first three Enzyme Commission (EC) code numbers, but accurate transfer of the fourth EC code number (substrate specificity) is much more difficult [32]. This study examines the D-mannonate dehydratase (ManD) subgroup of the enolase superfamily (ENS) to determine experimentally, on a large scale, how function diverges as sequence diverges in highly homologous enzymes. Our results illustrate the difficulty of accurately assigning function via homology-based methods and, also, provide insights into how different functions can arise in highly homologous enzymes.

Two conserved features are shared by members of the ENS: mechanism and structure. The mechanism is general base-catalyzed abstraction of a proton alpha to a carboxylate group of

the substrate to form an enediolate intermediate [23]. The enediolate intermediate is stabilized by coordination to an active site divalent metal cation (usually Mg^{2+}). Furthermore, members of the ENS share a common structural motif: an (α+β) capping domain that contains the residues that determine substrate specificity and a modified TIM-barrel domain ($(\beta/\alpha)_7\beta$-barrel) that contains the residues that mediate acid/base chemistry [25, 26]. Subgroups are differentiated by the conserved metal-binding residues at the ends of the 3rd, 4th, and 5th β-strands as well as conserved catalytic acid/base residues at the ends of the 2nd, 3rd, 6th, and/or 7th β-strands of the modified TIM-barrel domain. The ENS is particularly interesting because its members share a common mechanism and the same structural motif but are functionally diverse (e.g., β-elimination and 1,1-proton transfer reactions) [10, 27]. Thus, the ENS is a good model to investigate how function diverges as sequence diverges.

In 2007, Rakus and coworkers discovered the ManD subgroup of the ENS [28]. In that study, a protein from *Novosphingobium aromaticivorans* (*Na*ManD) was structurally characterized and discovered to catalyze the *syn*-dehydration of D-mannonate to 2-keto-3-deoxy-D-mannonate (equivalently, 2-keto-3-deoxy-D-gluconate) (EC 4.2.1.8 and Unitprot ID A4XF23). The Mg^{2+}-binding residues located at the ends of the 3rd, 4th, and 5th β-strands are Asp 210, Glu 236, and Glu 262, respectively. The general base that abstracts the 2-proton is the Tyr 159-Arg 147 dyad in the "150-180s" loop between the 2nd and 3rd β-strands; the general acid that facilitates departure of the 3-hydroxyl group is His 212 located at the end of the 3rd β-strand [28] (**Figure 2.1**). His 315 located at the end of the 7th β-strand that is hydrogen-bonded to the 5-hydroxyl group of the D-mannonate substrate is also conserved. Since this study, no other members of the ManD subgroup have been experimentally characterized. Given the

conservation of metal-binding residues, catalytic residues, and active site architecture, the assumption was that the entire subgroup is isofunctional.

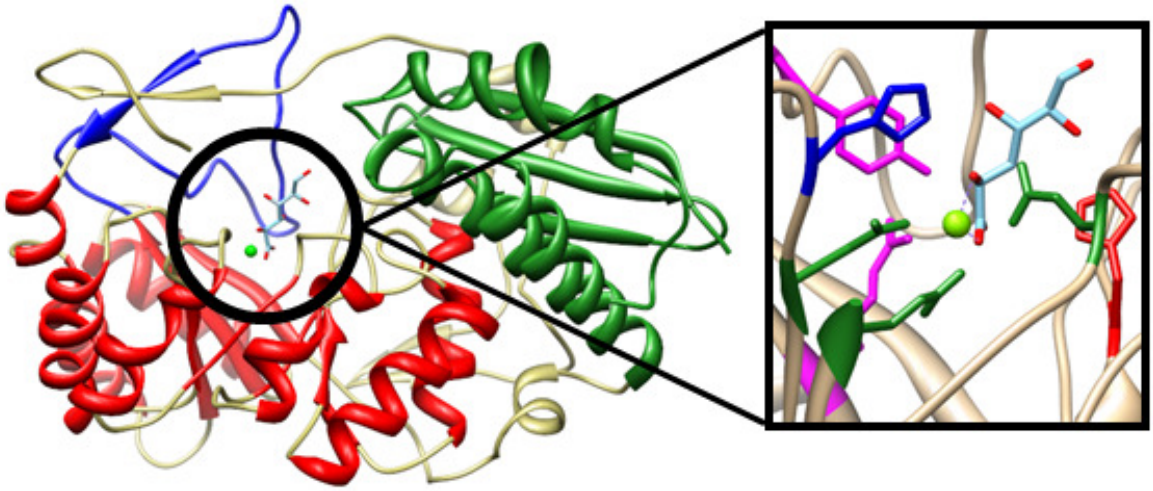
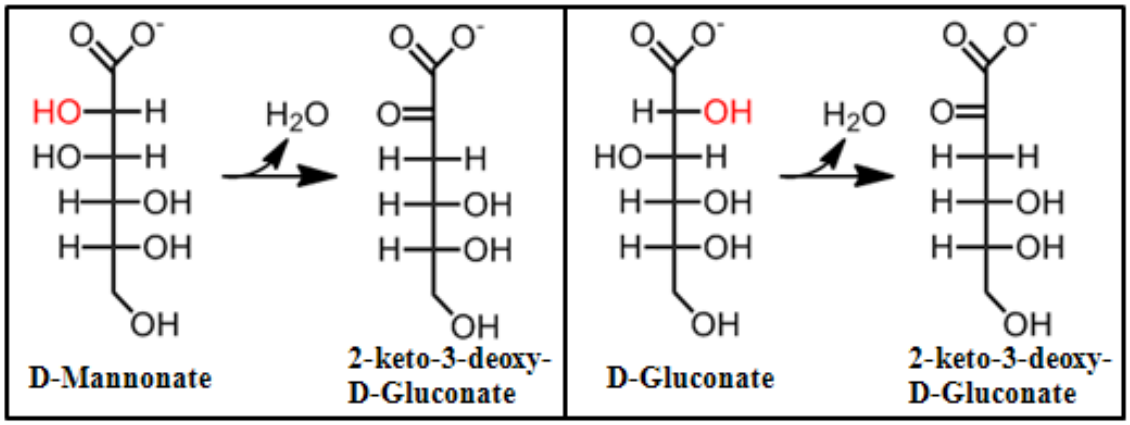


Figure 2.1. Denotes the structural features of the ManDs (PDB 2QJJ). On the left, the "150-180s" loop (blue), TIM barrel (red), and capping domain (green) are displayed. The right inset shows the active site residues: metal binding Asp210, Glu236, Glu262 (green); Tyr157-Arg149 catalytic dyad (magenta); acidic His212 (blue); and conserved His315 at the end of the 7th β-strand (red). The ligand, D-mannonate, is shown in light blue from the 2QJM structure.

In this study, we sought to investigate how function diverges as sequence diverges within this subgroup and to determine if the ManD subgroup is, in fact, isofunctional. When the target proteins for this study were selected in April 2011, the ManD subgroup included *Na*ManD [28] and 299 uncharacterized proteins that share ≥35% sequence identity [Structure-Function Linkage Database (http://sfld.rbvi.ucsf.edu/)]. [At the time of submission of this manuscript, the UniProtKB database contained the sequences for 2919 members of the ManD subgroup.] Forty three members representing the breadth of sequence-function space were produced as soluble proteins and screened for activity using a library of acid sugars. Surprisingly, we found that the ManD subgroup is *not* isofunctional; instead, in addition to ManDs it also contains D-gluconate

dehydratases (GlcDs) that catalyze the *anti*-dehydration of D-gluconate to 2-keto-3-deoxy-D-gluconate as well promiscuous proteins that catalyze both the ManD and GlcD reactions (**Scheme 2.1**). In addition, a wide range of catalytic efficiencies (values of k_{cat}/K_M) was



Scheme 2.1. The dehydration of D-mannonate/D-gluconate to 2-keto-3-deoxy-D-gluconate.

discovered. Using sequence similarity networks (SSNs) [13], the members with these divergent functions could be separated into isofunctional clusters. Furthermore, sixteen unique crystal structures (for a total of 36 unliganded and liganded structures) were solved to survey sequence and structure space; these revealed conserved active site structures but divergent conformations for the "150-180s" loops that contain the general basic Tyr-Arg dyads and close over the active site to sequester the substrate from solvent. Taken together, the functional and structural data provide a comprehensive description of how *in vitro* function diverges as sequence diverges.

2.2 Materials and methods

Cloning, Expression, and Purification of Targets (AECOM). pNIC28-BSA4-based expression vectors were transformed into BL21(DE3) *E. coli* containing the pRIL plasmid (Stratagene) and used to inoculate a 10mL 2xYT culture containing 25ug/mL Kanamycin and

34ug/mL Chloramphenicol. The cultures were allowed to grow overnight at 37°C in a shaking incubator. The overnight culture was used to inoculate 2L of PASM-5052 auto-induction media [33]. The culture was placed in a LEX48 airlift fermenter and incubated at 37°C for 4 hours and then at 22°C overnight. The culture was harvested and pelleted by centrifugation.

Cells were resuspended in lysis buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 20 mM imidazole, and 10% glycerol) and lysed by sonication. The lysate was clarified by centrifugation at 35,000g for 30 min. The protein was purified using an AKTAxpress FPLC (GE Healthcare). The lysate was loaded onto a 1 mL His60 column (Clontech), washed with 10 column volumes of lysis buffer, and eluted with buffer containing 20 mM HEPES, pH 7.5, 500 mM NaCl, 500 mM Imidazole, and 10% glycerol. This partially purified protein was loaded onto a HiLoad S200 16/60 PR gel filtration column equilibrated with SECB buffer (20 mM HEPES, pH7.5, 150 mM NaCl, 10% glycerol, and 5 mM DTT). The protein was analyzed by SDS-PAGE, flash frozen in liquid nitrogen, and stored at -80°C.

Expression and Purification of N-terminal His-tagged Proteins (UIUC). Genes in pET15b (Novagen) were expressed in *Escherichia coli* strain BL21. Small-scale cultures were grown at 37˚ C for 18 hrs in 5 mL of LB containing 100 μg/mL ampicillin and used to inoculate 1 L LB containing 100 μg/mL ampicillin. IPTG (500 μM) was added at $OD_{600\ nm}$ = 0.6-0.8 to induce expression. The induced cultures then were grown for an additional 18 hrs at 37˚ C. The cells were harvested by centrifugation at 5000 rpm for 10 minutes and resuspended in 70 mL of binding buffer (6 mM imidazole, 20 mM Tris-HCl, pH 7.9, 5 mM $MgCl_2$, and 500 mM NaCl). The resuspended cells were lysed by sonication and centrifuged at 17,000 rpm for 30 minutes. The supernatant was loaded onto a column of 50 mL chelating Sepharose Fast Flow (Amersham Biosciences) charged with Ni^{2+} and eluted with a linear gradient of imidazole (0 – 1 M over 600

mL). Fractions were analyzed using SDS-PAGE. Fractions containing protein at high purity (>90%) were combined and dialyzed against 4 L of buffer containing 100 mM imidazole, 20 mM Tris-HCl, pH 7.9, 10 mM $MgCl_2$, 150 mM NaCl, and 10% glycerol for 2 hrs at 4 ˚C. The protein was dialyzed in this manner a total of three times. Finally, the protein was concentrated to a maximum of ~10 mg/mL (depending on solubility) and flash-frozen using liquid nitrogen and stored at -80˚ C.

Expression and Purification of Tagless ManD Constructs (UIUC). The genes in pET17b (Novagen) were expressed in *Escherichia coli* strain BL21. Small-scale cultures were grown at 37 ˚C for 18 hrs in 5 mL of LB containing 100 μg/mL ampicillin and used to inoculate 1 L LB containing 100 μg/mL ampicillin. The 1L cultures were grown for an additional 18 hrs at 37 ˚C without induction. The cells were harvested by centrifugation at 5000 rpm for 10 minutes and resuspended in 70 mL of binding buffer. The resuspended cells were lysed by sonication and centrifuged at 17,000 rpm for 30 minutes. The supernatant was loaded onto a 300 mL DEAE Sepharose column (Amersham Biosciences) and eluted with a linear gradient of NaCl (0 - 1 M over 1.6 L) in 10 mM Tris-HCl, pH 7.9, containing 5 mM $MgCl_2$. Fractions containing the protein of interest at high purity were combined and dialyzed against 4 L buffer containing 10 mM Tris-HCl, pH 7.9, and 5 mM $MgCl_2$ for 2 hours at 4 ˚C. The dialyzed protein was then loaded onto a 30 mL Q-Sepharose column (Amersham Biosciences) and eluted with a linear gradient of NaCl (0 - 1 M over 500 mL) in 10 mM Tris-HCl,pH 7.9, containing 5 mM $MgCl_2$. Fractions containing the protein of interest at high purity were combined and dialyzed in 4 L buffer containing 10 mM Tris-HCl, pH 7.9, containing 5 mM $MgCl_2$ for 2 hrs at 4 ˚C. Ammonium sulfate was added to a final concentration of 1 M, and the sample was loaded onto a 30 mL phenylsepharose column (Amersham Biosciences). The protein was eluted with a

gradient of ammonium sulfate (1 - 0 M over 500 mL) in 10 mM Tris-HCl, pH 7.9, containing 5 mM $MgCl_2$. Fractions with pure protein (SDS-PAGE) were combined and dialyzed against 4 L buffer containing 10 mM Tris-HCl, pH 7.9, 5 mM $MgCl_2$, 150 mM NaCl, and 10% glycerol for 2 hrs at 4 ˚C. Finally, the protein was concentrated to a maximum of ~10 mg/ml (depending on solubility) and flash frozen in liquid nitrogen and stored at -80˚ C.

Screen for Dehydration. Reactions to test for enzymatic activity were performed in acrylic, UV transparent 96-well plates (Corning Incorporated) using a library of 72 acid sugars (**Figure A.1**). Reactions (60 µL) contained 50 mM HEPES, pH 7.9, 10 mM $MgCl_2$, 1 µM enzyme, and 1 mM of acid sugar substrate (blanks without enzyme). The plates were incubated at 30 ˚C for 16 hrs. A 1% semicarbazide reagent solution (240 µL) was added to each well and incubated for 1 hr at room temperature. The absorbancies were measured at 250 nm (ε = 10,200 $M^{-1}cm^{-1}$) using a Tecan Infinite M200PRO plate reader.

Kinetic Assays. Dehydration of D-mannonate and D-gluconate was quantitated using either a discontinuous assay with the semicarbazide reagent [34, 35] or a continuous, coupled-enzyme spectrophotometric assay. In the latter assay, the product was phosphorylated using 2-keto-3-deoxy-D-gluconate kinase (KdgK) and ATP; formation of ADP was measured using pyruvate kinase (PK) and L-lactate dehydrogenase (LDH). The assay (200 µL) at 25 °C contained 50 mM potassium HEPES, pH 7.5, 5 mM $MgCl_2$, 1.5 mM ATP, 1.5 mM PEP, 0.16 mM NADH, 9 units of PK, 9 units of LDH, 18 units of KdgK, and ManD/GlcD. Dehydration was quantitated by measuring the decrease in absorbance at 340 nm (ε = 6220 M^{-1} cm^{-1}). Low-activity enzymes were characterized using the discontinuous assay; high-activity ManDs were characterized using the coupled assay.

Site-Directed Mutagenesis (SDM). Mutants were constructed using primers designed with the Agilent Technologies online webserver (https://www.genomics.agilent.com/) and purchased from Bio-Synthesis, Inc. Forward and reverse primers containing the mutations of interest are listed in **Table A.1**. PCR reactions (30 μL) contained 1 mM $MgCl_2$, 1X Pfx Amp Buffer, 0.33 mM dNTP, 0.33 μM of FOR/REV primer, and 1.25 units Pfx polymerase (Invitrogen Platinum Pfx DNA Polymerase kit). The templates were 50 ng ManD-containing pET15b (*Na*ManD) or pET17b (*Cs*ManD). Amplifications were performed according to the manufacturer's guidelines. After addition of DpnI (10 units), the reactions were incubated for 4 hrs at 37 ˚C. The DpnI-digested products were purified by gel electrophoresis, extracted and transformed (electroporation, Bio-Rad Micropulsar Electroporator) into XL1 Blue competent cells. Finally, plasmids isolated from the transformants were sequenced to confirm the mutations.

Circular Dichroism of NaManD and CsManD Loop Mutants. The circular dichroism spectrum of a 10 μM solution of mutant enzyme in an optically clear borate buffer (50 mM boric acid, 100 mM KCl, 0.7 mM DTT, pH 8.0) was measured from 190-260 nm using a Jasco J-715 spectropolarimeter. Five replicate measurements were made.

Protein crystallization and X-ray diffraction data collection. Proteins were crystallized by the sitting-drop vapor diffusion method. The concentrated (usually 5-40 mg/ml) protein solutions (from 0.3 to 1 μL) were mixed with an equal volume of a precipitant solution and equilibrated at room temperature (~294 °K) against the same precipitant solution in clear tape-sealed 96-well INTELLI-plates (Art Robbins Instruments, Sunnyvale, CA). Crystallization was performed using either a TECAN crystallization robot (TECAN US, Research Triangle Park, NC) or a PHOENIX crystallization robot (Art Robbins Instruments) and four types of

commercial crystallization screens: the WIZARD I&II screen (Emerald BioSystems, Bainbridge Island, WA); the INDEX HT and the CRYSTAL SCREEN HT (both from Hampton Research, Aliso Viejo, CA); and the MCSG screen (Microlytic, Woburn, MA). The appearance of protein crystals has been monitored either by visual inspection or using a Rock Imager 1000 (Formulatrix, Waltham, MA) starting within 24 hours of incubation and again at weeks 1, 2, 3, 5, 8, and 12. Where necessary, the crystallization conditions were optimized manually using 24-well Cryschem sitting drop plates (Hampton Research). The crystallization conditions for all crystal structures are listed in the PDB methods tab and the Supplementary Information. The crystals were either directly frozen in liquid nitrogen or treated with a cryoprotectant (glycerol or ethylene glycol, 20-30%, vol/vol) before freezing.

The X-ray diffraction data for the frozen crystals were collected at 100 °K on the beamline X29A (National Synchrotron Light Source, Brookhaven National Laboratory, Upton, NY) using a wavelength of 1.075 Å or on the beamline 31I-D (LRL-CAT, Advanced Photon Source, Argonne National Laboratory, IL, USA) using a wavelength of 0.9793 Å. The diffraction data were processed and scaled with SCALA [36] (APS data) or HKL [37] (NSLS data). The crystal structures reported here were determine by molecular replacement using coordinates for similar structures from the PDB (listed in each PDB deposition as REMARK 200: STARTING MODEL) and PHASER MR software (the CCP4 program package suite [38]. Each structure was refined using the programs REFMAC [39] or PHENIX [40] and the resulting models were rebuilt manually using COOT visualization and refinement software [41]. The data collection and refinement statistics for all crystal structures are listed in the **Table A.2**.

2.3 Results and Discussion

Selection of Targets. In April 2011, the Structure-Function Linkage Database (SFLD; sfld.rbvi.ucsf.edu/) contained 300 sequences for the ManD subgroup of the ENS (*Na*ManD and 299 uncharacterized homologues). These sequences were used to generate a sequence similarity network (SSN) with a BLASTP e-value threshold of 10^{-80} (~35% sequence identity) (**Figure 2.2a**) [13, 42, 43]. As the BLASTP e-value threshold is decreased to 10^{-190}, the sequences segregate into several clusters sharing >70% sequence identity (**Figure 2.2c**).

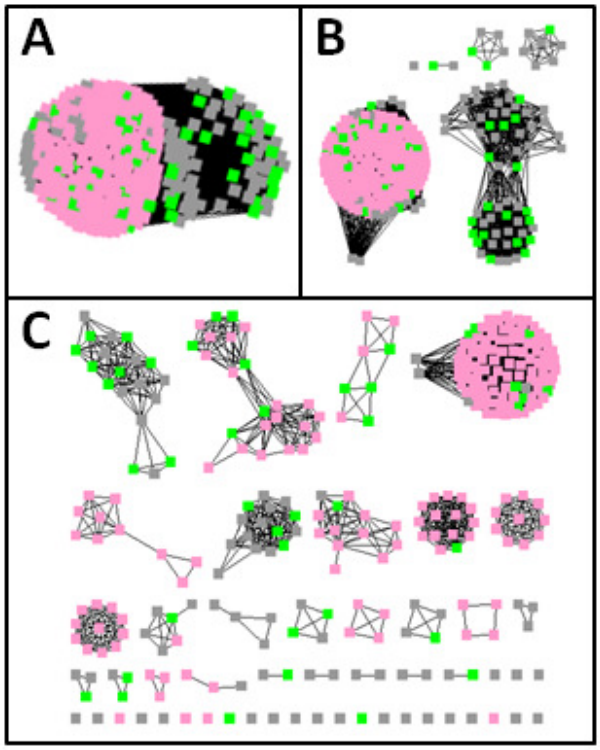


Figure 2.2. **S**equence similarity networks (SSNs) of the ManD subgroup of the ENS at several e-value thresholds to illustrate the effect of increasing stringency on clustering. **Panel A**, 10^{-80}, ~35% identity. **Panel B**, 10^{-120}, ~45% identity. **Panel C**, 10^{-190}, ~75% identity. Pink coloring indicates proteins predicted to be ManDs by the Structure Function Linkage Database. Green coloring indicates proteins that were purified and subjected to activity screening.

The genome neighborhoods of the genes encoding members of the ManD subgroup (± 10 genes) were analyzed to aid in target selection for protein production and structure determination. The genome neighborhoods of some members encode 2-keto-3-deoxy-D-gluconate kinase (KdgK) and 2-keto-3-deoxy-D-gluconate-6-P aldolase (KdgA). KdgK and KdgA metabolize the 2-keto-3-deoxy-D-gluconate product of the ManD reaction to pyruvate and D-glyceraldehyde 3-phosphate, indicating a catabolic role for the proximal member of the ManD subgroup. Alternatively, for some members the genome neighborhoods lack these enzymes but contain, for example, dehydrogenases, suggesting divergent catalytic and metabolic functions. Targets for protein production by the Protein Core of the Enzyme Function Initiative (EFI; enzymefunction.org), functional characterization by the University of Illinois, and structure determination by the Structure Core of the EFI were chosen from both types of genome neighborhoods. A large number of targets (115) were chosen with the anticipation that not all targets would produce soluble, purified proteins.

Substrate Screening. Of the 115 targets, 42 were produced as soluble, purified proteins (sharing less than 95% sequence identity). The ManD SSN in **Figure 2.2** highlights the diversity of purified proteins assayed. The proteins were screened for dehydration activity with a library of 72 acid sugars using a semicarbazide-based assay (**Figure A.1**) [30, 44, 45]. The catalytically active proteins (24 of the 42 screened) utilize D-mannonate, D-gluconate, or both as substrates; no other hits were observed with the acid sugar library. Positive hits were verified with ^{1}H NMR spectra of the products (2-keto-3-deoxy-D-mannonate/2-keto-3-deoxy-D-gluconate) before the proteins were subject to more in-depth analyses to determine kinetic constants. The ability of some members to catalyze the dehydration of D-gluconate was not expected (*vide infra*).

The kinetic characterizations revealed further unexpected divergence in function (**Table 2.1**). Among the newly characterized ManDs, seven dehydrate D-mannonate with catalytic efficiencies similar to that of *Na*ManD (k_{cat}/K_M = 10^3 to 10^4 $M^{-1}s^{-1}$). However, 16 targets showed low catalytic efficiencies (k_{cat}/K_M = 10^1 to 10^2 $M^{-1}s^{-1}$); 19 showed no detectable activity with any member of the acid sugar library. Three of the 12 proteins that dehydrate D-mannonate with low catalytic efficiency also dehydrate D-gluconate with low catalytic efficiency. Furthermore, four of the 23 targets with no activity on D-mannonate dehydrate D-gluconate. Thus, the functionally characterized members were assigned into three categories according to catalytic efficiency and substrate specificity: 1) high-activity (k_{cat}/K_M = 10^3 to 10^4 $M^{-1}s^{-1}$) and specific for D-mannonate; 2) low-activity (k_{cat}/K_M = 10^1 to 10^2 $M^{-1}s^{-1}$) and specific for either D-mannonate or D-gluconate or promiscuous for both; and 3) no-activity with either D-mannonate or D-gluconate (or any acid sugar in the library). The SSN constructed with a threshold of 10^{-190} (~75% sequence identity) segregates groups with different catalytic efficiencies and substrate specificities (**Figure 2.3**).

Divergence in Activity. Members with different *in vitro* activities are assumed to have different *in vivo* functions. Physiologically, the dehydration of D-mannonate to 2-keto-3-deoxy-D-mannonate is found in the D-glucuronate degradation pathway in which D-glucuronate is isomerized to 5-keto-D-mannonate and then reduced to D-mannonate. The D-mannonate is dehydrated, phosphorylated, and cleaved by an aldolase to form pyruvate and glyceraldehyde 3-phosphate. When this pathway was discovered in *E. coli* and *Erwinia carotovora*, dehydration of D-mannonate was found to be catalyzed by a dehydratase, UxuA, that is not a member of the ENS [29, 46]. Therefore, the discovery that members of the ManD subgroup dehydrate D-mannonate with high catalytic efficiency implies convergent evolution of function in different

Table 2.1. Kinetic Parameters of Assayed ENS Mannonate Dehydratases

Cluster	Uniprot ID	D-Mannonate k_{cat} (s^{-1})	D-Mannonate k_{cat}/K_M ($M^{-1}s^{-1}$)	D-Gluconate k_{cat} (s^{-1})	D-Gluconate k_{cat}/K_M ($M^{-1}s^{-1}$)	End of 7th β-strand	UxuA?
1	A5KUH4	---	---	---	---	Pro	Yes
1	C9NUM5	---	---	---	---	Pro	No
1	C9Y5D5	---	---	---	---	Pro	Yes
1	D0KC90	---	---	---	---	Pro	Yes
1	A4W7D6	---	---	---	---	Pro	Yes
1	D0X4R4	---	---	---	---	Pro	Yes
1	A6AMN2	---	---	---	---	Pro	Yes
1	Q6DAR4	---	---	---	---	Pro	Yes
1	C6DI84	---	---	---	---	Pro	Yes
6	B8HCK2	---	---	---	---	Pro	No
9	C9CN91	---	---	---	---	Pro	Yes
9	C8ZZN2	---	---	---	---	Pro	Yes
10	C7PW26	---	---	---	---	Pro	No
Singleton	A6M2W4	---	---	---	---	Pro	Yes
Singleton	Q2CIN0	---	---	---	---	Pro	Yes
Singleton	A8RQK7	---	---	---	---	Pro	Yes
Singleton	C6CVY9	---	---	---	---	Gly	Yes
Singleton	C9A1P5	---	---	---	---	Pro	Yes
Singleton	B5GCP6	---	---	---	---	Pro	No
3	A6VRA1	0.02 ± 0.001	160	---	---	Ala	Yes
3	E1V4Y0	0.03 ± 0.006	20	0.05 ± 0.004	20	Pro	Yes
3	B3PDB1	0.03 ± 0.002	100	---	---	Ala	Yes
3	*Cs*ManD/Q1QT89	0.02 ± 0.0005	5	0.04 ± 0.006	40	Pro	Yes
4	Q8FHC7	0.02 ± 0.001	10	---	---	Pro	Yes
4	A4WA78	0.02 ± 0.002	30	---	---	Pro	Yes
4	B1ELW6	0.02 ± 0.001	20	---	---	Pro	Yes
4	D8ADB5	0.01 ± 0.002	30	---	---	Pro	Yes
4	J7KNU2	0.01 ± 0.001	20	---	---	Pro	Yes
4	B5RAG0	0.01 ± 0.001	50	---	---	Pro	Yes
5	D4GJ14	---	---	0.04 ± 0.003	120	Pro	Yes
5	B5R541	---	---	0.05 ± 0.003	80	Pro	Yes
5	B5QBD4	---	---	0.02 ± 0.0005	150	Pro	Yes
5	C6CBG9	0.04 ± 0.002	50	0.03 ± 0.002	60	Pro	Yes
Singleton	D9UNB2	0.004 ± 0.001	60	---	---	Pro	Yes
8	D7BPX0	---	---	0.01 ± 0.002	40	Pro	No
2	Q1NAJ2	2 ± 0.07	4200	---	---	Ala	No
2	Q9AAR4	1 ± 0.006	12300	---	---	Ala	No
2	Q9A4L8	0.65 ± 0.02	1200	---	---	Ala	No
2	B0T4L2	0.3 ± 0.01	1200	---	---	Ala	No
2	B0T0B1	2 ± 0.2	12100	0.003 ± 0.001	5	Ala	No
2	A5V6Z0	4 ± 0.2	2900	0.01 ± 0.001	10	Ala	No
2	*Na*ManD/A4XF23	1.3 ± 0.1	3200	---	---	Ala	No
7	G7TAD9	0.8 ± 0.03	4400	---	---	Ala	No

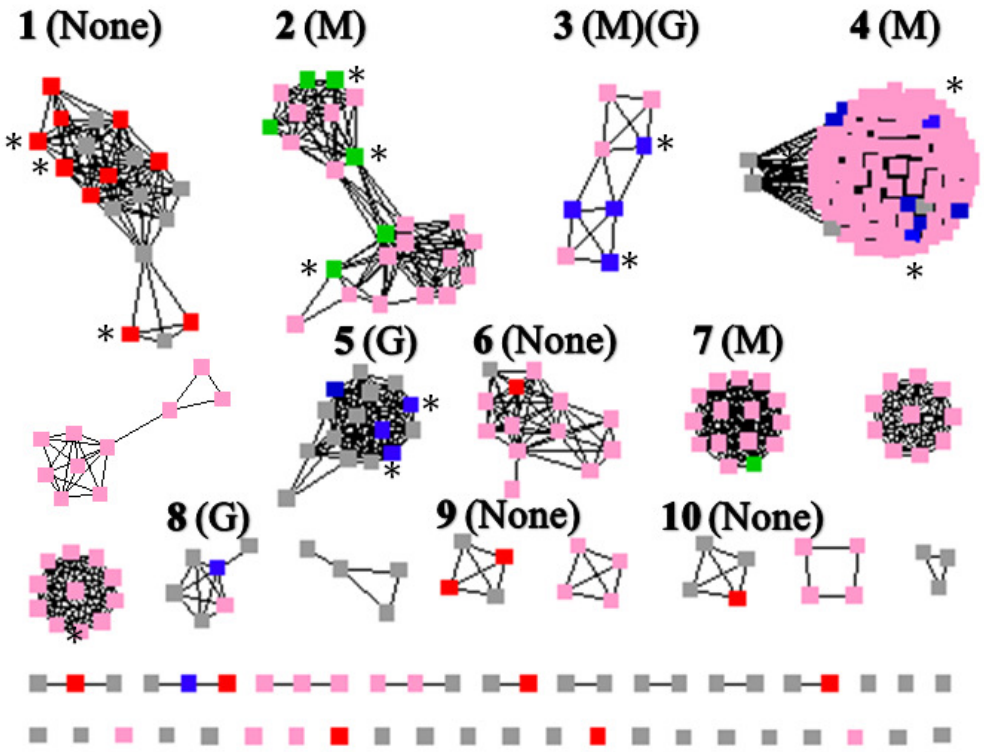


Figure 2.3. SSN showing the distribution of high- (green), low- (blue), and no-activity (red) proteins along with substrate specificities (M, D-mannonate; G, D-gluconate; M/G, D-mannonate and D-gluconate). Proteins for which structures were determined by marked with asterisks. This SSN also shows that the identities of Pro and Ala residues associated with different substrate specificities for D-mannonate and D-gluconate are separated at an e-value threshold of at 10^{-190}: Clusters 1, 4, 5, 6, 8, 9, and 10 contain Pro; Clusters 2 and 7 contain Ala; and Cluster 3 contains both. Pro-containing clusters exhibit low or no dehydration activity; Purely Ala-containing clusters exhibit high dehydration activity with D-mannonate.

superfamilies within the D-glucuronate catabolic pathway. Interestingly, the genomes of all of the organisms encoding high-activity ManDs lack the gene encoding UxuA; however, the genomes of the majority of organisms with low- or no-activity ManDs have a gene encoding UxuA (**Table 2.1**). This suggests that the high-activity ManDs perform the same role as UxuA in the encoding organisms; however, the low-activity members have a different metabolic function. In those organisms that encode low- or no-activity members of the ManD subgroup but no UxuA, growth on D-glucuronate likely is enabled by an alternate catabolic pathway, such as the uronate dehydrogenase/KduI pathway [47, 48]. Therefore, members with low *in vitro* catalytic efficiencies likely have different *in vivo* metabolic functions, even if they can dehydrate D-mannonate. In work that will be described elsewhere, we are characterizing some of these divergent physiological functions.

D-Gluconate Dehydration. Although, in retrospect, the discovery of D-gluconate as a substrate for some members is not surprising because D-gluconate and D-mannonate are epimers at carbon-2 so they yield the same dehydration product, this stereochemical difference requires that a base other than the Tyr-Arg dyad in the "150-180s" loop abstract the 2-proton from D-gluconate. To investigate which residue could function as the D-gluconate specific base, D-mannonate and D-gluconate were modeled into the active site of the member from *Chromohalobacter salexigens* (*Cs*ManD) (Uniprot ID Q1QT89) that dehydrates both D-mannonate and D-gluconate (PDB code 3BSM). This was accomplished by superposing the structures of *Na*ManD with D-mannonate in its active site (2QJM), Uniprot ID B5R541 with D-gluconate in its active site (3TWB), and unliganded *Cs*ManD (3BSM) (**Figure 2.4**). Based on this comparison, we hypothesized that the conserved His after the 7th β-strand is the base in D-gluconate dehydration (His 315 in *Cs*ManD).

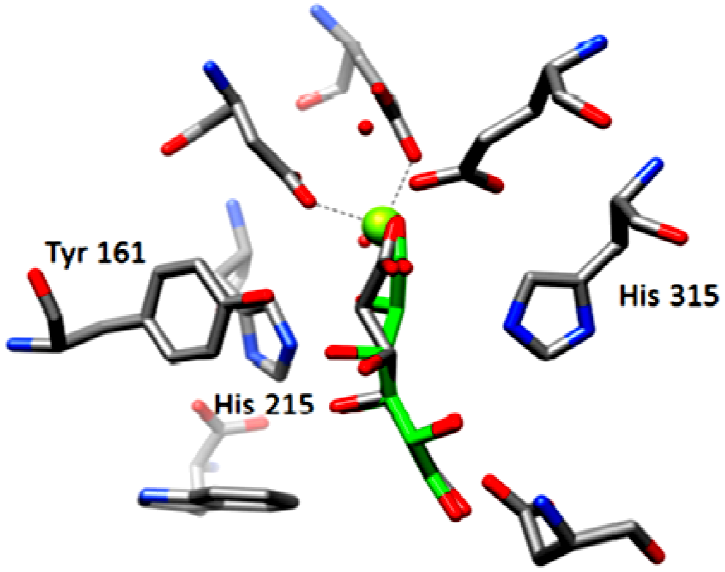


Figure 2.4. A superposition of a structure with D-mannonate bound in the active site (2QJM, *Na*ManD) with one with D-gluconate bound in the active site (3TWB; *Cs*ManD). In 2QJM, Tyr 161 is the general base that abstracts the 2-proton and His 215 is the general acid that catalyzes the departure of the 3-OH group from D-mannonate. In 3TWB, His 315 is proposed to be the general base that abstracts the 2-proton from D-gluconate or hydrogen bonds with the C5 hydroxyl of D-mannonate. The ε-nitrogen of His315 is 3.0Å from the C5 hydroxyl of D-mannonate and 3.1Å from C2 of D-gluconate. Both distances are appropriate for proton abstraction or hydrogen bonding.

Site-directed mutagenesis was performed to convert His 315 in *Cs*ManD to either Asn or Gln. His 315 is conserved in all members of the ManD subgroup, including *Na*ManD, because it hydrogen bonds to 5-hydroxyl group of D-mannonate (**Figure A.2**) [28]. Both mutants abolished dehydration activity with D-gluconate. However, the H315Q mutant maintained wild-type catalytic efficiency with D-mannonate (**Table 2.2**). In contrast, the H315N mutant was inactive

with D-mannonate, presumably because it is not able to hydrogen bond to the 5-hydroxyl group. These studies support the suggested role of His 315 as the base for dehydration of D-gluconate.

Table 2.2. Kinetic Parameters of *Cs*ManD H315 Mutant

	D-Mannonate			D-Gluconate		
Protein	WT	H315Q	H315N	WT	H315Q	H315N
k_{cat}/K_M ($M^{-1}s^{-1}$)	10	10	NA	40	NA	NA

NA = no activity

Structure analysis. "New" crystal structures were solved for 12 sequence diverse members of the subgroup; structures previously were available for 4 other members. Taken together, a total of 36 unliganded and liganded structures are now available for members of the ManD subgroup (**Table A.3**). These structures were used to identify the general base that initiates dehydration of D-gluconate and also yielded insights into how structure diverges as sequence diverges. An overlay of one structure from each structure-containing cluster is shown in **Figure 2.5** (the overlay includes only structures with ordered "150-180s" loops). The structures of the modified TIM-barrel and capping domains are highly conserved, although the conformations of the "150-180s" loop are divergent. The sequences of this loop are also highly variable (**Figure A.3**).

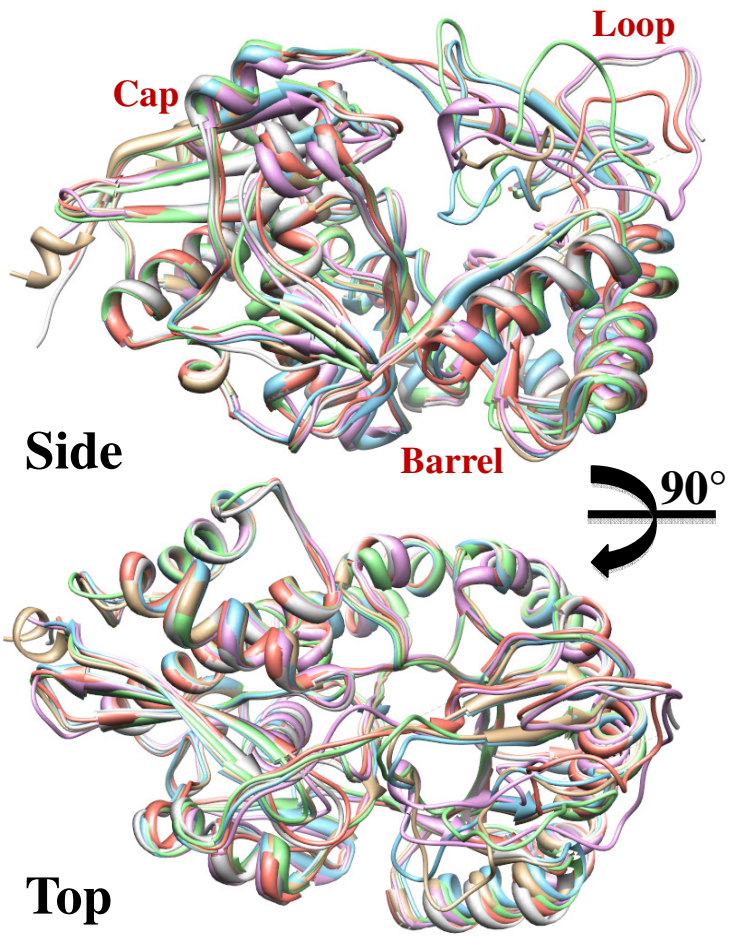
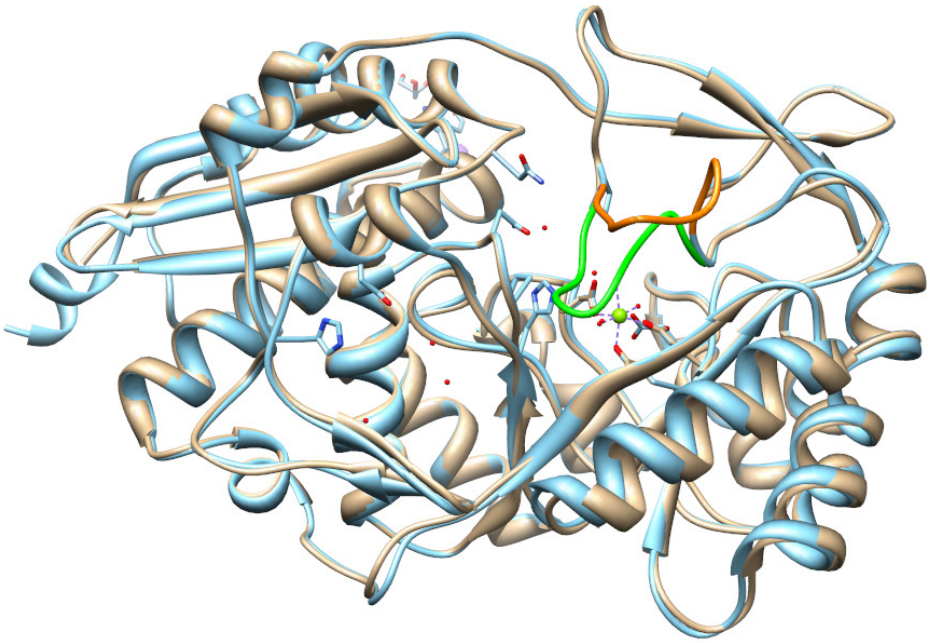


Figure 2.5. An overlay of *Na*ManD (blue), *Cs*ManD (tan), Uniprot ID Q8FHC7 (green), Uniprot ID B5R541 (magenta), Uniprot ID A5KUH4 (red), and Uniprot ID A6M2W4 (grey) showing the overall structural homology. The "150-180s" loops are conformationally distinct.

Initially, the "150-180s" loops were proposed to contain the substrate specificity determinants in analogy with the "20s" loops in other ENS members[49]. Of the 36 structures available (**Table A.3**), 9 have an ordered "150-180s" loop covering the active site, 3 with a substrate bound. Seven of the 11 structures with disordered "150-180s" loops also have a substrate bound. The disorder of the loop, even in the presence of substrate, suggests that the

substrate and the "150-180s" loop are not interacting strongly. In the structures with a bound substrate and an ordered loop, the only hydrogen bonds to the substrate involve ordered water molecules and backbone amide groups. This suggests a role other than determining substrate specificity for the "150-180s" loops.

To determine the importance of the sequence of the "150-180s" loops, segments of the loops in *Cs*ManD (A164 to E172) and *Na*ManD (V161 to E169) that are proximal to the substrate were mutated to AGAGGAGAG (**Figure 2.6**) to eliminate any ionic or hydrogen-bonding interactions with the substrate. Both mutants were expressed and purified as



***Na*ManD WT**	TGVPGIKDAYGVGRGKLYYEPADASLPSVTG
***Na*ManD GA Loop Mutant**	TGVPGIKDAYGAGAGGAGAGPADASLPSVTG
***Cs*ManD WT**	VQSGVPGIETTYGVAKTPGERYEPADSSLPA
***Cs*ManD GA Loop Mutant**	VQSGVPGIETTYGVAGAGGAGAGPADSSLPA

Figure 2.6. An overlay of *Cs*ManD (4F4R, blue ribbon, orange loop) and *Na*ManD (2QJJ, tan ribbon, green loop) showing the regions of their "150-180s" loops that were mutated. The sequences of the loops are given with the area of mutagenesis highlighted with yellow.

soluble proteins; correct folding was verified via circular dichroism and x-ray crystallography (**Figure A.4 and A2.5**). Screening of both mutants with the acid sugar library showed a complete loss of activity. We conclude that this loop is important for catalysis, although its exact contribution is unknown.

Structural differences between D-mannonate/D-gluconate specific ManDs. The structures of proteins with divergent functions are very similar. A superposition of the structures of a high-activity ManD (*Na*ManD), a low activity ManD (Uniprot ID Q8FHC7), a low-activity GlcD (Uniprot ID B5R541), and a member with no activity (Unitprot ID A5KUH4) reveals that the identities and positions of the active site residues are conserved (**Figure 2.7**). In addition, the

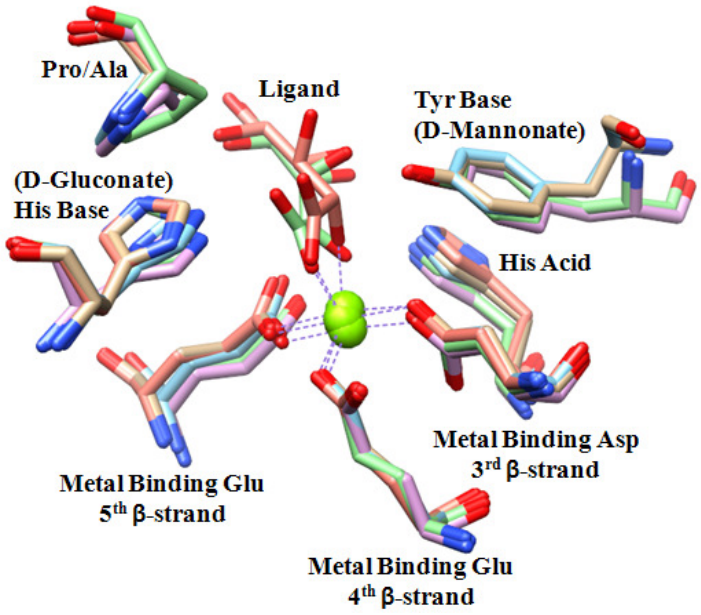


Figure 2.7. An overlay of the active sites of *Na*ManD (2QJJ, high-activity, D-mannonate specific - red), *Cs*ManD (4F4R, low-activity, promiscuous for D-mannonate/D-gluconate - blue), Uniprot ID D4GJ14 (3T6C, low-activity, D-gluconate specific - green), and Uniprot ID A4W7D6 (3TJI, no-activity - magenta). The metal binding and acid/base residues are superimposable. The Pro/Ala dimorphism is also shown. The ligands are D-mannonate from 2QJM (red) and D-gluconate from 3T6C (green).

residues within 6 Å of the substrate are conserved. The primary site of divergence in the active site is a Pro/Ala substitution two residues after the conserved His at the end of the 7th β-strand that hydrogen bonds to the 5-hydroxyl group of the substrate in the ManD reaction or is the general base in the GlcD reaction (**Figure 2.4**). When this substitution is mapped onto the SSN, low- and no activity proteins possess Pro at this position but high-activity ManDs possess an Ala (**Figure 2.3**). The single exception is cluster 3, which is recently separated from cluster 2 (i.e. connected at an e-value of 10^{-184}). Cluster 2 includes high-activity ManDs that contain Ala; cluster 3 contains low-activity proteins that contain either Ala or Pro. Members that contain Pro dehydrate both D-mannonate and D-gluconate, but members with Ala dehydrate only D-mannonate. Two of the 8 high-activity ManDs also dehydrate D-gluconate, but with very low catalytic efficiency. We hypothesize that the proteins in cluster 3 may be intermediates in the evolution of the GlcD function.

This hypothesis was investigated by constructing the *Na*ManD A314P and *Cs*ManD P317A mutants. The A314P mutant of *Na*ManD maintained its specificity for D-mannonate, but with a reduced catalytic efficiency (3200 $M^{-1}s^{-1}$ to 60 $M^{-1}s^{-1}$). The P317A mutant of *Cs*ManD maintained low-activity on D-mannonate and D-gluconate; the catalytic efficiency for D-mannonate is increased (from 5 $M^{-1}s^{-1}$ to 100 $M^{-1}s^{-1}$), but that for D-gluconate is somewhat decreased (from 40 $M^{-1}s^{-1}$ to 5 $M^{-1}s^{-1}$). Thus, we conclude that 1) Ala favors D-mannonate dehydration and disfavors D-gluconate dehydration; and 2) Pro disfavors D-mannonate dehydration. A restriction of backbone flexibility may explain the observed change in substrate specificity and catalytic efficiency, with the flexibility associated with Ala allowing

hydrogen-bonding to the substrate and the rigidity associated with Pro promoting general-base catalysis.

Proteins with no activity. The genes encoding many inactive members of the ManD subgroup are neighbors of genes encoding 2-keto-3-deoxy-D-gluconate kinase and 2-keto-3-deoxy-D-gluconate-6-phosphate aldolase that are downstream of ManD in the D-glucuronate catabolic pathway (**Figure A.6**). However, they have no activity with D-mannonate or the three carbon-2 or -3 epimers that would be dehydrated to 2-keto-3-deoxy-D-mannonate (D-altronate, D-allonate, or D-gluconate). Variations in pH, temperature, salt concentration, osmolarity, osmolytes, and metal cations as well as the presence of dithiothreitol and nucleotide mono-, di-, and triphosphates were tested to determine whether these enzymes may be subject to unanticipated regulation; however, no changes in activity were observed. D-Mannonate 6-phosphate and D-gluconate 6-phosphate also were tested, but no activity was observed. Perhaps these proteins function in multienzyme complexes (channeling) or utilize an acid sugar that is not present in our library.

2.4 Conclusions

The ManD subgroup of the ENS is an excellent example of highly homologous proteins that have divergent catalytic efficiencies and substrate specificities. Unexpectedly, we discovered that the ManD subgroup is not isofunctional: in addition to the ManDs, some members catalyze the dehydration of D-gluconate and others are promiscuous for dehydration of both D-mannonate and D-gluconate. Clearly, automated methods would provide misleading/incorrect annotations that would be of limited/no value in deducing their metabolic functions. We have also determined that the structural determinations of substrate specificity are

both indirect and subtle: a Pro/Ala substitution appears to be the major determination of specificity.

In addition, the role of the sequence divergent and conformationally flexible "150-180s" loop is uncertain. The side chains of the loop makes no direct contacts with the substrate, so the loop does not appear to be a determinant of substrate specificity as has been well-established for members of the muconate lactonizing enzyme (MLE) and mandelate racemase (MR) subgroups in the ENS. Although we attempted to determine whether the loop is involved in protein-protein interactions/substrate channeling, we could not obtain any conclusive results.

We also discovered that the catalytic efficiencies of members of the subgroup are highly variable, despite conservation of active site residues and structures. This data may provide important insights for the metabolic engineering community. The data illustrates how closely related sequences can perform different reactions, and therefore, may help guide studies which aim to redesign proteins for use in new pathways.

Enzymological dogma has been that enzymes have evolved to achieve catalytic perfection, i.e., the reactions are diffusion-controlled. *In vitro* experiments alone can and will not provide biological insights into why the values of k_{cat}/K_M for some of the members of the subgroup are "low". We have selected several of these for future biological/metabolic characterization so that we might be able to better understand the relationship between catalytic efficiencies and metabolic requirements.

CHAPTER 3: Identification of the In vivo Function of the High Efficiency D-Mannonate Dehydratase in *Caulobacter crescentus* NA1000 from the Enolase Superfamily

ABSTRACT: The D-mannonate dehydratase (ManD) subgroup of the enolase superfamily contains members of varying catalytic activities (high efficiency, low efficiency, or no activity) which dehydrate D-mannonate and/or D-gluconate to 2-keto-3-deoxy-D-gluconate (Wichelecki et al. *Biochemistry* 2014, *53*, 2722-2731). Despite extensive *in vitro* characterization, the *in vivo* physiological role of a ManD has yet to be established. In this study, we report the *in vivo* functional characterization of a high efficiency ManD from *Caulobacter crescentus* NA1000 (UniProt ID B8GZZ7) by *in vivo* discovery of its essential role in D-glucuronate metabolism. This *in vivo* functional annotation may be extended to ~50 additional proteins.

Reprinted (adapted) with permission from Wichelecki, D. J., Graff, D. C., Al-Obaidi, N., Almo, S. C., and Gerlt, J. A. Biochemistry. 2014 Jul 1;53(25):4087-9.

3.1. Introduction

The enolase superfamily (ENS) has been the subject of function discovery-based research since it was demonstrated that its homologous members catalyze different chemical reactions[50]. Although all members of the ENS catalyze a common partial reaction (general base-catalyzed abstraction of a proton alpha to a carboxylate group to form a Mg^{2+}-stabilized enediolate intermediate), their overall reactions and chemistries can be quite diverse (e.g., β-elimination and 1,1-proton transfer reactions)[27],[10]. Nevertheless, the members of ENS

share significant structural homology: they share a $(\beta/\alpha)_7(\beta)$-barrel domain for acid/base catalysis and an ($\alpha+\beta$) capping domain for substrate specificity[26, 30, 34, 44, 45, 51].

When the D-mannonate dehydratase (ManD) subgroup of the ENS was discovered in 2007, all of its members were assumed to catalyze the dehydration of D-mannonate to 2-keto-3-deoxy-D-gluconate[28]. The Asp, Glu, and Glu ligands for the essential Mg^{2+} are located at the ends of the 3rd, 4th, and 5th β-strands, respectively, of the barrel domain. The general base catalyst is a Tyr-Arg dyad located at the start of the "150-180s" loop (between the 2nd and 3rd β-strands); the general acid catalyst is a His located at the end of the 3rd β-strand[28].

A large-scale *in vitro* survey of sequence-function space in the ManD subgroup was supported by the Enzyme Function Initiative (EFI; U54GM093342), a large-scale, collaborative, project focused on devising tools and strategies to address and correct the widespread non- and misannotation present in today's automatically curated protein databases[10];[31]. For confident functional/physiological assignment, the EFI seeks not only *in vitro* enzymatic activity (values of k_{cat} and k_{cat}/K_M) but also *in vivo* metabolic function.

The ManD subgroup of the ENS (in Pfam families PF02746 and PF13378) initially was thought to be isofunctional, but the survey of the sequence space revealed dehydration activity with D-mannonate and/or D-gluconate[52]. The members of the ManD subgroup have a range of catalytic activities and efficiencies: high efficiency (specific to D-mannonate with a k_{cat}/K_M of 10^3 to 10^4 $M^{-1}s^{-1}$), low efficiency (D-mannonate and/or D-gluconate with a k_{cat}/K_M of 10^1 to 10^2 $M^{-1}s^{-1}$), and no activity[52]. However, the *in vivo* physiological function had not been assigned to any member of the ManD subgroup. The only metabolic pathway in which dehydration of D-mannonate is known to occur is the catabolism of D-glucuronate (Figure 1), in which the

ManD reaction is performed by the well-characterized UxuA (**u**tilization of he**xu**ronate **A**)[29],[53] that is included in Pfam family PF03786.

One high efficiency ManD (Uniprot ID B8GZZ7; PDB 3VCN and 4GME) was shown to be specific for D-mannonate with a catalytic efficiency of 1.2 x 10^4 $M^{-1}s^{-1}$ [52]. The UxuAs in D-glucuronate catabolism have catalytic efficiencies of ~10^3 $M^{-1}s^{-1}$. Interestingly, low efficiency/no activity ManDs are encoded by genomes that encode a member of the UxuA family; however, genomes that encode a high efficiency ManD do not encode a UxuA[52]. This suggested that high efficiency ManDs are analogs for UxuAs, with low efficiency/no activity ManDs likely having different physiological functions.

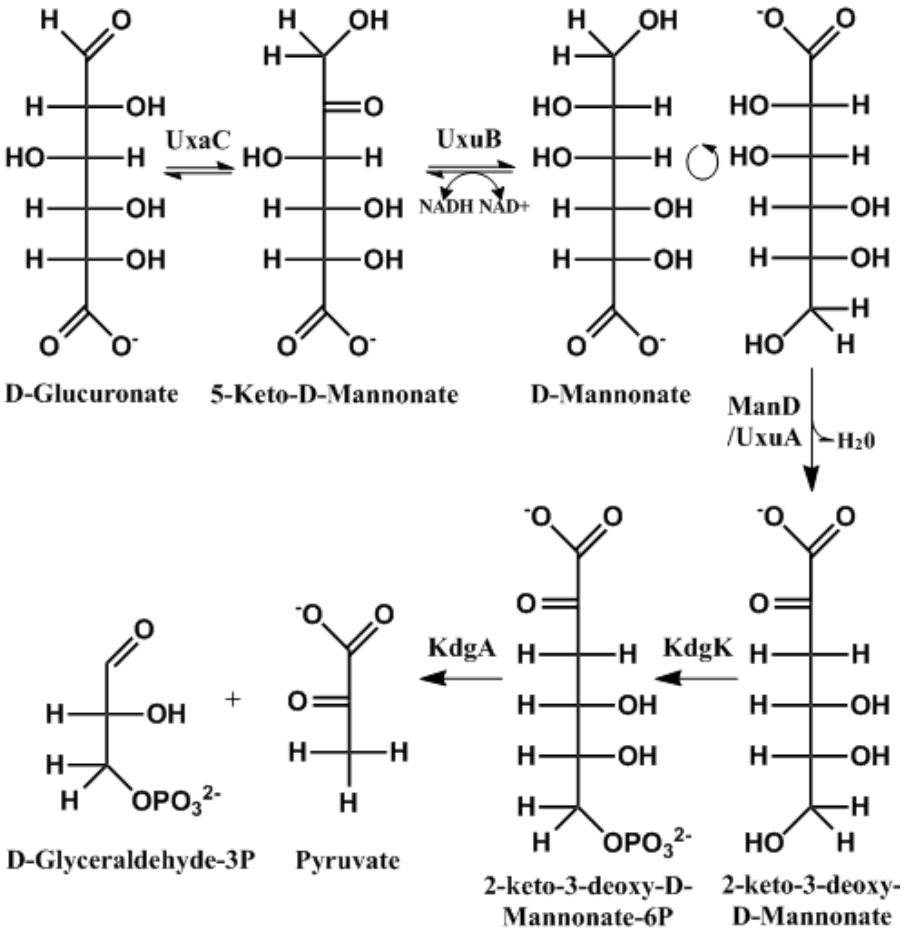


Figure 3.1. Degradation pathway of D-glucuronate in *E. coli*. The dehydration of D-mannonate to 2-keto-3-deoxy-D-gluconate is performed by UxuA.

Of the high efficiency ManDs that were characterized *in vitro*, we selected B8GZZ7 for *in vivo* study because its encoding organism, *Caulobacter crescentus* NA1000, is amenable to genetic manipulation[54]. *Caulobacter crescentus* is an important organism in the study of asymmetric cell division and prokaryotic development because it can differentiate into mobile swarmer cells and immobile stalk cells[54]. Species of *Caulobacter* utilize aquatic vegetation as carbon source in freshwater environments[54]; D-glucuronate makes up a large part of plant cell-

wall biomass[55]. Therefore, we hypothesized that *Caulobacter crescentus* NA1000 could utilize D-glucuronate as a carbon source via B8GZZ7, a high efficiency ManD, despite the absence of UxuA.

3.2 Materials and methods

qRT-PCR of B8GZZ7/Mannosidase Genes. Caulobacter crescentus NA1000 (culture from Smit lab, University of British Columbia) was grown to an OD of 0.4 - 0.5 in M2 minimal salts media (6.1 mM Na_2HPO_4, 3.9 mM KH_2PO_4, 9.3 mM NH_4Cl, 0.5 mM $MgSO_4$, 10 μM $FeSO_4$, 0.5 mM $CaCl_2$) and 10 mM D-glucuronate (Sigma-Aldrich) or 10 mM D-glucose (Sigma-Aldrich). 500 μl of cells were incubated in 1 ml of RNA Protect (Qiagen) for 5 min at room temperature. Cells were pelleted at 15,000 rpm and the supernatant was removed. mRNA was then purified from the cells using an RNeasy Mini Kit (Qiagen). The RNA was cleaned up using RNase-free DNase (Qiagen) following the manufacturer's protocol. Purity was verified using 30 μL PCR reactions consisting of 50 ng mRNA, 1 mM $MgCl_2$, 1X Pfx amplification buffer, 2X PCR enhancer, 0.33 mM dNTP, 0.33 μM of FOR/REV primers (B8GZZ7_RTPCR_FOR and B8GZZ7_RTPCR_REV), and 1.25 units Pfx polymerase (Invitrogen Platinum Pfx DNA Polymerase kit). The PCR reaction was run on a gel to check for amplification. The concentration was determined spectrophotometrically at 280 nm. cDNA was made using Protoscript First Strand (New England Biolabs) and 1 μg mRNA. Manufacturer's protocol was followed. The qRT-PCR reaction was performed using the Light Cycler 480 SYBR Green I Master Kit (Roche) and a Light Cycler 480 II (Roche) per the manufacturer's protocol. **Table 1** contains sequences of the qRT-PCR primers (labeled as RTPCR). The RpoD (RNA

polymerase sigma factor) primers were included to normalize the results based on the amount of cDNA added.

Construction of the Knockout of the Gene Encoding Q9AAR. The knockout of the gene encoding B8GZZ7 in *C. crescentus* NA1000 was constructued using overlap extension and a suicide vector (pK19mobsacB, from ATCC# 87097)[56]. 600 p upstream and downstream of the gene encoding B8GZZ7 (Locus tag CC_0532) start and stop codons was amplified via 30 μL PCR reactions consisting of 50 ng *C. crescentus* NA1000 genomic DNA, 1 mM $MgCl_2$, 1X Pfx Amp Buffer, 2X PCR enhancer 0.33 mM dNTP, 0.33 μM of FOR/REV primers [**Table 1**: B8GZZ7_BAMHI_FOR (upstream), B8GZZ7_BAMHI_Extend_REV (upstream),

Table 3.1. qRT-PCR primers used in this study

Primer name	Sequence
B8GZZ7_RTPCR_FOR	5'- CGACCGCCAAGTATCTCAAC -3'
B8GZZ7_RTPCR_REV	5'- AGCCAGAACAGGCGATAGG -3'
Mannosidase_RTPCR_FOR	5'- GCCAATCTGGACGAGGAAG -3'
Mannosidase_RTPCR_FOR	5'- ATGGTGTGGGAGCGATAGAG -3'
RpoD_RTPCR_FOR	5'- CTGATGCCGTAGCTGTCGG -3'
RpoD_RTPCR_REV	5'- GGCCTATGAAGGCCTGTCG -3'
UxaC_RTPCR_FOR	5'- ACGTGTTCAGCCAGGTGTTC -3'
UxaC_RTPCR_REV	5'- GTGATGCTGAAGGCTCTCGT -3'
UxuB_RTPCR_FOR	5'- GAAACCGCGCACATCATC -3'
UxuB_RTPCR_REV	5'- CATGGTCGACAGCATCACAC -3'
KdgK_RTPCR_FOR	5'- CCGGTGAAGTGAACTGGGAAA -3'
KdgK_RTPCR_REV	5'- GCGGTAGTTCAGGTCGTAG -3'
KdgpA_RTPCR_FOR	5'- CGAGATCGTCAAGGTGTTCC -3'
KdgpA_RTPCR_REV	5'- CTTGGCGTCCAGCATCTCT -3'
B8GZZ7_BamHI_FOR	5'- CAAATGGCTCGACCACCGGATCCACGGTCACTATTCGCAG -3'
B8GZZ7_BamHI_Extend_REV	5'- GGCCGCGCCCCGGCCGACAGGGGGGGACCTGGCGTTGCTGAGTGCGGCC-3'
B8GZZ7_EcoRI_REV	5'- CGACGCATACGGATTCTCATGACGGGCGC -3'
B8GZZ7_EcoRI_Extend_FOR	5'- GGCCGCACTCAGCAACGCCAGGTCCCCCCCTGTCGGCCGGGGCGCGGCC-3'
B8GZZ7_NdeI_FOR	5'-CTCAGCAACGCCAGGCATATGCTCAAGATCATCG-3'
B8GZZ7_BamHI_REV	5'-CCCCGGATCCCAGGGGGTTACCAGTTGAAC-3'

B8GZZ7_EcoRI_REV (downstream), B8GZZ7_EcoRI_Extend_FOR (downstream)], and 1.25 units Pfx polymerase (Invitrogen Platinum Pfx DNA Polymerase kit). These amplifications contained homologous extension regions. These upstream and downstream fragments were then PCR amplified again with the same reaction mixture except the 50ng of each fragment was used as template and B8GZZ7_BAMHI_FOR and B8GZZ7_EcoRI_REV primers were used. The product was digested by BamHI and EcoRI (New England Biolabs) and ligated into BamHI- and EcoRI-digested pK19mobsacB to form pK19mobsacB:B8GZZ7. Successful ligations were transformed into *E. coli* XL1-Blue for storage. The knockout of the gene encoding B8GZZ7 was constructued via homologous recombination by transforming in pK19mobsacB: B8GZZ7 into *C. crescentus* NA1000 via electroporation[57]. PYE plates (0.2% peptone, 0.1% yeast extract, 0.01% CaCl2, 0.02% MgSO4, and 1.2% agar) containing kanamycin (50 μg/ml) were used to identify single-crossover events. Four rounds of subculturing were used to encourage double-crossover events. Double-crossover events were identified by growing the subcultured *C. crescentus* on PYE plates with 10% sucrose. The colonies were then PCR amplified using the previous PCR mixture, but using cells as template. B8GZZ7_BAMHI_FOR and B8GZZ7_EcoRI_REV primers were used. The resulting amplification was sequenced for verification of the knockout.

Complementation of the Knockout of the Gene Encoding B8GZZ7. B8GZZ7 was amplified from 50 ng *C. crescentus* NA1000 genomic DNA using the same reaction mixture as in the qRT-PCR methods section and the primers B8GZZ7_NdeI_FOR and B8GZZ7_BamHI_REV. The amplification product was digested with NdeI and BamHI (New England Biolabs) and ligated into NdeI/BamHI-digested pSRK-Kan[58] (Farrand laboratory, UIUC). Successful ligations were transformed into *E. coli* XL1-Blue for storage. For

complementation, successful ligations were transformed into wild type *C. crescentus* and the ΔB8GZZ7 strain as previously described.

C. crescentus NA1000 Wildtype, ΔB8GZZ7, and Complementation Growth Curves. Wildtype *C. crescentus* NA1000, the ΔB8GZZ7 strain, the B8GZZ7-complemented ΔB8GZZ7 strain, and the ΔB8GZZ7-complemented wild type strain were grown to an optical density of 0.4 to 0.5 in PYE media. The cells were pelleted and resuspended in M2 minimal salts media then used to inoculate cultures of M2 minimal salts media containing 10 mM D-glucuronate and 500 μM IPTG (Sigma-Aldrich) with a normalized amount of cells. Growth curves were made using a Bioscreen C instrument (Growth Curves USA). Cells were grown at 30°C for 120 hours with continuous shaking.

Fluorescence–based thermal shift assay (ThermoFluor). The library consists of 188 ligands in duplicate wells, plus 8 control wells (protein without compound) (**Table B.2**). Compound stocks are dissolved in ThermoFluor running buffer (100 mM HEPES pH 7.5, 150 mM NaCl) to a concentration of 1 mM. The protein of interest is diluted to 10 μM in 5x SYPRO orange (Sigma-Aldrich). 1 μl of the 5x SYPRO orange protein mix is added to 9 μl of compound dissolved in ThermoFluor running buffer in a 384 clear well plate. Using "Applied Biosystems 7900HT Fast Real-Time PCR System" (Life Technologies) the temperature was increased from 22 to 99°C at a rate of 3°C/min. As the protein is melting under the effect of increasing temperature, the machine will scan the fluorescence signal from each well at multiple time points creating a point curve for each reaction. Using the start point and peak point of melting temperature, we calculate the middle of the curve point using a Boltzmann fit for each reaction (T_M value). We subtract that value from the average of the 8 control well's T_M to get the delta T_M

for each well. A hit is considered significant if the ΔT_M is greater than or equal to 4°C or 3 times the range value of the control wells T_M.

3.3 Results and Discussion

The gene encoding B8GZZ7 is located in a genome neighborhood that is not indicative of D-glucuronate metabolism. Nearby genes encode catabolic enzymes for N-acetyl-glucosamine-6-phosphate, a mannosidase, and a GntR transcriptional regulator (Figure 2). Nevertheless, the

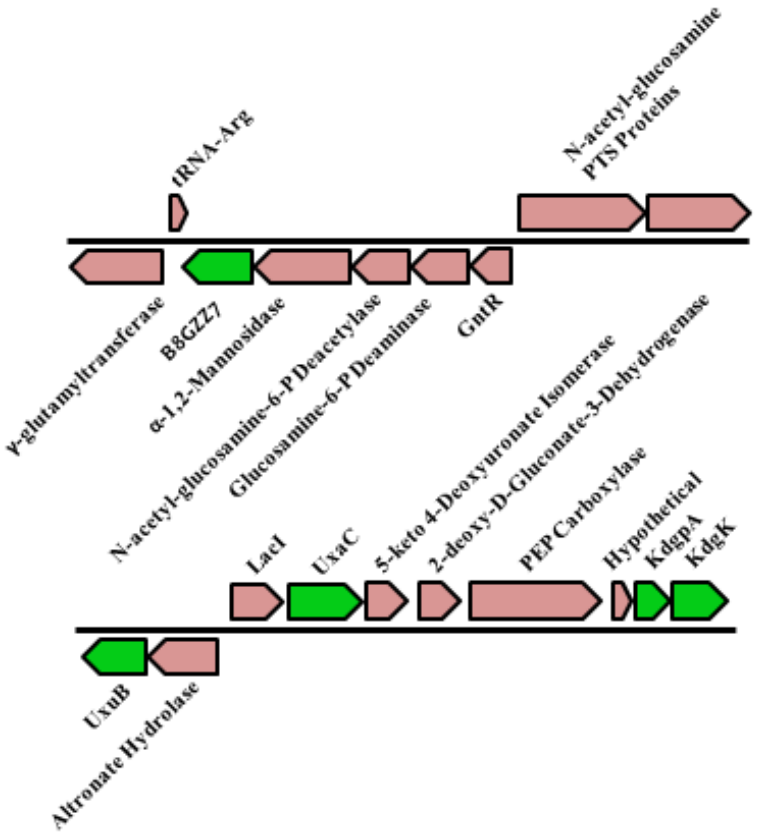


Figure 3.2. The genome neighborhoods of B8GZZ7, a high efficiency ManD, (top) and canonical D-glucuronate catabolism genes (bottom). The genes directly involved in D-glucuronate metabolism are highlighted in green.

ability of D-glucuronate to up-regulate the gene encoding B8GZZ7 was investigated via qRT-PCR. Expression of the gene encoding B8GZZ7 was induced > 1000-fold when *C. crescentus* NA1000 was grown on D-glucuronate as carbon source (relative to growth on D-xylose) (Figure 3). Importantly, the gene encoding the mannosidase immediately upstream of the gene encoding

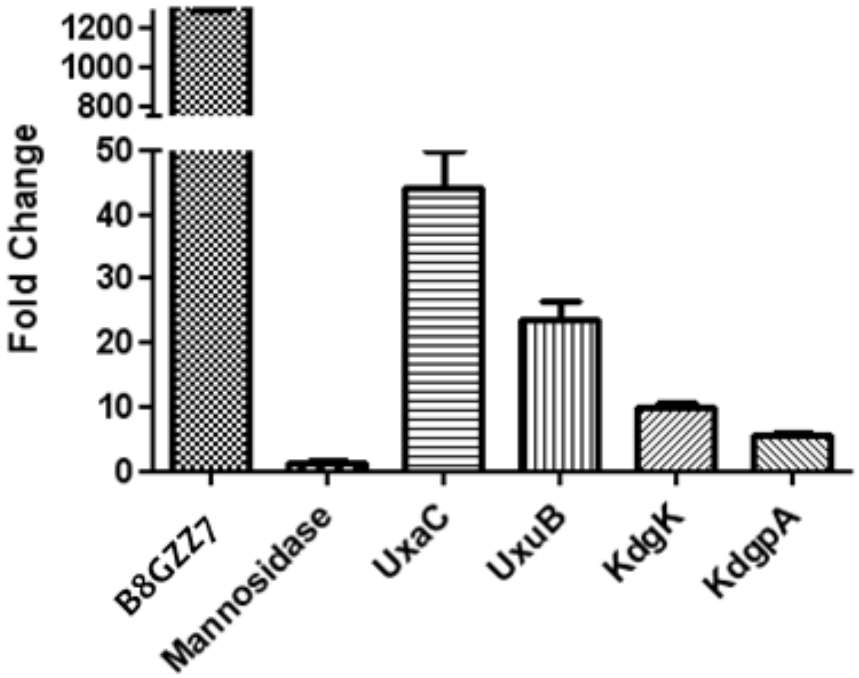


Figure 3.3. Up-regulation of the genes encoding B8GZZ7 and the D-glucuronate catabolism genes shown in Figure 2.

B8GZZ7 was not upregulated (Figure 3); therefore, the gene encoding B8GZZ7 is transcribed separately from its genome neighbors and is likely not involved in N-acetyl-glucosamine catabolism. Absence of co-transcription of the genome neighborhood was further verified by the lack of a PCR product when cDNA was amplified using primers inside of the genes encoding both B8GZZ7 and the mannosidase (such that the potential product would span the gap between the two genes). Thus, we conclude that genomic context can be misleading for functional

assignment in the absence of definitive information on regulation and the identities of transcriptional units.

ThermoFluor[14] ligand binding studies were performed on the GntR transcriptional regulator in the genome neighborhood of B8GZZ7 using a library of acid sugars, including D-glucuroante and D-mannonate, as well as other metabolites (Table B.2). This assay yielded no stabilizing ligands, providing further evidence that the gene encoding B8GZZ7 is regulated separately.

The genes encoding the remaining enzymes in the canonical D-glucuronate pathway are located elsewhere in the genome, closely clustered with other sugar metabolizing genes (Figure 2). The genes encoding D-glucuronate isomerase (UxaC), fructuronate reductase (UxaB), 2-keto-3-deoxy-D-gluconate kinase (KdgK), and 2-keto-3-deoxy-D-gluconate-6-phosphate aldolase (KdgpA) were upregulated when grown on D-glucuronate (relative to growth on D-xylose) (Figure 3). These qRT-PCR data provide compelling evidence that B8GZZ7 is performing the ManD reaction in the canonical D-glucuronate catabolic pathway.

The knockout of the gene encoding B8GZZ7 showed no growth on D-glucuronate (Figure B.1). Complementation of the knockout with an IPTG-inducible gene encoding B8GZZ7 (pSRK-Kan)[58] restored growth on D-glucuronate. Although the growth rate of the complemented knockout strain was slower than that of wild-type *C. crescentus* NA1000 on D-glucuronate, the wild-type strain transformed with the same complementation vector showed a similar growth defect (Figure B.1), likely due to toxicity of plasmid-based expression of B8GZZ7. The requirement of B8GZZ7 for growth on D-glucuronate and complementation of the growth defect by supplying B8GZZ7 *in trans* verifies that B8GZZ7 is the ManD for D-glucuronate metabolism in *C. crescentus* NA1000.

We visualize functional diversity in the ManD subgroup using sequence similarity networks[52]. Using an E-value threshold of 10^{-190}, corresponding to ~75% identity, the clusters are isofunctional as judged by a sequence alignment[52]. B8GZZ7 is a member of an isofunctional cluster that includes 55 additional proteins (Figure B.2 and Table A3.3). None of the organisms that encode these proteins encode the UxuA in D-glucuronate catabolism, although the genes encoding all of the other enzymes needed for catabolism of D-glucuronate via the pathway shown in Figure 1 are present in the genomes.

The genomes of a few of these organisms, including *C. crescentus* NA1000, encode two high efficiency ManDs. In *C. crescentus* NA1000, the second ManD (UniProt ID B8H1R9) is not involved in D-glucuronate catabolism: the knockout of the gene encoding the B8H1R9 retains the ability to utilize D-glucuronate as carbon source (Figure B.1); also, the gene encoding B8H1R9 is not upregulated by D-glucuronate. Apparently, the ManD activity of B8H1R9 is essential for another metabolic function that is currently unknown.

We are confident that we can transfer the now verified physiological function of B8GZZ7 to the high efficiency ManDs that are encoded by organisms with a single orthologue—all of these organisms encode the other enzymes in D-glucuronate catabolism. For those organisms that encode two high efficiency ManDs, genome context is sufficient to assign function in some cases, e.g., *Sphinogomonas wittichii* and *Sphingomonas* sp. MM-1; in other cases, such as *C. crescentus* NA1000 described in this manuscript, transcriptomic and genetic experiments are needed to assign the physiological function.

3.4. Conclusions

Previously, no physiological function for a ManD in the ENS had been determined. We now have used *in vivo* approaches to establish that B8GZZ7, a high efficiency ManD in *Caulobacter cresecentus* NA1000, participates in D-glucuronate catabolism. Thus, we conclude that the high efficiency ManDs in the ENS are examples of convergent evolution of function with the ManDs in the UxuA family.

This study demonstrates that 1) the *in vitro* enzymatic activity alone cannot be used to infer *in vivo* physiological function for an "unknown" enzyme discovered in a genome project, and 2) genome context also is not sufficient to infer the *in vivo* physiological function. In due course, we will report the *in vivo* physiological functions for members of the ManD subgroup that have low catalytic efficiencies for dehydration of D-mannonate and D-gluconate.

CHAPTER 4: INVESTIGATING THE PHYSIOLOGICAL ROLES OF LOW-EFFICIENCY D-MANNONATE AND D-GLUCONATE DEHYDRATASES IN THE ENOLASE SUPERFAMILY: PATHWAYS FOR THE CATABOLISM OF L-GULONATE AND L-IDONATE

ABSTRACT: The sequence/function space in the D-mannonate dehydratase subgroup (ManD) of the enolase superfamily was investigated to determine how enzymatic function diverges as sequence identity decreases (Wichelecki et al. *Biochemistry* 2014, *53*, 2722-2731). That study revealed that members of the ManD subgroup vary in substrate specificity and catalytic efficiency: high-efficiency (k_{cat}/K_M = 10^3 to 10^4 $M^{-1}s^{-1}$) for dehydration of D-mannonate, low-efficiency (k_{cat}/K_M = 10^1 to 10^2 $M^{-1}s^{-1}$) for dehydration of D-mannonate and/or D-gluconate, and no-activity. Characterization of high-efficiency members revealed that these are ManDs in the D-glucuronate catabolic pathway (analogs of UxuA; Wichelecki et al. *Biochemistry* 2014, *53*, 4087-4089). However, the genomes of organisms that encode low-efficiency members of the ManDs subgroup encode UxuAs; therefore, these must have divergent physiological functions. In this study, we investigated the physiological functions of three low-efficiency members of the ManD subgroup and identified a novel physiologically relevant pathway for L-gulonate catabolism in *Chromohalobacter salexigens* DSM3043 as well as cryptic pathways for L-gulonate catabolism in *Escherichia coli* CFT073 and L-idonate catabolism in *Salmonella enterica subsp. enterica serovar Enteritidis* str. P125109. However, we could not identify physiological roles for the low-efficiency members of the ManD subgroup, allowing the suggestion that these pathways may be either evolutionary relics or the starting points for new metabolic potential.

Reprinted (adapted) with permission from Wichelecki, D. J., Vendiola, J. A. F., Al-Obaidi, N., Almo, S. C., and Gerlt, J. A. Accepteded to Biochemistry August 2014.

4.1. Introduction

As the protein databases continue to expand (release 2014_07 of UniProt contains 80,370,243 nonredundant sequences) and, in parallel, the homologous families identified by Pfam increase in size and functional diversity (Pfam release 27.0 describes 14,831 families and 515 clans), the problem of assigning function based on sequence identity/similarity is becoming increasingly more challenging. For the past 25 years, we have developed the enolase superfamily (ENS) as a paradigm to characterize and understand the structural bases for divergence of function in functionally diverse superfamilies[28],[30],[34],[23]. The members of the ENS share not only a conserved structure, an (α+β) capping domain that determines substrate specificity and a modified TIM-barrel domain [$(\beta/\alpha)_7\beta$-barrel] that contains the acid/base residues that determine the reaction mechanism, but also a conserved mechanism that is initiated by abstraction of a proton from the carbon adjacent to a carboxylate group to generate a Mg^{2+}-stabilized enolate anion intermediate[25],[26]. Our recent studies have focused on studying how function diverges as sequence diverges in homologous subgroups[59].

The D-mannonate dehydratase subgroup (ManD) of the ENS recently was found to include members that dehydrate D-mannonate and/or D-gluconate to 2-keto-3-deoxy-D-gluconate (equivalently, 2-keto-3-deoxy-D-mannonate)[28],[59]. [The subgroup was named by the function of the first member that was functionally and structurally characterized, when (2007), the protein sequence databases were considerably smaller than they are today[28].] Also, the members of the ManD subgroup display varying catalytic efficiencies: high-efficiency

(specific for D-mannonate with a k_{cat}/K_M of 10^3 to 10^4 $M^{-1}s^{-1}$), low-efficiency (D-mannonate and/or D-gluconate with a k_{cat}/K_M of 10^1 to 10^2 $M^{-1}s^{-1}$), and no activity[59]. Although we recently reported the physiological role of the high-efficiency ManDs[60], dehydration of D-mannonate in the D-glucuronate catabolic pathway in organisms that lack UxuA, the roles of the low-efficiency members of the ManD subgroup have not been known.

In this article, we describe *in vitro* characterization of two metabolic pathways that are encoded by the genes that are proximal to those that encode low-efficiency members of the ManD subgroup. The pathways utilize two successive dehydrogenase reactions (oxidation followed by reduction) to catalyze epimerization of carbon-5 of a six-carbon acid sugar, one for conversion of L-gulonate to D-mannonate (*Escherichia coli* CFT073 and *Chromohalobacter salexigens* DSM3043) and the second for conversion of L-idonate to D-gluconate (*Salmonella enterica subsp. enterica serovar Enteritidis* str. P125109). *In vivo* studies using genetic deletions (knockouts) and transcriptomics established that the L-gulonate pathway in *C. salexigens* DSM3043 allows growth on L-gulonate as carbon source. However, we were unable to demonstrate a physiological requirement for the ManD (Uniprot ID Q1QT89; *Cs*ManD[61]) that catalyzes the dehydration of D-mannonate (k_{cat}/K_M, 5 M^{-1} s^{-1}) and D-gluconate (k_{cat}/K_M of 40 M^{-1} s^{-1})[59]. We also describe cryptic pathways for the catabolism of L-gulonate and L-idonate in *E. coli* CFT073 and *S. enterica subsp. enterica serovar Enteritidis* str. P125109, respectively, that are encoded by the same genome neighborhoods that encode the low-efficiency ManD and GlcD, respectively, that could participate in the pathways.

4.2 Materials and methods

Cloning, expression, and purification. The genes encoding the L-gulonate 5-dehydrogenase from *Halomonas elongata* DSM 2581 (*He*GulDH, Uniprot ID E1V4Y1, an orthologue of the L-gulonate 5-dehydrogenase from *C. salexigens* DSM3043, *Cs*GulDH), the L-idonate dehydrogenase from *S. enteriditis* (*Se*IdoDH, Uniprot ID B5R538), and the D-gluconate dehydrogenase from *S. enteriditis* (*Se*GlucDH, Uniprot ID B5R540) were synthesized (Genscript) and codon-optimized for expression in *Escherichia coli*; the genes were synthesized with 5'-NdeI and 3'-BamHI restriction sites and were received in the vector pUC57. The pUC57 constructs were transformed into *E. coli* XLI Blue for subcloning and storage. Purified vector was NdeI/BamHI (New England Biolabs) digested and ligated into NdeI/BamHI digested pET15b (Novagen). The pET15b constructs were transformed into *E. coli* BL21 (DE3) for expression. The proteins were purified from 1 L cultures using a chelating Sepharose Fast Flow (Amersham Biosciences) column charged with Ni^{2+} as previously described[59], concentrated (*He*GulDH to 15 mg/mL, *Se*IdoDH to 8.1 mg/mL, and *Se*GlcDH to 4.7 mg/mL), flash-frozen using liquid nitrogen, and stored at -80 ˚C prior to use.

The gene encoding RspB, the L-idonate 5-dehydrogenase from *E. coli* CFT073 (Uniprot ID Q8FHC8), was amplified from *E. coli* CFT073 genomic DNA (ATCC 700928) using RspB_FOR and RspB_REV primers (**Table C.1**). The PCR (30 μL) contained 50 ng template, 1 mM $MgCl_2$, 1X Pfx Amp Buffer, 0.33 mM dNTP, 0.33 μM of each primer, and 1.25 units Pfx polymerase (Invitrogen Platinum Pfx DNA Polymerase kit). Amplifications were performed according to the manufacturer's guidelines. The amplified product was digested with NdeI/BamHI (New England Biolabs) and ligated into NdeI/BamHI digested pET15b (Novagen). The pET15b *RspB* construct was transformed into *E. coli* BL21 (DE3) for expression. RspB was

purified from a 1 L culture using a chelating Sepharose Fast Flow (Amersham Biosciences) column charged with Ni^{2+} as previously described[59]. The protein was concentrated to 4.2 mg/mL, flash-frozen using liquid nitrogen, and stored at -80 ˚C prior to use.

The gene encoding RspD, the D-mannonate 5-dehydrogenase from *E. coli* CFT073 (Uniprot ID Q8FHD0), was amplified from *E. coli* CFT073 genomic DNA (ATCC 700928) using RspD_FOR and RspD_REV primers (**Table C.1**). The PCR (30 µL) contained 50 ng template, 1 mM $MgCl_2$, 1X Pfx Amp Buffer, 0.33 mM dNTP, 0.33 µM of each primer, and 1.25 units Pfx polymerase (Invitrogen Platinum Pfx DNA Polymerase kit). Amplifications were performed according to the manufacturer's guidelines. The amplified product was digested with SacI/BamHI (New England Biolabs) and ligated into SacI/BamHI digested pET17b (Novagen). The pET17b *RspD* construct was transformed into *E. coli* BL21 (DE3) for expression. RspD was purified from a 1 L culture using DEAE Sepharose, Q-Sepharose column, and phenyl sepharose columns (all Amersham Biosciences) as previously described[59]. The protein was concentrated to 13 mg/mL, flash-frozen using liquid nitrogen, and stored at -80 ˚C prior to use.

The gene encoding *Cs*GntR, the GntR transcriptional factor in the genome neighborhood of *Cs*ManD (Uniprot ID Q1QT90), was amplified from *C. salexigens* DSM3043 genomic DNA (ATCC BAA-138) using CsGntR_FOR and CsGntR_REV primers (**Table C.1**). The PCR was the same as for RspD described above but with the CsGntR primers. The amplified product was digested with NdeI and BamHI (New England Biolabs) and ligated into similarly digested pET15b. The pET15b *Cs*GntR construct was transformed into *E. coli* BL21 (DE3) for expression. *Cs*GntR was purified from a 1 L culture using a chelating Sepharose Fast Flow (Amersham Biosciences) column charged with Ni^{2+} as previously described[59]. The protein

was concentrated to 12 mg/mL, flash-frozen using liquid nitrogen, and stored at -80 ˚C prior to use.

*Cs*UxuA was obtained from Albert Einstein College of Medicine as previously described[59].

Screen for Oxidation Activity of HeGulDH, SeGlcDH, SeIdoDH, RspB, and RspD. Reactions to test for oxidation activity were performed in acrylic, UV transparent 96-well plates (Corning Incorporated) using a library of 72 acid sugars (**Figure C.1**). Reactions (60 μL) contained 50 mM HEPES, pH 7.9, 10 mM $MgCl_2$, 1 mM NAD^+, 1 μM enzyme, and 1 mM acid sugar substrate (blanks without enzyme). The plates were incubated at 30 ˚C for 16 hrs. The absorbancies were measured at 340 nm ($\varepsilon = 6220\ M^{-1}\ cm^{-1}$) using a Tecan Infinite M200PRO plate reader. Products were verified via ^1H NMR.

Screen for Dehydration Activity of CsUxuA . Reactions to test for dehydration activity were performed in acrylic, UV transparent 96-well plates (Corning Incorporated) using a library of 72 acid sugars (**Figure C.1**) as previously described[59].

Kinetic Assays for HeGulDH, SeIdoDH, RspB, and RspD. Oxidation of L-gulonate, L-idonate, or D-mannonate was monitored using a continuous spectrophotometric assay. For HeGulDH, the assays (200 μL), at 25 °C, contained 50 mM potassium HEPES, pH 7.9, 5 mM $MnSO_4$, 3 mM NAD^+, and 200 nM *He*GulDH. For the remaining enzymes, the assays (200 μL), at 25 °C, contained 50 mM potassium HEPES, pH 7.9, 5 mM $MgCl_2$, 3 mM NAD^+, and either 200 nM RspB, 200 nM *Se*IdoDH, or 20 nM RspD (lower concentration because of greater catalytic efficiency). Substrates varied from 12.5 μM to 30 mM depending on the enzymes K_M. Oxidation was quantitated by measuring the decrease in absorbance at 340 nm ($\varepsilon = 6220\ M^{-1}\ cm^{-1}$).

Kinetic Assays for SeGlcDH. Reduction of 5-keto-D-gluconate (fructuronate) was monitored using a continuous spectrophotometric assay. The assay (200 µL), at 25 °C, contained 50 mM potassium HEPES, pH 7.9, 5 mM $MgCl_2$, 200 µM NADH, and 200 nM *Se*GlcDH. Substrate varied from 50 µM to 30 mM. Oxidation was quantitated by measuring the decrease in absorbance at 340 nm ($\varepsilon = 6220\ M^{-1}\ cm^{-1}$).

Kinetic Assay of CsUxuA. Dehydration of D-mannonate was monitiored using a continuous, coupled-enzyme spectrophotometric assay as previously described[59].

Quantitative RT-PCR of CsManD, RspABCD, and SeGlcD Genome Neighborhoods. *C. salexigens* DSM3043 was grown to an optical density (absorbance at 600 nm) of 0.4 - 0.5 in M9 minimal salts medium plus additional NaCl (1.7 M NaCl, 6.1 mM Na_2HPO_4, 3.9 mM KH_2PO_4, 9.3 mM NH_4Cl, 0.5 mM $MgSO_4$, 0.5 mM $CaCl_2$) and 10 mM D-mannonate, 10 mM L-gulonate, 10 mM D-glucuronate, or 10 mM D-glucose. Cells were pelleted at 15,000 rpm, and the supernatant was removed. mRNA was purified from the cells using an RNeasy Mini Kit (Qiagen). The RNA was further purified using RNase-free DNase (Qiagen) following the manufacturer's protocol. Purity was verified using 30 µL PCR reactions consisting of 50 ng mRNA, 1 mM $MgCl_2$, 1X Pfx Amp Buffer, 2X PCR enhancer 0.33 mM dNTP, 0.33 µM of primers (CsRpoD_RTPCR_FOR and CsRpoD_RTPCR_REV, **Table C.1**), and 1.25 units Pfx polymerase (Invitrogen Platinum Pfx DNA Polymerase kit). The PCR reactions were electrophoresed on an agarose gel to check for amplification. cDNA was prepared using Protoscript First Strand (New England Biolabs) and 1 µg mRNA; the manufacturer's protocol was followed. The qRT-PCR reaction was performed using the Light Cycler 480 SYBR Green I Master Kit (Roche) and a Light Cycler 480 II (Roche) according to the manufacturer's protocol.

The qRT-PCR primers are listed in **Table C.1**. C_p values were analyzed using the Light Cycler 480 application, and fold changes calculated in Microsoft Excel.

Salmonella enterica serovar Enteritidis str. P125109 was grown was grown to an optical density (absorbance at 600 nm) of 0.4 - 0.5 in M2 minimal salts medium (6.1 mM Na_2HPO_4, 3.9 mM KH_2PO_4, 9.3 mM NH_4Cl, 0.5 mM $MgSO_4$, 0.5 mM $CaCl_2$, and 10 μM $FeSO_4$) and either 10 mM L-gulonate/D-glucose or 1 – 100 mM L-idonate/D-gluconate/D-glucose. The following steps are identical to those taken with *C. salexigens.* The primers for *SeRspABCD and SeGlcD* qRT-PCR are listed in **Table C.1**.

Knockout Construction in C. salexigens DSM3043. The knockouts of genes in *C. salexigens* DSM3043 were constructed using overlap extension and a suicide vector (pK19mobsacB, ATCC# 87097)[56]. The DNA ~600 bp upstream and downstream of the gene of interest (GOI) was cloned using overlap extension to create a single linear product that perfectly excises the GOI as previously described (primers in **Table C.1**)[60]. This PCR product was digested with either EcoRI/HindIII or BamHI/HindIII (New England Biolabs) and ligated into similarly digested pK19mobsacB. The ligation products with confirmed sequences were transformed into *E. coli* WM6029 (obtained from Professor William Metcalf at the University of Illinois) for conjugation. *E. coli* WM6029 plus pK19mobsacB containing the GOI and *C. salexigens* DSM3043 were grown to an optical density (absorbance at 600 nm) of 0.4 - 0.5 in 3 mL PYE media[54] containing 500 mM NaCl and 50 μg/mL diaminopimelic acid (conjugation media). The cells were pelleted at 4500 rpm for 5 min, the supernatant was removed, and the cells were resuspended in 3 mL of conjugation media. Aliquots (250 μL) of both *E. coli* WM6029 plus pK19mobsacB containing the GOI and *C. salexigens* DSM3043 were mixed in a 1.5 mL eppendorf tube and pelleted at 4500 rpm for 5 min. The supernatant was removed, and

the cells were resuspended in 100 µL of conjugation media. The resuspended cells were inoculated onto a nitrocellulose filter, placed on a plate of conjugation medium plus 1.2% agar, and incubated overnight at 30 °C. The overnight conjugation was resuspended in 1 mL of PYE medium containing 500 mM NaCl. An aliquot (40 µL) of was plated onto PYE medium containing 500 mM NaCl, 4x Kn (200 µg/mL), and 1.2% agar for selection of single cross-over events. Single cross-over events were verified by colony PCR. Colonies with successful single-crossover events were subcultured onto a plate of PYE medium containing 500 mM NaCl and 1.2% agar. Single colonies were plated onto PYE medium containing 500 mM NaCl, 20% sucrose, and 1.2% agar. Colonies were probed for double-crossover events via colony PCR. Successful double cross-over events were verified by isolating the genomic DNA (Qiagen DNeasy® Blood and Tissue Kit) and sequencing the knockout region.

Knockout construction in Salmonella enterica serovar Enteritidis str. P125109. All knockouts were performed via the method described by Datsenko and Wanner for *E. coli[62]*. Primers are listed in **Table C.1**.

Growth Curves of WT and Knockout Strains in Chromohalobacter salexigens DSM3043. All strains were grown to an optical density (absorbance at 600 nm) of 0.4 - 0.5 in 3 mL PYE medium containing 1.7 M NaCl (50 µg/mL streptomycin for conjugations). The cells were pelleted at 5000 rpm for 5 min and resuspended in 3 mL of M9 minimal salts medium containing 1.7 M NaCl. The cells were inoculated into triplicate 300 µL cultures of M9 minimal salts medium containing 1.7 M NaCl and 10 mM of either D-mannonate or L-gulonate. Growth curves were recording using a Bioscreen C instrument (Growth Curves USA). Cells were grown at 37 °C for 120 hours with continuous shaking.

Fluorescence–based Thermal Shift Assay for CsGntR (ThermoFluor). The library for screening the specificity of CsGntR contained 188 ligands in duplicate wells plus 8 control wells (protein without compound) (**Figure C.2**). The assays were performed as previously described[60]. A hit is considered significant if the ΔT_M is greater $\geq$ 4°C or 3 times the standard deviation of the T_Ms for the control wells.

4.3 Results and Discussion

In vitro pathway for CsManD. The genome neighborhood for *Cs*ManD suggests a pathway for D-glucuronate metabolism that involves dehydration of D-mannonate (**Figure 4.1**)[29]. Specifically, the genome proximal fructuronate reductase that reduces fructuronate to D-mannonate and UxuA[53] that dehydrates D-mannonate to 2-keto-3-deoxy-D-mannonate (not a member of the ENS) are involved in D-glucuronate metabolism in many eubacteria[60]. However, *C. salexigens* DSM3043 uses an alternate pathway for D-glucuronate metabolism initiated by oxidation of D-glucuronate to D-glucaro-1,5-lactone[63],[64],[65] (**Figure 4.2**).

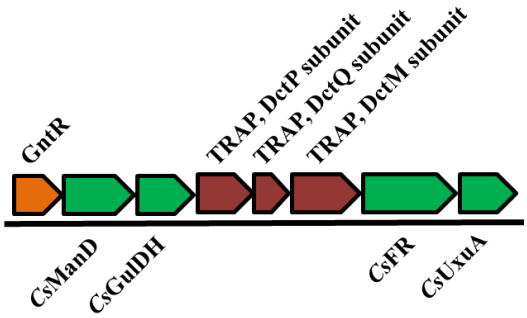


Figure 4.1. Genome neighborhood for *Cs*ManD in *Chromohalobacter salexigens* DSM3043. Carbohydrate metabolism genes are highlighted in green, TRAP transporters in red, and the GntR transcriptional regulator in orange.

The genome neighborhood encoding the predicted uronate dehydrogenase (Udh; Uniprot ID Q1QUN7) also encodes D-glucarolactone cycloisomerase (Gci; Uniprot ID Q1QUN5), a member of the amidohydrolase superfamily (presumably D-galactarolactone isomerase; Uniprot ID Q1QUN6), D-glucarate dehydratase (GlucD, Uniprot ID Q1QUM5), and 5-dehydro-4-deoxy-D-glucarate dehydratase/decarboxylase (Uniprot ID Q1QUM4) (**Figure C.3**). These enzymes are involved in an alternate pathway for D-glucuronate degradation[66],[67].

The physiological relevance of these enzymes/genes was probed using RT-PCR, comparing growth on D-glucuronate to that on D-glucose. The genes encoding all of these enzymes are upregulated; however, the genome neighborhood encoding *Cs*ManD is not upregulated (**Figure C.4).** Further studies are needed to determine whether *C. salexigens* DSM3043 utilizes the newly discovered Gci-based catabolic pathway[66],[67] to metabolize D-glucuronate or the previously described pathway in which D-glucarolactone is hydrolyzed to D-glucarate that is dehydrated by GlucD[63],[64],[65]; the Gci and "amidohydrolase" in *C. salexigens* DSM3043 are members of the same Pfam families as those characterized by Bouvier and coworkers[66], PF02746 and PF01979, respectively. Irrespective of the precise pathway, we conclude that *C. salexigens* DSM 3043 metabolizes D-glucuronate via conversion to D-glucarolactone/D-glucarate and not D-mannonate.

In addition to fructuronate reductase, a second alcohol dehydrogenase is encoded by the *Cs*ManD genome neighborhood. Its presence suggested that the genome neighborhood enables another acid sugar to be utilized as carbon source, with the oxidation product an intermediate in the ManD-utilizing D-glucuronate catabolic pathway, e.g., oxidation at carbon-5 of L-gulonate produces fructuronate (**Figure 4.2**). Therefore, we hypothesized that the genome neighborhood of *Cs*ManD is responsible for L-gulonate metabolism. A pathway for L-gulonate catabolism in

E. coli K-12 was suggested by Cooper in 1980, but the gene/enzyme responsible for oxidation of L-gulonate was not identified[68]. Furthermore, Cooper was only able to see growth of *E. coli* K-12 after spontaneous mutations had occurred[68].

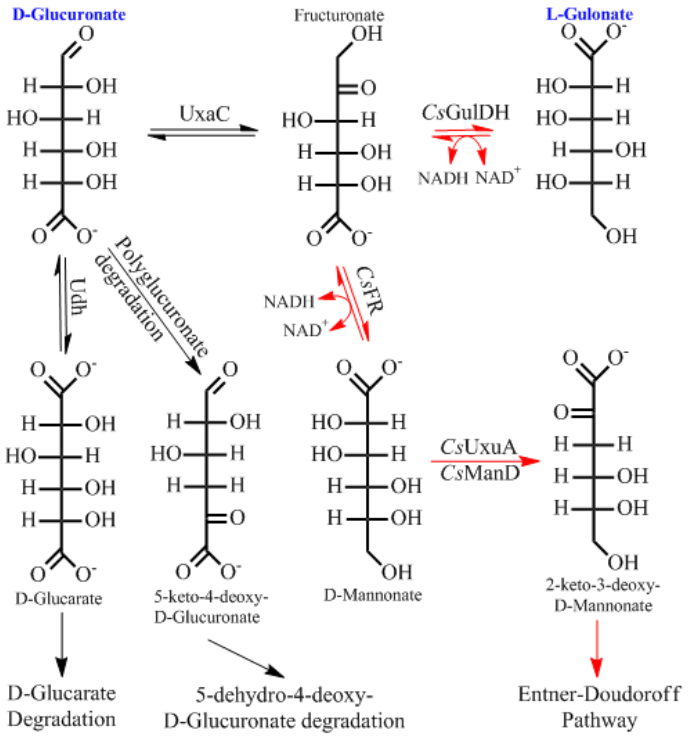


Figure 4.2. Catabolic pathways for D-glucuronate in eubacteria. The pathway hypothesized for degradation of L-gulonate in *C. salexigens* is highlighted in red: *Cs*GulDH (alcohol dehydrogenase), *Cs*FR (fructuronate reductase), *C*sManD (D-mannonate dehydratase), and UxuA (D-mannonate dehydratase). Starting compounds are labeled in blue (D-glucuronate and L-gulonate).

Functional Assignment in the CsManD Genome Neighborhood: CsUxuA. The functions of the proteins encoded by the genome neighborhood must be determined to assign a physiological role to *Cs*ManD. *Cs*UxuA (Uniprot ID Q1QT83) was screened for dehydration

activity using a library of 72 acid sugars (**Figure C.1**). The only hit was D-mannonate that was dehydrated to 2-keto-3-deoxy-D-mannonate with a catalytic efficiency of 1.5 x 10^3 M^{-1} s^{-1} (**Table 4.1**). Therefore, it was unclear whether UxuA, *Cs*ManD, or both were responsible for D-mannonate dehydration in this pathway.

Table 4.1. Kinetics for Enzymes Encoded by the *Cs*ManD Operon

Protein	**Substrate**	**k_{cat} (s^{-1})**	**k_{cat}/K_M ($M^{-1}s^{-1}$)**
*Cs*ManD[59]	D-mannonate	0.02 ± 0.0005	5.0 x 10^0
*Cs*UxuA	D-mannonate	1.6 ± 0.1	1.5 x 10^3
*He*GulDH (*Cs*GulDH)	L-gulonate	1.8 ± 0.13	6.0 x 10^2
RspD (*Cs*FR)	D-mannonate	30 ± 1.8	2.8 x 10^5

Functional Assignment within the CsManD Genome Neighborhood: CsGulDH. The alcohol dehydrogenase (*Cs*GulDH; Uniprot ID Q1QT88) in the *Cs*ManD genome neighborhood was not amenable to purification, despite several cloning strategies. A presumably orthologous alcohol dehydrogenase from the closely related *Halomonas elongata* DSM 2581 (*He*GulDH; Uniprot ID E1V4Y1) is 76% identical and 86% similar to *Cs*GulDH. Additionally, the genome neighborhood of *He*GulDH is identical to the *Cs*GulDH neighborhood. The purified *He*GulDH was screened for oxidation activity using the library of acid sugars (**Figure C.1**). Multiple hits were observed, with L-gulonate exhibiting the most efficient oxidation (k_{cat}/K_M = 6 x 10^2 $M^{-1}s^{-1}$; **Table 4.1**). All three L-gulonate dehydrogenases discussed in this manuscript (*Cs*GulDH, *He*GulDH, and RspB) are in Pfam family PF08240.

Functional Assignment within the CsManD Genome Neighborhood: CsFR. The fructuronate reductase (*Cs*FR; Uniprot ID Q1QT84) in the *Cs*ManD genome neighborhood also was not amenable to purification. *Cs*FR is 43% identical and 59% similar to a D-mannonate oxidoreductase in *Escherichia coli* CFT073 (RspD, Uniprot ID Q8FHD0); both proteins belong

to Pfam PF01232. RspD was purified and screened for oxidation activity with the acid sugar library described previously (**Figure C.1**). RspD exhibited specific oxidation of D-mannonate to fructuronate with a catalytic efficiency of 3 x 10^5 M^{-1} s^{-1} (**Tables 4.1 and 4.2**). Although the sequence similarity is not large, the similar genome contexts (*vide infra*) and enzymatic activities of CsGulDH/CsFR, and HeGulDH/RspD allow confident transfer of function between the two proteins, respectively. Considering their *in vitro* activities, these proteins provide support for the L-gulonate catabolic pathway proposed by Cooper[68]. Therefore, we pursued *in vivo* validation to corroborate the *in vitro* data and confirm the pathway.

Table 4.2. Kinetics for Enzymes Encoded by the RspABD Operon

Protein	**Substrate**	**k_{cat} (s^{-1})**	**k_{cat}/K_M ($M^{-1}s^{-1}$)**
RspA[59]	D-mannonate	0.02 ± 0.001	1.0 x 10^1
RspB	L-gulonate	3.9 ± 0.5	1.4 x 10^3
RspD	5-keto-D-mannonate	30 ± 1.8	2.8 x 10^5

Transcript Analysis of the CsManD Genome Neighborhood. Transcript analysis via qRT-PCR of mRNA revealed that all of genes in the neighborhood of *Cs*ManD (**Figure 4.1**) were upregulated ~20-300 fold during growth on both D-mannonate and L-gulonate relative to that on D-glucose (**Figure 4.3**). Mutual upregulation in combination with close proximity and identical orientation implies these genes are in an operon, but northern blots were not performed for confirmation. These findings support the claim that the CsManD gene cluster is responsible for L-gulonate metabolism *via* D-mannonate.

ThermoFluor for CsGntR. The GntR transcriptional regulator for the L-gulonate utilization operon (*Cs*GntR; Uniprot ID Q1QT90) was expressed, purified, and analyzed by ThermoFluor using a library of 188 ligands (sugars, amino acids, and various metabolites)

(**Figure C.2**). Both D-mannonate (ΔT_m of 7 °C) and L-gulonate (ΔT_m of 2 °C) were among the top 10 hits (**Table C.2**). The ligand specificity of the GntR provides further support of the involvement of the operon in growth on L-gulonate and D-mannonate.

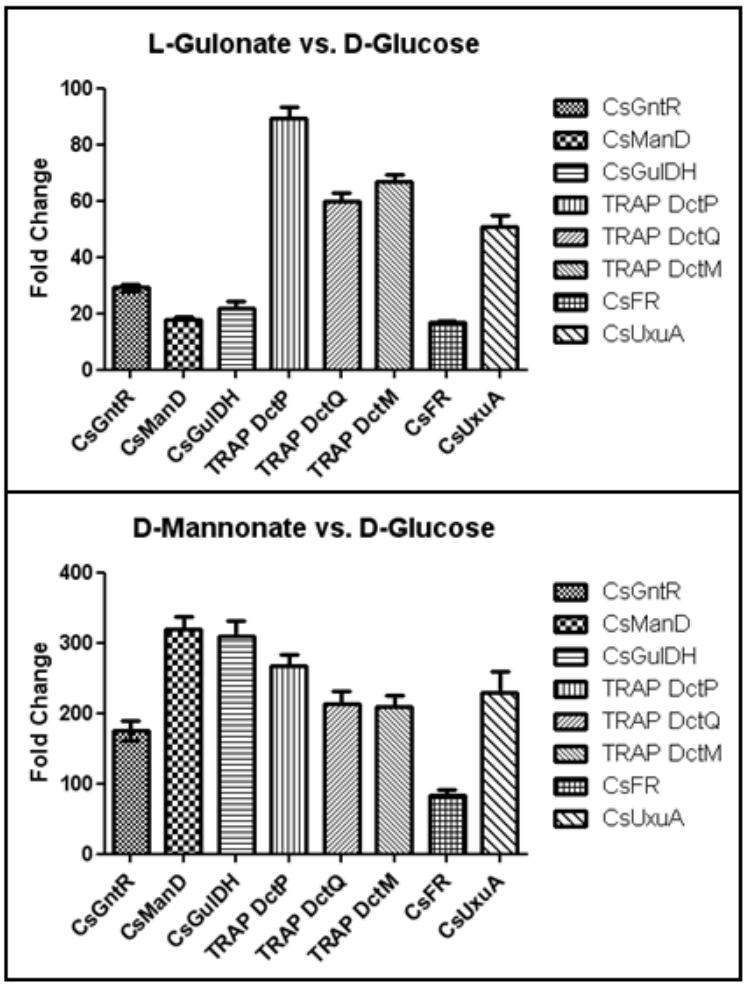


Figure 4.3. RT-PCR data for genes in the *Cs*ManD operon for cells grown on L-gulonate (top) or D-mannonate (bottom) versus growth on D-glucose. Upregulation is observred for all genes on both L-gulonate and D-mannonate.

Knockout Studies in the CsManD Genome Neighborhood. Gene deletion/knockouts (KOs) were constructed to further probe the proposed L-gulonate pathway. The following KO strains were constructed: Δ*Cs*GulDH, Δ*Cs*FR, Δ*Cs*, Δ*Cs*ManD, Δ(*Cs*ManD,*Cs*UxuA), and Δ(*Cs*FR, *Cs*UxuA); all of the genes were removed from the genome via homologous recombination. Their growth phenotypes using either D-mannonate or L-gulonate as sole carbon source were recorded (**Figure 4.4**).

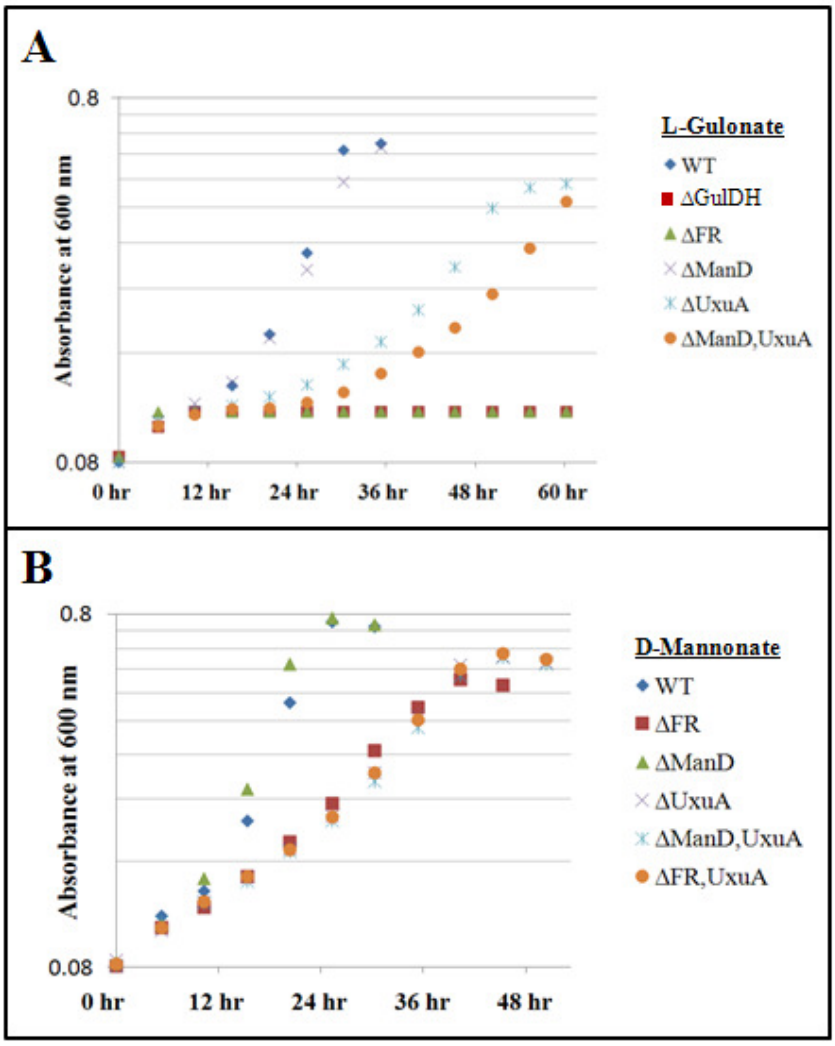


Figure 4.4. Growth curves for wild type *C. salexigens* DSM3043 and various knockout strains. Growth was recorded in M9 minimal medium with 1.7 M NaCl and either L-gulonate (Panel A) or D-mannonate (Panel B) as the sole carbon source.

For both D-mannonate and L-gulonate, the wild type strain and the Δ*Cs*ManD KO had identical growth curves, indicating *Cs*ManD is not required for the catabolism of either carbon source. In contrast, the Δ*Cs*UxuA KO resulted in slower growth with both carbon sources, suggesting *Cs*UxuA is the predominant ManD for both L-gulonate and D-mannonate catabolism. That the Δ(*Cs*ManD, *Cs*UxuA) double KO was able to grow on D-mannonate and L-gulonate implicates a slower, as-of-yet unidentified pathway for D-mannonate catabolism. Interestingly, the Δ(*Cs*ManD,*Cs*UxuA) double KO exhibited growth retardation (relative to either single KO) when grown on L-gulonate, suggesting that *Cs*ManD plays a role in the alternate D-mannonate metabolic pathway when L-gulonate is the carbon source.

Most importantly, the Δ*Cs*GulDH, Δ*Cs*FR, and Δ(*Cs*FR,*Cs*UxuA) knockout strains did not grow on L-gulonate. This provides convincing evidence that *Cs*GulDH, *Cs*FR, and *Cs*UxuA are responsible for the catabolism of L-gulonate. For D-mannonate, the growth rates for Δ*Cs*FR and are very similar to that for Δ*Cs*UxuA, which likely indicates a polar effect on *Cs*UxuA when *Cs*FR is knocked out (the genes encoding *Cs*FR and *Cs*UxuA are separated by 9 base pairs). Furthermore, the growth of the Δ(*Cs*FR, *Cs*UxuA) double KO on D-mannonate confirms that the *Cs*UxuA deletion strain does not grow by conversion of D-mannonate to fructuronate/L-gulonate and then funneling into an undiscovered catabolic pathway. The unidentified pathway utilizes genes not included in the *Cs*ManD genome neighborhood.

Other in vitro pathways implicated by low-efficiency ManDs: RspA. An *in vitro* pathway for the catabolism of L-gulonate was derived from the genome neighborhood of a low-efficiency ManD discovered in *Eschericia coli* CFT073. This pathway is encoded by four genes: *RspA*[59], a low-efficiency ManD (Uniprot ID Q8FHC7); *RspB*, annotated as an alcohol dehydrogenase

(Uniprot ID Q8FHC8); *RspC*, annotated as a hypothetical metabolite transporter (Uniprot ID Q8FHC9); and *RspD*, annotated as an oxidoreductase (Uniprot ID Q8FHD0) (**Figure 4.5**).

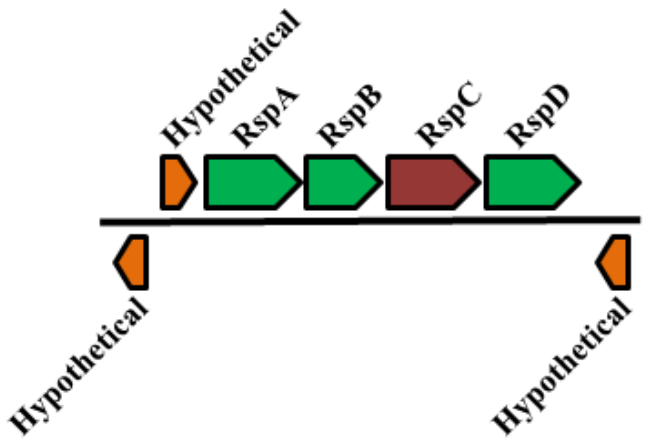


Figure 4.5. Genome neighborhood of the gene cluster in *E. coli* CFT073. RspA, low-efficiency D-mannonate dehydratase (Uniprot ID Q8FHC7); RspB, L-gulonate dehydrogenase (Uniprot ID Q8FHC8); and RspD, fructuronate reductase (Uniprot ID Q8FHD0) are highlighted in green. The hypothetical metabolite transporter (RspC, Uniprot ID Q8FHC9) is highlighted in red. Hypothetical genes of unknown function are highlighted in orange.

The genome neighborhood does not encode a UxuA, although one is encoded elsewhere in the genome. RspA (ManD), RspB (L-gulonate 5-dehydrogenase), and RspD (D-mannonate 5-dehydrogenase) were found to catalyze the *in vitro* conversion of L-gulonate to 2-keto-3-deoxy-D-gluconate (**Table 4.2**). Although RspA, RspB, and RspD convert L-gulonate to 2-keto-3-deoxy-D-mannonate/2-keto-3-deoxy-D-gluconate *in vitro*, no upregulation or KO phenotype for growth on L-gulonate was observed in either *E. coli* K-12 or *Salmonella enterica serovar Enteritidis* str. P125109 where identical gene clusters are found: *E. coli* K-12 does not

grow on L-gulonate until spontaneous mutations occur[68]; and, while the *RspABCD* gene cluster in *Salmonella enterica serovar Enteritidis* str. P125109 is negligibly upregulated (**Figure C.5),** knockouts of *RspA* and *RspB* grow to turbidity overnight on L-gulonate.

Other in vitro pathways implicated by low-efficiency ManDs: SeGlcD. Bausch and coworkers discovered a novel pathway for L-idonate metabolism in *E. coli* K-12 in which L-idonate is converted to D-gluconate via oxidation and subsequent reduction of the C5 hydroxyl group[69]. D-Gluconate is then phosphorylated and dehydrated to form 2-keto-3-deoxy-D-gluconate-6-phosphate, which is cleaved by an aldolase to pyruvate and glyceraldehyde-3-phosphate. Interestingly, *Salmonella enterica subsp. enterica serovar Enteritidis* str. P125109 encodes this L-idonate catabolic pathway as well as a similar pathway that is encoded by genes proximal to that encoding a low-efficiency GlcD (Uniprot ID B5R541; *Se*GlcD; k_{cat}/K_M = 80 M^{-1} s^{-1} [59])(**Figure 4.6 and 4.7**).

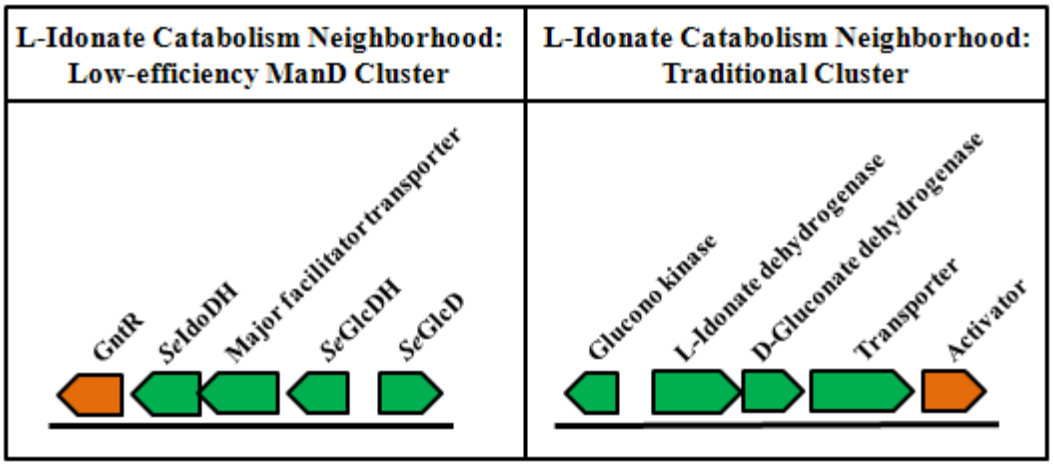


Figure 4.6. The genome neighborhoods of the traditional L-idonate catabolic pathway and the low-efficiency ManD-containing L-idonate catabolic pathway present in *Salmonella enterica subsp. enterica serovar Enteritidis* str. P125109.

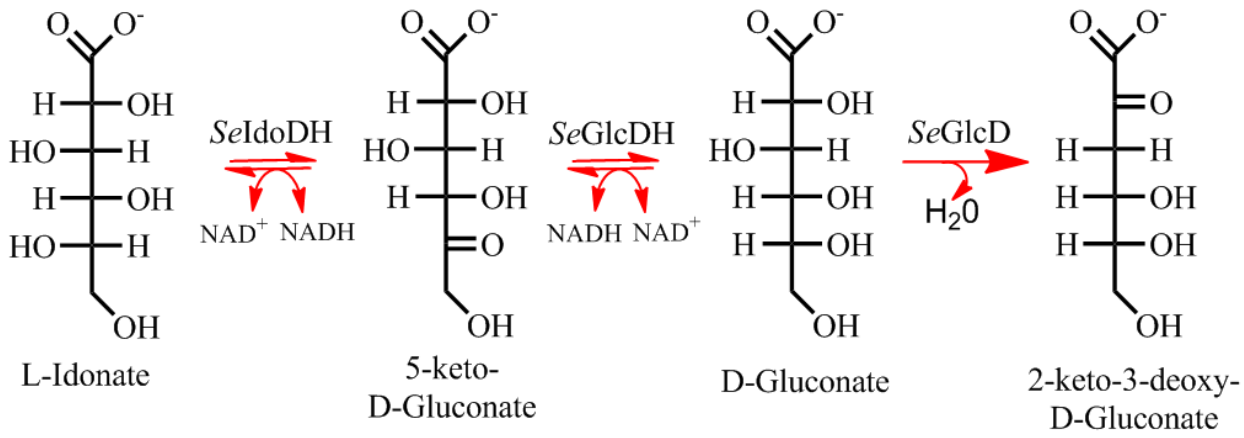


Figure 4.7. The *in vitro* catabolic pathway for consumption of L-idonate present in *Salmonella enterica subsp. enterica serovar Enteritidis* str. P125109.

In the latter pathway, L-idonate is converted to D-gluconate via two dehydrogenases as previously described (SeIdoDH and SeGlcDH; Uniprot IDs B5R538 and B5R540, respectively), but the subsequent steps of phosphorylation and dehydration apparently have shifted order: D-gluconate is dehydrated by *Se*GlcD to 2-keto-3-deoxy-D-gluconate which can be catabolized by either the non-phosphorylated or phosphorylated Entner-Doudoroff pathway (**Table 3**). We have been unable to determine a physiological role for this pathway; it is a cryptic pathway, i.e., growth on L-idonate or D-gluconate at concentrations as large as 100 mM did not result in upregulation of the genes (**Figures A4.6 and A4.7**).

Table 4.3. Kinetics for Enzymes Encoded by the *Se*GlucD Operon

Protein	Substrate	k_{cat} (s^{-1})	k_{cat}/K_M ($M^{-1}s^{-1}$)
*Se*IdoDH	L-idonate	1.2 ± 0.09	4.8×10^2
*Se*GlcDH	5-keto-D-gluconate	3.4 ± 0.5	3.8×10^3
*Se*GlcD[59]	D-gluconate	0.05 ± 0.003	8.0×10^1

As for the pathways for catabolism of L-gulonate using 1) *Cs*GulDH, *Cs*FR, *Cs*UxuA, and *Cs*ManD or 2) RspA, RspB, and RspD, the pathway for catabolism of L-idonate using *Se*IdoDH, *Se*GlcDH, and *Se*GlcD demonstrates the successive use of two dehydrogenases to epimerize the configuration of carbon-5. Although all three pathways are found via genome neighorhoods involving the low-efficiency ManDs, the physiological purposes, if any, of these dehydratases has yet to be discovered.

4.4 Conclusions

Our investigations into the role of the low-efficiency *Cs*ManD discovered a novel pathway for the catabolism of L-gulonate in *C. salexigens* DSM3043 via oxidation/reduction of carbon-5 (**Figure 4.2**). This pathway was proposed more than 30 years ago for *E. coli* K-12 when it was observed that spontaneous mutations allowed growth on L-gulonate, although the enzyme/gene for L-gulonate oxidation was not identified[68]. Based on sequence homology to *Cs*GulDH, *He*GulDH, and RspB, we propose that the likely L-gulonate 5-dehydrogenase in *E. coli* K-12 for which Cooper described the activity is Uniprot ID P38105 (44% identical to CsGulDH, 46% identical to *He*GulDH, and 98% identical to RspB).

We have determined that low efficiency D-gluconate and D-mannonate dehydratases need not be required for growth, although the their genome neighborhoods encode enzymes sufficient to constitute catabolic pathways for L-idonate and L-gulonate, respectively. Perhaps,

these “cryptic” pathways have been silenced as the result of lack of selective pressure or, alternatively, are evolving to meet new metabolic needs. Enzyme promiscuity has long been hypothesized to be the starting point of evolution[70]. A similar view, which is supported by laboratory evolution, is that the emergence of new functions is a gradual process where the evolution of a low-efficiency, promiscuous activity is the most favored option rather than starting “from scratch”[71]. This ideology supports the proposal that low-efficiency GlcDs are in evolutionary flux toward a new function if one assume that the high-efficiency ManD activity is the progenitor function in the ManD subgroup. These hypotheses imply that low-efficiency ManDs are such because they are under no selective pressure: they may be derived from high-efficiency ManDs and are becoming less efficient; or they are relics that have not yet evolved to high-efficiency because of a lack of selective pressure. Or, it is possible that we have not yet discovered the true physiological role of the low-efficiency ManDs/GlcDs. Finally, assessing the physiological roles of these pathways also may be complicated because in Nature *E. coli* and *S. enteriditis* are members of bacterial communities, i.e., these pathways may be activated by unknown metabolites produced by other members of the community.

CHAPTER 5: ENZYMATIC AND STRUCTURAL CHARACTERIZATION OF rTS γ PROVIDES INSIGHTS INTO THE FUNCTION OF rTS β

ABSTRACT: In humans, the gene encoding a reverse thymidylate synthase (*rTS*) is transcribed in the reverse direction of the gene encoding the thymidylate synthase (*TS*) that involved in DNA biosynthesis. Three isoforms are found: α, β, and γ, with the transcript of the α-isoform overlapping with that of *TS*. rTSβ has been of interest since the discovery of its overexpression in methotrexate and 5-fluorouracil resistant cell lines. Despite more than 20 years of study, none of the rTS isoforms have not been biochemically or structurally characterized. In this study, we identified rTSγ as an L-fuconate dehydratase and determined its high-resolution crystal structure. Our data provide an explanation for the observed difference in enzymatic activities between rTSβ and rTSγ, enabling more informed proposals for the possible function of rTSβ in chemotherapeutic resistance.

Reprinted (adapted) with permission from Wichelecki, D. J., Froese, D. S., Kopec, J., Muniz, J. R. C., Yue, W. W., and Gerlt, J. A. Biochemistry. 2014 Apr 29;53(16):2732-8.

5.1. Introduction

In establishing a quantitative PCR assay for human thymidylate synthase (*TS*), a partially overlapping gene at the chromosome location 18p11.32, designated *reverse thymidylate synthase* (*rTS*), was found to be transcribed in the reverse direction with marginal overlap between the 3'-untranslated region of *rTS* and the last intron of *TS* [72]. Since then, two other isoforms have been identified from alternatively spliced mRNA, yielding a total of three isoforms: rTSα

(UniprotKB G2MQH2), rTSβ (G2MQH3), and rTSγ (Q7L5Y1) [73, 74]. All three proteins share the same 341 C-terminal amino acid sequence but differ in their N-terminal sequences and lengths. rTSα appends an additional 20 residues to the N-terminal of the shared sequence making it a 361 residue polypeptide. rTSβ has a different N-terminal sequence of 75 residues, i.e., 416 residue polypeptide, while rTSγ appends an additional 27 amino acids to the N-terminus of rTSβ to form a 443 residue polypeptide (**Figure D.1**). The *rTS* genes have been of clinical interest in the cancer research community for over 20 years, yet their functions remain largely unknown. The main interest lies in the correlation of cell lines resistant to the common chemotherapeutics methotrexate and 5-fluorouracil, with rTSβ overexpression [73-78]. Methotrexate is a competitive inhibitor of dihydrofolate reductase (DHFR), and 5-fluorouracil is an irreversible inhibitor of TS. Both DHFR and TS play central roles in the synthesis of dTMP, a precursor for DNA synthesis. In clinical studies, rTSβ was found to be expressed in breast cancer tissue but not the surrounding tissues [79]. Furthermore, a statistically significant correlation was found between the level of rTSβ expression and a decrease in the five year survival rate of colon cancer patients [76]. The 27 residue longer N-terminus of rTSγ, compared to rTSβ, is proposed to constitute a mitochondrial signaling sequence [80], suggesting that rTSβ and rTSγ serve similar enzymatic functions. The function of rTSα, the shortest of the three proteins, remains to be postulated.

Based on sequence homology and inspection of key catalytic residues, the three rTS isozymes are members of the mandelate racemase (MR) subgroup of the enolase superfamily (ENS) that is known to catalyze the racemization of mandelate as well as dehydration of various acid sugar [81]. All ENS members catalyze the abstraction of a proton alpha to a carboxylate group of the substrate to form a Mg^{2+} stabilized enediolate anion, although the overall reaction is

not conserved [82]. Furthermore, members of the ENS share common structural motifs that form the catalytic machinery, including a capping domain for substrate specificity and a $(\beta/\alpha)_7\beta$-barrel domain for acid/base chemistry [24, 83].

In order to provide functional insights into the medically relevant rTSβ and facilitate future studies, we isolated recombinant human rTSβ and rTSγ isoforms, identified rTSγ as an L-fuconate dehydratase using a medium throughput enzymatic screen, and determined the crystal structure of rTSγ at 1.74 Å resolution.

5.2. Materials and methods

*Cloning, expression and purification of human rTSγ (*Hs*rTSγ).* A DNA fragment containing residues 1-440 of *Hs*rTSγ (IMAGE clone: 3454185) was subcloned into the pNIC28-Bsa4 vector (GenBank accession EF198106) incorporating an N-terminal TEV-cleavable His_6-tag. The plasmid was transformed into *E. coli* BL21(DE3)-R3-pRARE2 cells, grown overnight at 18 °C in 12 L of TB medium after induction by 0.1 mM IPTG. Cells were harvested, lysed in buffer A (50 mM HEPES, pH 7.4, containing 500 mM NaCl, 5% glycerol, 10 mM imidazole, 0.5 mM TCEP, and an EDTA-free protease inhibitor), and centrifuged to remove insoluble debris.

The initial stage of purification involved passing the clarified cell extract through 2.5 ml Ni-NTA resin pre-equilibrated with buffer A.Bound protein was eluted with buffer B (buffer A + 250 mM imidazole). The fractions containing *Hs*rTSγ were applied onto a HiLoad 16/60 Superdex 200 column pre-equilibrated with GF buffer (10 mM HEPES, pH 7.4, containing 500 mM NaCl, 5% glycerol, and 0.5 mM TCEP). Fractions containing *Hs*rTSγ were treated with TEV protease overnight at 4 °C, and passed over Ni-Sepharose resin pre-equilibrated with GF

buffer. The tagless protein was then diluted to 50 mM NaCl, and applied to a 1 mL Resource-S column pre-equilibrated with IEX buffer (50 mM Tris-HCl pH 8.5, containing 50 mM NaCl). Protein was eluted with a linear gradient of 0-500 mM NaCl and concentrated to 35 mg/mL for storage at -80 °C. Further purification details can be found on: http://www.thesgc.org/structures/4a35.

Crystallization, data collection, and structure determination. Crystals of *Hs*rTSγ were grown by vapor diffusion at 20 °C. A sitting drop containing 100 nL protein (35 mg/mL) and 50 nL well solution was equilibrated against well solution containing 25% (v/v) PEG3350 and 0.1 M Bis-Tris, pH 5.5. Crystals were mounted in the presence of 25% (v/v) ethylene glycol and flash-cooled in liquid nitrogen. Diffraction data for the native *Hs*rTSγ crystal were collected at the Diamond Light Source beamline I03 and processed with the XDS and SCALA. *Hs*rTSγ crystallized in the hexagonal space group $P6_522$ with one polypeptide in the asymmetric unit. The structure was solved by single-wavelength anomalous dispersion (SAD) phasing.

Crystals were derivatized by incubation with reservoir solution supplemented with 1 mM thiomersal for 60 min. Diffraction data were collected at the Diamond Light Source beamline I03, and processed and scaled with XDS [84] and Scala [85], respectively. SHELXD [86] identified two heavy atom sites, and after phase refinement in SHARP [87] and subsequent density modification with SOLOMON, [88] an electron density map of excellent quality was obtained and substantial parts of the model were automatically built with ARP/wARP [89]. The resulting model was refined against the native data, where manual model rebuilding was carried out with Coot [90], and structure refinement with BUSTER (Global Phasing Ltd, Cambridge, UK).

*Cloning, expression, and purification of human rTSβ (*Hs*rTSβ).* The gene encoding *Hs*rTSβ was PCR amplified from the *Hs*rTSγ:pNIC28-Bsa4 plasmid using platinum *Pfx* polymerase (Invitrogen). The PCR reaction (30 µL) contained 50 ng template, 1 mM $MgCl_2$, 1X Pfx Amp Buffer, 0.33 mM dNTP, 0.33 µM of each primer (forward primer 5'-TGGGTACCGAGAACCTGTACTTCCAACATATGCACACGGACCC-3' and reverse primer 5'-CAGTGGTGGTGGTCGTGGTGCTCGAGT3'), and 1.25 units Pfx polymerase (Invitrogen Platinum Pfx DNA Polymerase kit). Amplifications were performed according to the manufacturer's guidelines. The amplification product was digested by NdeI and EcoRI (New England Biolabs) and ligated into NdeI/EcoRI digested pET17b (Novagen). The rTSβ gene in pET17b was expressed in *Escherichia coli* BL21 (DE3). Small-scale cultures were grown at 37 ˚C for 18 hrs in 5 mL of LB containing 100 µg/mL ampicillin and used to inoculate 1 L LB containing 100 µg/mL ampicillin. The 1 L cultures were grown for an additional 18 hrs at 37 ˚C without induction. The cells were harvested by centrifugation at 5000 rpm for 10 minutes and resuspended in 70 mL of binding buffer (6 mM imidazole, 20 mM Tris-HCl, 5 mM $MgCl_2$, and 500 mM NaCl, pH 7.9). The resuspended cells were lysed by sonication and centrifuged at 17,000 rpm for 30 minutes. The supernatant was loaded onto a 300 ml DEAE-Sepharose column (Amersham Biosciences) and eluted with a NaCl gradient (0 - 1 M over 1.6 L) in 10 mM Tris-HCl, pH 7.9, containing 5 mM $MgCl_2$. Fractions were analyzed using SDS-PAGE. Fractions that contained rTSβ were combined and dialyzed for 2 hrs at 4˚C against 4 L of 10 mM Tris-HCl, pH 7.9, containing 5 mM $MgCl_2$. The dialyzed protein was then loaded onto a 30 ml Q-Sepharose column (Amersham Biosciences) and eluted with a NaCl gradient (0 - 1 M over 500 mL) in 10 mM Tris-HCl, pH 7.9, containing 5 mM $MgCl_2$. Fractions were analyzed using SDS-PAGE. Fractions that contained *Hs*rTSβ at high purity were combined and dialyzed for 2

hrs at 4˚C against 4 L of 10 mM Tris-HCl, pH 7.9, containing 5 mM $MgCl_2$. Ammonium sulfate was added to a final concentration of 1 M, and the protein was loaded onto a 30 mL phenylsepharose column (Amersham Biosciences). The protein was eluted with a gradient of ammonium sulfate (1 - 0 M over 500 mL) in 10 mM Tris-HCl, pH 7.9, containing 5 mM $MgCl_2$. Fractions were analyzed using SDS-PAGE. Fractions with pure *Hs*rTSβ were combined and dialyzed for 2 hr at 4˚C against 4 L of 10 mM Tris-HCl pH 7.9, containing 5 mM $MgCl_2$, 100 mM NaCl, and 10% glycerol. Finally, the protein was concentrated to 7.6 mg/mL, flash frozen using liquid N_2, and stored at -80 ˚C prior to use.

Screen for dehydration. Reactions to test for dehydration activity for *Hs*rTSβ and *Hs*rTSγ were performed in acrylic, UV transparent 96-well plates (Corning Incorporated) using a library of 72 acid sugars (**Figure D.2**). Reactions (60 μL total volume) contained 50 mM HEPES, pH 7.9, 10 mM $MgCl_2$, 1 μM enzyme, and 1 mM of acid sugar substrate (blanks with no enzyme). The plates were incubated at 30 ˚C for 16 hr. After incubation, 240 μL of a semicarbazide solution (1% semicarbazide w/v, 1% sodium acetate w/v) was added to each well and the plate was incubated for 1 hr at room temperature. The absorbance at 250 nm was measured (semicarbazone ε = 10,200 M^{-1} cm^{-1}) using an Infinite M200 PRO microplate reader (Tecan Group Ltd.).

Kinetic assays of Hs*rTSγ*. Kinetic constants for the dehydration of L-galactonate, D-arabinonate, L-arabarate, and D-ribonate were measured using a discontinuous assay containing the semicarbazide assay [35, 91]. Kinetic constants for the dehydration of L-fuconate were determined using a coupled enzyme assay as described by Yew and coworkers [92].

Differential scanning fluorimetry. *Hs*rTSβ and *Hs*rTSγ were assayed for shifts in melting temperature as previously described [93]. Each protein (5 μg) was assayed as purified and, also, in the presence of 5 mM $MgCl_2$ or 5 mM $MgCl_2$ and 5 mM D-erythronohydroxamate.

Methotrexate and 5-fluorouracil Assays. Methotrexate (Sigma-Aldrich) or 5-fluorouracil (Sigma-Aldrich) (10 mM) was incubated (800 μL) with 50 mM deuterated Tris-DCl, pD 7.9, 5 mM $MgCl_2$ and 1 μM purified protein (rTSβ or rTSγ) in D_2O for 48 hr at 37 °C. Immediately following incubation, samples were analyzed for deuterium incorporation into methotrexate or 5-fluorouracil via ^{1}H-NMR. In the case of a reaction that occurs without deuterium incorporation (with methotrexate), the optical activity at 589 nm was measured for an 800 μL reaction containing 10 mM methotrexate, 50 mM HEPES, pH 7.9, 10 mM $MgCl_2$, and 1 μM rTSβ in H_2O. Measurements were made at room temperature using a Jasco P-1010 polarimeter (Jasco Inc.) conFigured with a halogen lamp and 589 nm sodium d-line filter.

5.3. Results and discussion

Characterization of rTSγ. A sequence similarity network of all proposed acid sugar dehydratases within the ENS is shown in **Figure D.3**. Clusters in the network were assigned specific acid sugar dehydratase functions based on homology to known acid sugar dehydratases, including conservation of metal binding and catalytic residues. *Hs*rTSγ, together with the canonical FucD from *Xanthomonas campestris* (*Xc*FucD), is found within the fuconate dehydratase (FucD) clustered at an e-value threshold of 10^{-80} (~40% identity) (**Figure D.3**). *Hs*rTSγ and *Xc*FucD diverge as the e-value threshold is decreased to 10^{-180} (~70% identity) (**Figure 5.1**) [92]. The sequences of *Hs*rTSγ and *Xc*FucD are 52% identical and 71% similar. In the mannonate dehydratase subgroup of the ENS, proteins that are 67% identical and 79% can

catalyze different enzymatic reactions [52]. Therefore, experimental characterization is necessary to assign an enzymatic function to *Hs*rTSγ. *Hs*rTSγ was screened for dehydration activity with a library of 72 acid sugars. Positive screening hits were verified via ^{1}H-NMR, and steady-state kinetic constants were obtained the confirmed substrates. The following sugars were positive hits in the screen: L-fuconate, L-galactonate, D-arabinonate, L-arabarate, and D-ribonate (**Figure 5.2**). All sugars with the exception of L-arabarate were similarly identified as substrates for *Xc*FucD substrates [92]. L-Fuconate shows the greatest catalytic efficiency (k_{cat}/K_M = 2.5 x 10^3 M^{-1} s^{-1}), which is an order of magnitude greater than that for the second best substrate, L-galactonate (k_{cat}/K_M = 1.0 x 10^2 M^{-1} s^{-1}) (Table 1). Together with conservation of active site catalytic and metal binding residues, the kinetic

Table 5.1. Steady-State Kinetic Parameters for rTSγ

Substrate	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M (M^{-1} s^{-1})
L-fuconate	0.5 ± 0.006	0.2 ± 0.01	2.5 x 10^3
L-galactonate	0.3 ± 0.01	3.0 ± 0.2	1.0 x 10^2
L-arabinarate	0.3 ± 0.004	4.0 ± 0.5	7.5 x 10^1
D-arabinonate	0.04 ± 0.002	2.0 ± 0.1	2.0 x 10^1
D-ribonate	0.002 ± 0.001	0.4 ± 0.2	5.0

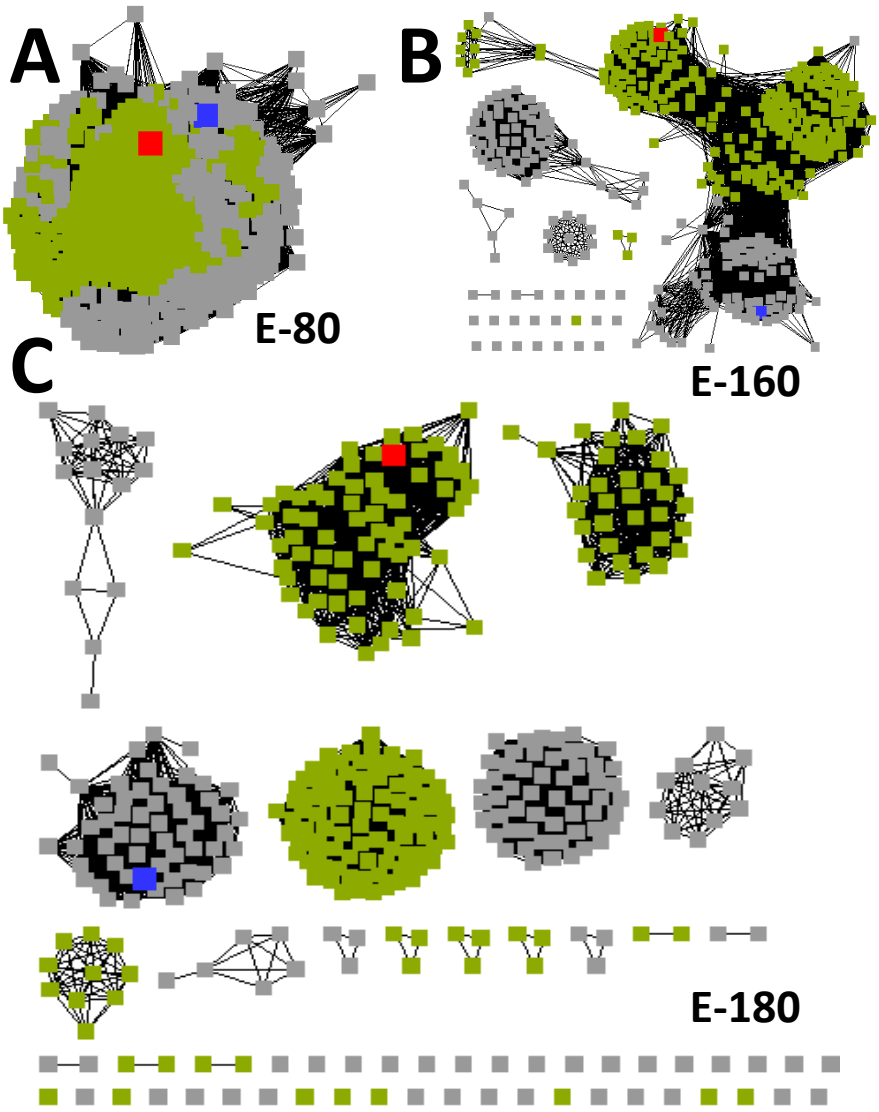


Figure 5.1. Panel A, sequence similarity network for the L-fuconate dehydratase subgroup at an e-value threshold of 10^{-80} (~40% identity). Panel B, network at an e-value threshold of 10^{-160} (~65% identity). Panel C, network at an e-value threshold of 10^{-180} (~70% identity). The nodes for *Xc*FucD (PDB 2HXT) and rTSγ are colored red and blue, respectively.

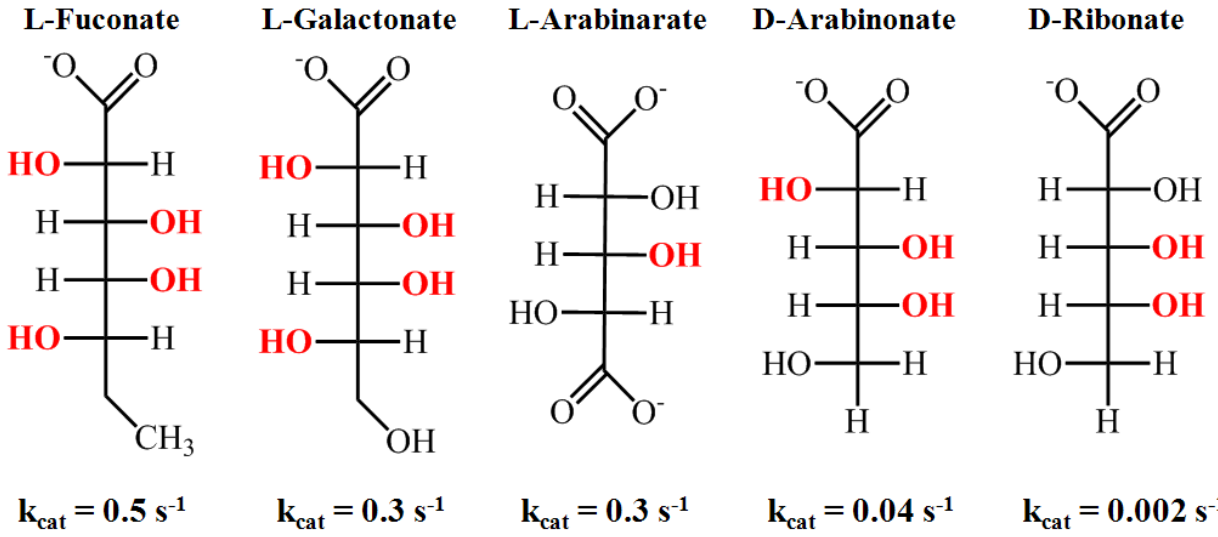


Figure 5.2. Structures of the top dehydration screening hits for rTSγ. The first order rate constants for dehydration are shown below the corresponding acid sugar. Carbons with conserved stereochemistry to L-fuconate have their hydroxyl groups highlighted in red.

data establish rTSγ as an L-fuconate dehydratase. In humans, as well as bacteria, L-fuconate dehydratase is important for the metabolism of L-fucose, which is found on the cellular surface of mammalian, insect, and plant cells [92, 94, 95].

The k_{cat}/K_M value for L-fuconate in *Hs*rTSγ is 10-fold lower than for *Xc*FucD (4.5×10^4 M^{-1} s^{-1}) [92]. This difference in catalytic efficiency may not be surprising given the moderate percent identity shared between the two enzymes. Additionally, the ten-fold reduction could also be explained by post-translational modifications in native *Hs*rTSγ that may not be present in the recombinant protein expressed in bacteria.

Crystal structure of Hs*rTSγ.* The crystal structure of *Hs*rTSγ with Mg^{2+} was solved by Hg-SAD phasing to a resolution of 1.74 Å (**Table 5.2**, PDB 4A35). The structure confirms its membership within the MR subgroup of the ENS, with *Xc*FucD being the closest structural

relative, as identified by the DALI server [92] (root-mean-square deviation of 1.03 Å for 430 C^{α} atoms, Z-score 62.0, PDB 2HXT) (**Figure 5.3a**). As a member of the MR subgroup, *Hs*rTSγ contains the following canonical signature sequences in the core $(\beta/\alpha)_7\beta$-barrel domain that are essential for catalysis (**Figure D.1**): 1) a KxK motif (Lys 220-Val 221-Lys 222; *Hs*rTSγ residue numbering hereafter) at the end of the 2nd β-strand of the barrel domain for base-catalyzed proton abstraction, 2) acidic residues Asp 250, Glu 276, and Glu 305 at the ends of the 3rd , 4th , and 5th β-strands of the barrel domain, respectively, which provide the ligands for the essential Mg^{2+} (**Figure 5.3b**), and 3) a His-Asp dyad (His 355-Asp 328) at the ends of the 7th and 6th β-strands of the barrel domain, respectively, which is the general acid catalyst for dehydration of the enediolate intermediate [96]. The RMSD of these conserved residues between the *Hs*rTSγ and *Xc*FucD structures is 0.31 Å, supporting their catalytic roles in *Hs*rTSγ. Based on the characterization of *Xc*FucD, *Hs*rTSγ is assumed to utilize Lys 222 as the general basic catalyst for proton abstraction and His 355 as the general acid catalyst for dehydration (**Figure 5.3b**).

Table 5.2. Data collection and refinement statistics for rTSγ

Data collection	
Space group	$P6_522$
No. of molecules in asym. unit	1
Cell dimensions	
a (Å)	84.77
b (Å)	84.77
c (Å)	316.30
βº	γ=120
Resolution (Å)	1.74
No. of unique reflections	69165
R_{merge}	0.04 (0.79)
Completeness (%)	99.0 (93.9)
Refinement	
Resolution (Å)	20.00-1.74 (1.78-1.74)
R_{cryst}	0.152 (0.240)
R_{free}	0.176 (0.288)
No. atoms	
Protein	3468
Waters	594
Bound ligands	MG, EDO
Ligand atoms	30
R.m.s deviations	
Bond lengths (Å)	0.015
Bond angles (°)	1.11
PDB accession code	4A35

Bracketed values represent highest resolution shell.

Liang and coworkers previously postulated that *Hs*rTSγ was simply an isoform of *Hs*rTSβ appended with a mitochondrial signaling sequence at its N-terminus [80]. Structural analysis of *Hs*rTSγ reveals that the extra 27 residues in the γ isoform comprise the first β-strand within the capping domain, a short alpha-helix, and a large portion of the "20s" loop. This β-strand is part of a three-stranded β-sheet in the capping domain and packs closely against the C-terminus, while the "20s" loop, a conserved structural feature in MR subgroup members, is known to confer substrate specificity while occluding solvent from the active site [97-100] (**Figure 5.4**). Considering the structural integrity of this region and its close proximity to the active site, we propose that the extra 27 residues are not involved in mitochondrial targeting, but instead are necessary for the enzyme architecture and also possibly confer L-fuconate dehydratase activity to *Hs*rTSγ. This is supported by the lack of definitive mitochondrial targeting signals identified from various bioinformatics servers (data not shown).

Pursuit of rTSβ Function. The importance of the N-terminal 27 residues for *Hs*rTSγ reactivity and integrity implies that *Hs*rTSβ, lacking this region, may have a deformed active site (e.g. highly solvent exposed, **Figure D.4**) and be devoid of L-fuconate dehydratase activity. Accordingly, we pursued enzymatic characterization of recombinantly produced *Hs*rTSβ. Our initial attempt to subclone the gene encoding *Hs*rTSβ, as described for *Hs*rTSγ (i.e. N-terminal His-tagged fusion protein), resulted in insoluble protein. This suggests the importance of the N-terminal 27 residues for protein stability, in addition to its role in enzyme activity, and is also consistent with our observations that truncation constructs of *Hs*rTSγ lacking the N-terminal 5, 8, or 12 residues rendered the γ isoform to be highly insoluble in *E. coli* (**Figure D.5**). We managed to subclone *Hs*rTSβ as a tag-less construct which resulted in a low level of expression in *E. coli*. The purified protein was screened for dehydration activity on the library of 72 acid sugars as

described above. Based on the expectation that the complete 20s loop is necessary for catalysis, no dehydration activity was detected for *Hs*rTSβ. We also characterized *Hs*rTSβ and *Hs*rTSγ by differential scanning fluorimetry (**Table D.1**) and showed that *Hs*rTSβ is more thermolabile than *Hs*rTSγ (ΔT_m = 5 ^oC) and, unlike *Hs*rTSγ, is not thermally stabilized by the addition of D-erythronohydroxamate, an enediolate intermediate analog, thereby suggesting that *Hs*rTSβ does not bind the ligand. Taken together, our data substantiate the hypothesis that the N-terminal 27 amino acid residues in rTSγ are necessary for catalytic activity and confer stability to the protein.

Several studies have observed uncharacteristically high levels of *Hs*rTSβ expression in tumor cell lines that also exhibit resistance to either methotrexate or 5-fluorouracil [73-78]. This correlation suggests the possibility that *Hs*rTSβ, which does not exhibit the L-fuconate dehydratase activity of *Hs*rTSγ, could instead harbor an alternative metabolic activity on these small molecules in a manner that inactivates them. The observation that *Hs*rTSβ expression was found to be induced by TS inhibitors indicates the protein may indeed be acting on these drug molecules [73][75, 101]. To explore this possibility, methotrexate and 5-fluorouracil were analyzed for their ability to act as substrates for *Hs*rTSγ. Methotrexate has a proton alpha to the carboxylate group of its glutamate side chain, which could be racemized by an ENS member. Because methotrexate has more than twice the molecular weight of L-fuconate, a truncation of the "20s" loop, such as that seen in the β isoform, could expose a larger binding surface to accommodate methotrexate in the active site. Reactions were performed in D_2O to determine if exchange of the alpha proton (racemization) were occurring. Also, 5-fluorouracil was tested for defluorination. No incorporation of deuterium was observed via ^1H NMR in the presence of either *Hs*rTSγ or *Hs*rTSβ. Furthermore, to test a racemization mechanism in which deuterium

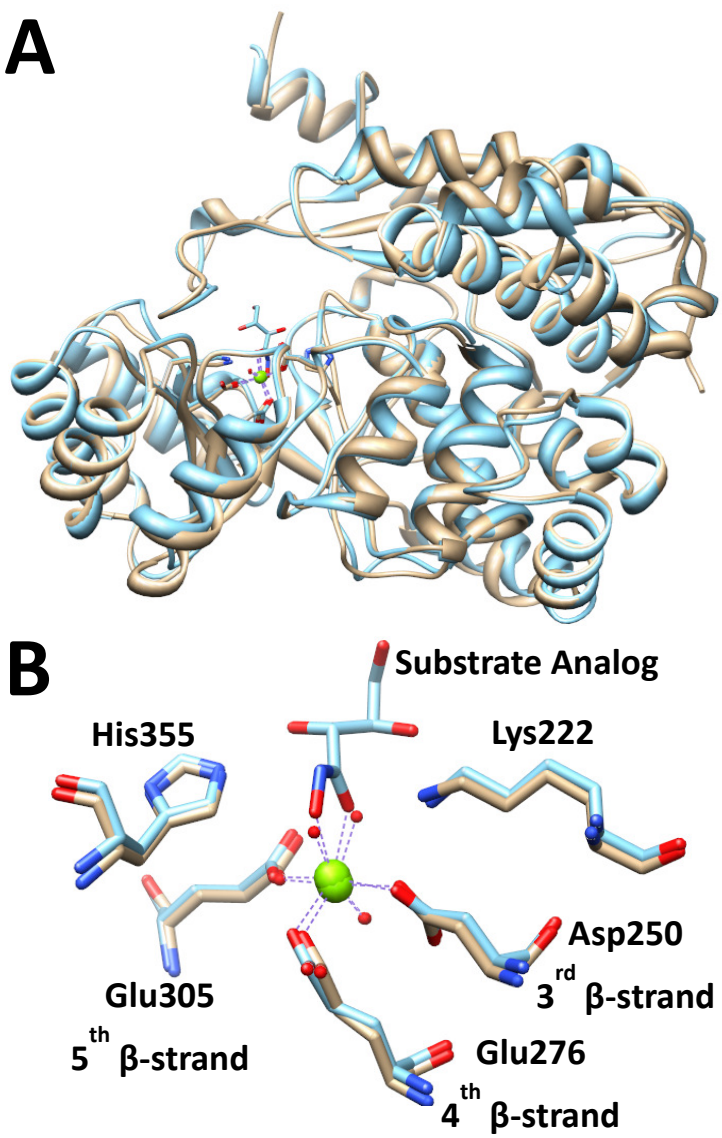


Figure 5.3. Panel A, an overlay of the structure of rTSγ (PDB 4A35, tan) and *Xc*FucD bound to the substrate analog L-erythronohydroxamate (PDB 2HXT, blue). The Cα RMSD is 1.03 Å. Panel B shows an overlay of the active site residues of the structures. The Cα RMSD for these metal binding and catalytic residues is 0.31 Å. Residue number is from the structure of rTSγ.

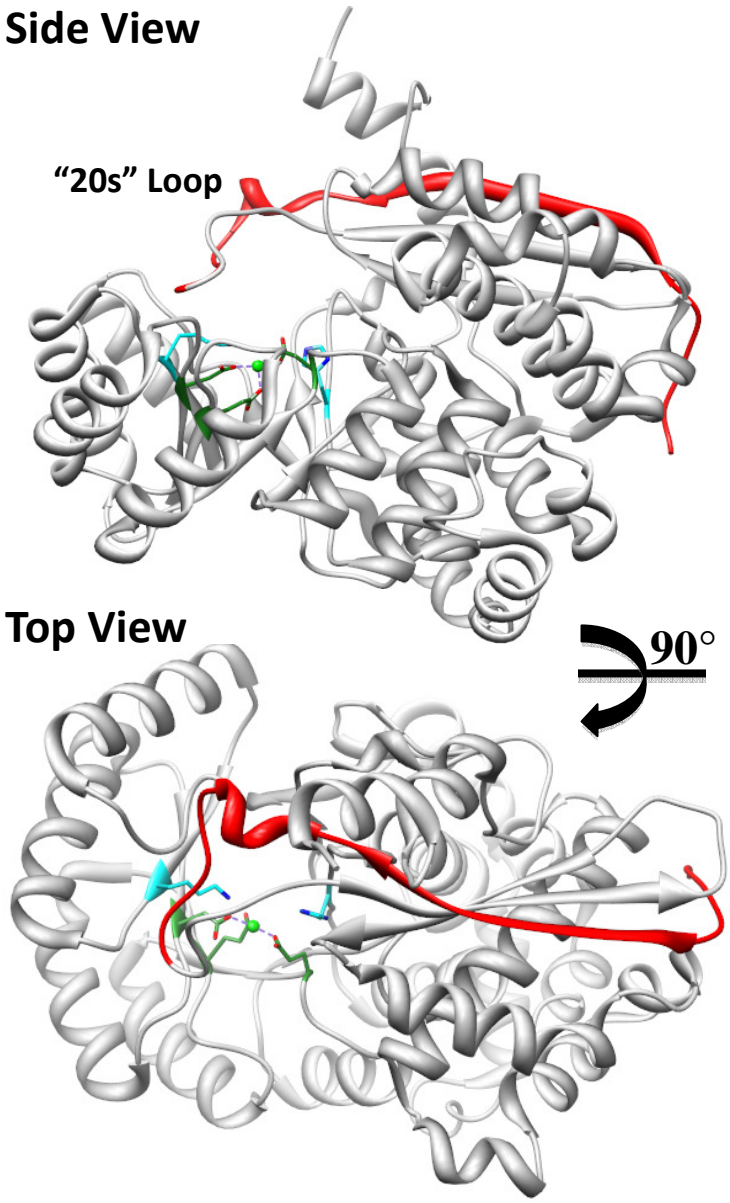


Figure 5.4. The structure of rTSγ (PDB 4A35) from side and top views. Each structure has the first 27 amino acids, which would be missing in rTSβ, highlighted in red. The metal binding residues are highlighted in green and catalytic residues in cyan.

exchange does not occur, methotrexate was incubated with *Hs*rTSβ or *Hs*rTSγ in a polarimeter, but no change in optical rotation was observed. These negative results, in conjunction with the lack of identification of an acid sugar substrate, allow us to conclude that rTSβ is not involved in chemical modification of methotrexate or 5-fluorouracil.

Dolnick proposed that *Hs*rTSβ could function by transferring a carboxylate group from S-adenosylmethionine to a lipophilic acceptor [101-103]. We find this unlikely, considering that MR subgroup members catalyze the abstraction of a proton alpha to a carboxylate group, which is not a partial reaction known to result in decarboxylation. Based on the *in vitro* function of *Hs*rTSγ and sequence homology to members of the MR subgroup, *Hs*rTSβ likely lacks an enzymatic activity. This, however, does not prohibit *Hs*rTSβ from participating in protein-protein interactions that may produce the observed resistance phenotype. In fact, TS can physically associate with proteins of the same molecular weight as rTSα and rTSβ, and conversely, rTSα and rTSβ could coimmunoprecipitate with TS [75]. Furthermore, preliminary data suggesting that *rTS* proteins interact with dihydrofolate reductase (DHFR) could explain the correlation between rTSβ overexpression and methotrexate resistance [103]. The characterization of *Hs*rTSγ as an L-fuconate dehydratase, coupled with the inability of *Hs*rTSβ to catalyze this reaction, points to a mechanism of resistance that does not involve catalysis, but rather interactions with one of the above proposed partners. Future *in vitro* studies are needed to confirm the rTS/TS or rTS/DHFR protein-protein interactions and characterize their effect on TS or DHFR function.

5.4. Conclusions

Human rTSγ has been characterized *in vitro* as an L-fuconate dehydratase that belongs to the mandelate racemase subgroup of the enolase superfamily. The additional N-terminal amino

acids in rTSγ compared to rTSβ are important for catalysis and stability, refuting previous proposals that rTSγ is merely a variant of rTSβ with an N-terminal mitochondrial signaling domain. rTSβ has been implicated in the resistance of tumor cell lines to methotrexate and 5-fluorouracil. Unfortunately, rTSβ does not catalyze any chemical modification of these chemotherapeutics. Therefore, we believe attempts to sensitize resistant cells lines via inhibitors to rTSβ will not be successful. Based on the crystal structure of rTSγ and the enzymatic screening performed in this study, we do not believe rTSβ is catalytically active. Therefore, hypotheses that involve protein-protein interactions are more likely the source of the observed phenotypes and should be investigated further.

CHAPTER 6: CONCLUSIONS

6.1. Mannonate dehydratases in the enolase superfamily

6.1.1. Sequential versus functional divergence in the ManDs

In this study it was found that the D-mannonate dehydratase subgroup of the enolase superfamily (ManD) was not only multifunctional [with the discovery of the D-gluconate dehydratase activity (GlcD)], but also displayed a variance in catalytic efficiency: 1) high efficiency ($k_{cat}/K_M = 10^3$ to 10^4 M^{-1} s^{-1}) and specific for D-mannonate; 2) low efficiency ($k_{cat}/K_M = 10^1$ to 10^2 M^{-1} s^{-1}) and specific for either D-mannonate, D-gluconate, or promiscuous for both; and 3) no-activity with either D-mannonate or D-gluconate (or any acid sugar in the library). When these activities are mapped onto a sequence similarity network (SSN), one can visualize how function diverges as sequence diverges in the subgroup [52]. Surprisingly, these distinct functions do not separate until a fairly stringent cut-off is applied: >70% sequence identity. The deconvolution of function in this subgroup exemplifies how razor thin the line is between structure and function, or rather how sensitive the balance is between enzyme sequence and catalysis.

6.1.2. ManD/GlcD impact on synthetic biology/metabolic engineering

The information gathered on the ManD/GlcD disparity in the ManD subgroup makes an impact in the field of metabolic engineering, where researchers work to optimize cellular processes with the end goal of increasing production of a certain substance. Generically, this is done by one of two methods: improving flux through the metabolic process or removing flux

through pathways which deplete from the pathway of interest [104]. Additionally, synthetic biology can be coupled with metabolic engineering to devise new pathways [105]. In the area of advanced biofuel production, one sees a combination of these strategies since advanced biofuels [longer chain alcohols (≥ C4) and fatty-acid based fuels] are generally produced in such low abundance naturally [106]. Researchers are utilizing synthetic biology to design metabolically engineered biosynthetic pathways or redesign naturally occurring pathways to increase yields of advanced biofuels [107-110]. The subtle differences observed in ManD/GlcD activity can help synthetic biologists discern how novel chemistry can arise in similar protein scaffolds; knowledge which can then be applied to the development or optimization of novel biosynthetic pathways.

6.1.3. Physiological role of high efficiency ManDs

Prior to this dissertation, there was no verified physiological role for ManDs. It was discovered that a ManD in *Caulobacter crescentus* NA1000 (D-mannonate dehydration to 2-keto-3-deoxy-D-mannonate with a k_{cat}/K_M of 1.2 x 10^4 M^{-1} s^{-1} [52]) was responsible for D-mannonate dehydration in D-glucuronate metabolism via RT-PCR, phenotypic knockout studies, and complementation [111]. The protein UxuA (**u**tilization of he**xu**ronate **A**) was originally characterized as the mannonate dehydratase in this pathway [29]. It was discovered that every organism with a high efficiency ManD lacks UxuA in its genome [52]. Alternate pathways exist to metabolize D-glucuronate (www.metacyc.org); therefore, this incidence of convergent evolution was not obvious and needed experimental verification. Now validated, this physiological function can be confidently applied to all high efficiency ManDs discovered, thus

far resulting in the simultaneous annotation of 55 enzymes that were largely misannotated previously [52].

6.1.4. Physiological activity of low efficiency ManDs

The discovery of low-efficiency GlcDs and ManDs calls into question what value of k_{cat}/K_M is physiologically relevant: Can a low efficiency enzyme be physiologically relevant if enough substrate is present in the cell/enough enzyme is produced to overcome a low k_{cat}? Or is low efficiency an indication that other factors are needed for full activity (i.e. cofactors, protein-protein interactions, different substrate, etc.)? Organisms with low efficiency ManDs generally also contain the *Uxu*A gene within their genomes and, therefore, are thought to be involved in pathways other than D-glucuronate metabolism. In *Chromohalobacter salexigens* DSM3043, a low efficiency ManD (dehydration of D-mannonate with a k_{cat}/K_M of 5 M^{-1} s^{-1} and dehydration of D-gluconate with a k_{cat}/K_M of 5 M^{-1} s^{-1}: *Cs*ManD) was investigated with the aim of finding a physiological role. The genome neighborhood surrounding *Cs*ManD was indicative of L-gulonate or D-mannonate metabolism. Additionally, a UxuA protein is present in this genome neighborhood. Through extensive *in vivo* experiments (RT-PCR and phenotypic knockout studies), it was proven that the UxuA is physiologically relevant to the metabolism of L-gulonate and D-mannonate (preferred pathway). *Cs*ManD was not found to be physiologically relevant for either pathway. While this result does not answer the question put forth, it does reveal a novel physiological role for a previously characterized enzyme – representing a unique form of functional discovery.

Similar studies have been performed in *Salmonella enterica enterica serovar Enteritidis* P125109 with a redundant L-idonate metabolizing pathway involving a low efficiency GlcD

(unpublished data). This ManD also failed to display a physiological role in our phenotypic knockout studies. Given these two test cases and the findings of Huisman and Kolter [112], it is postulated that low efficiency ManDs are not directly involved in metabolism, but may have a, heretofore undiscovered, regulatory role within the cell related to starvation sensing.

6.1.5. No-activity ManDs

Many no-activity ManDs are located in genome neighborhoods containing genes encoding components of downstream metabolism of the ManD product, 2-keto-3-deoxy-D-mannonate: 2-keto-3-deoxy-D-mannonate kinase and 2-keto-3-deoxy-D-mannonate-6-phosphate aldolase. Despite this fact, any attempt to elicit enzyme activity *in vitro* failed. Variations in pH, temperature, salt concentration, osmolarity (glycerol/betaine), and metal cations, as well as the presence of the reducing agent dithiothreitol and nucleotide mono-, di-, and triphosphates were tested to see if these enzymes were activated by any of these conditions/molecules [52]. Furthermore, C6 phosphorylated D-mannonate, D-gluconate, D-allonate, and D-altronate were tested for activity with Uniprot ID A5KUH4 (no-activity protein from *Vibrionales bacterium* SWAT-3 with downstream metabolism genes nearby) since they are all susceptible to dehydration to 2-keto-3-deoxy-D-mannonate-6-phosphate (unpublished data). Still, no activity was observed. Many of these organisms have UxuAs in their genomes so it is unclear which carbon source would require ManD for metabolism; therefore, no knockout studies have been attempted. Furthermore, none of the no-activity proteins characterized herein are from organisms which are considered genetically tractable. The leading hypothesis is that these enzymes function in multienzyme complexes (channeling) or utilize a substrate our lab does not yet have access to (presumabley an acid sugar or carbohydrate due to genome contexts).

6.1.6. Reflections on annotation

In this dissertation, the previously described *in vitro* characterization of the ManDs covered the majority of the sequence space within the ManD subgroup. As such, a clear picture of the boundaries between function and sequence identity was observed. A log(BLAST E-value) of -190 was needed to fully separate the diverse functions of the ManD subgroup. This corresponds to ~75% identity (**Figure 6.1**). In 2002, Rost examined the current state of automated annotations and found that most group's calculations had found that between 50 and 70% sequence identity should be used to transfer function with confidence [113]. Furthermore, in 2003 it was concluded that over 60% sequence identity was needed to transfer all four EC numbers with 90% accuracy [32]. Both of these conclusions fall short of the ~75% identity required by the ManD subgroup to accurately transfer function. Therefore, the ManD subgroup is a strong example of how difficult it is to annotate proteins via strictly homology-based methods. Some experimental characterization is still required, but it is necessary to couple that with intelligent interpretation of SSNs to transfer the functions accurately to a large number of homologous proteins.

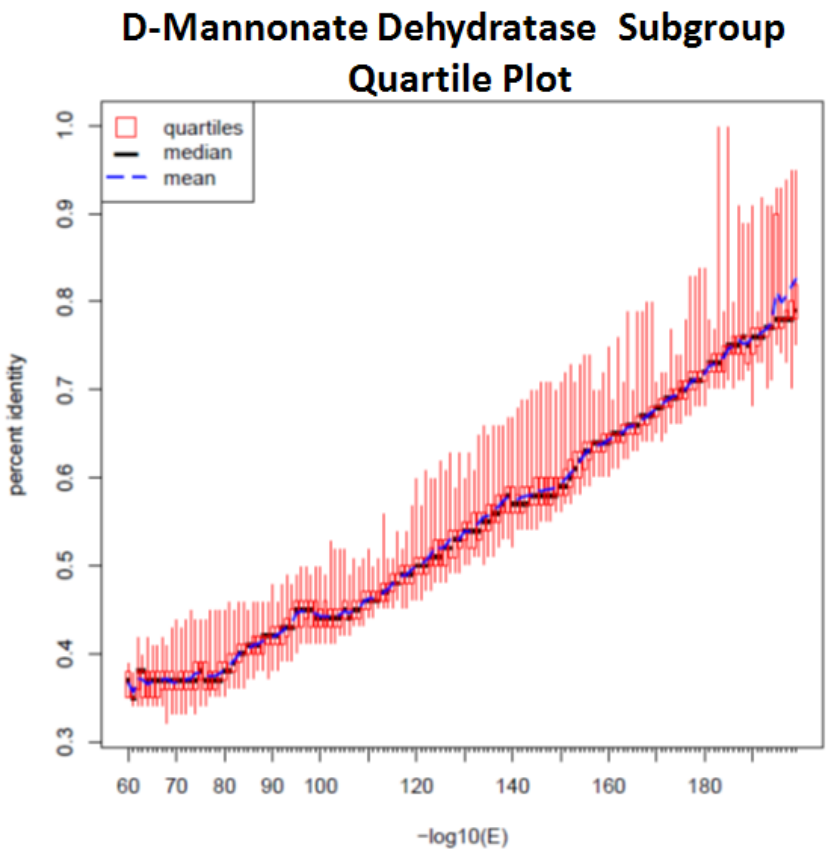


Figure 6.1. A quartile plot for the ManD subgroup. This displays the relationship between -log(BLAST E-value) and sequence identity. The "functional cut-off", which determines isofunctional clusters for this subgroup, has a –log10(E) value of 190 or ~75% identity.

6.1.7. Hypotheses on enzyme evolution in the ManD subgroup

When one visualizes the experimentally characterized members of the ManD subgroup with a phylogenetic tree (http://www.phylogeny.fr/), a clear dichotomy is seen (**Figure 6.2**). Standouts like EFI Target IDs 501675 and 501645 can be explained by their highly divergent nature: singletons in the ManD SSN at a log(BLAST E-value) of -190. More interestingly, this phylogenetic tree illustrates that the no activity proteins and low activity GlcDs diverge from the D-mannonate dehydrating ManDs immediately in the phylogenetic tree. The next branch on the

ManD side separates the high activity and low activity ManDs with the exception of the promiscuous ManD/GlcDs which branch with the high activity ManDs. The ManD subgroup phylogenetic tree taken as a whole can be seen as a snapshot of the subgroup's evolution, where some organisms have no need for a ManD in D-glucuronate metabolism and have thusly "repurposed" it. The logical evolutionary flow would be from high activity ManD to GlcD or no-activity protein; evidence for this comes in the form of the gene duplications present in some organisms with high-efficiency ManDs [52]. Specifically in *Caulobacter crescentus* NA1000, two high-efficiency ManDs exist (Uniprot IDs B8GZZ7 and B8H1R9), but only one was found to be physiologically relevant (B8GZZ7) [111]. These genes are highly similar (~80% identical), which indicates gene duplication may have taken place and B8H1R9 is in the process of evolving a new function.

The His residue at the end of the 7^{th} β-strand is a prime example of how evolution has occurred in this subgroup. It shows how evolution has taken advantage of a conserved residue that could serve multiple purposes and derived a new function from those possibilities (i.e. hydrogen bonding partner vs catalytic residue). Granted, GlcDs have low catalytic efficiency, but perhaps their transformation is not yet complete. Rather than explaining the lack of activity seen in no-activity proteins with the limitations of our modest acid sugar library, it is possible they are in the midst of their own "repurposing" and have yet to fully evolve their next function.

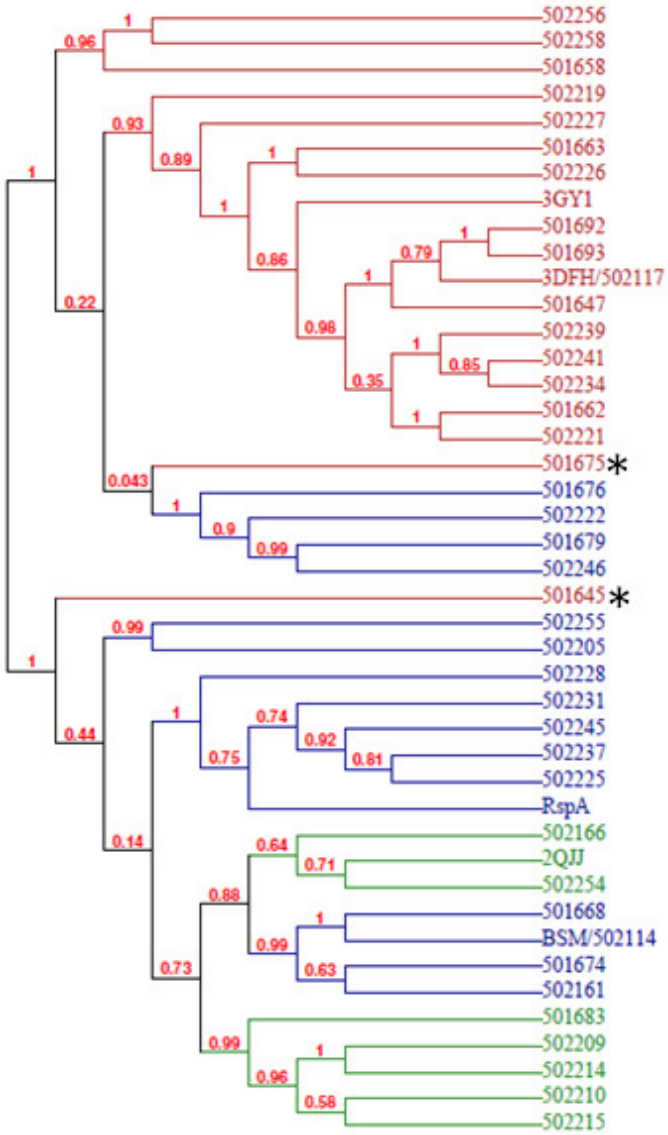


Figure 6.2. A phylogenetic tree for the characterized members of the ManD subgroup. High activity ManDs are in green. Low activity ManDs in blue. No activity proteins in red. It should be noted that 501676 to 502246 are specific to D-gluconate. Highly divergent ManDs [singletons on the ManD SSN at a –log(BLAST E-value) of 190] are marked with an asterisks.

6.2. L-Fuconate dehydratases in the mandelate racemase subgroup of the enolase superfamily

6.2.1. rTSγ characterization and insights into rTSβ overexpression phenotypes

It was discovered that human reverse thymidylate synthase γ (rTSγ) acts as an L-fuconate dehydratase, which catalyses the *in vitro* dehydration of L-fuconate to 2-keto-3-deoxy-L-fuconate: k_{cat} of 0.5 ± 0.006 s^{-1} and k_{cat}/K_M of 2.5 x 10^3 M^{-1} s^{-1} [114]. This is the only human enolase superfamily member to have been enzymatically characterized to date, other than the namesake enolase. rTSβ, an isoform of rTS γ, has been linked to cancer via implications of involvement in the resistance of tumor cell lines to methotrexate and 5-fluorouracil [73-78]. Unfortunately, we were unable to find any reactivity with rTSβ on either chemotherapeutic compound or our acid sugar library. The 27 N-terminal amino acids of rTSγ, which are missing in rTSβ, were found here to be essential for catalysis and ,therefore, are unlikely a mitochondrial signaling sequence for rTSβ as was previously hypothesized [80]. Based on the crystal structure of rTSγ and the truncated "blank" region observed in the rTSβ sequence, we do not believe rTSβ is catalytically active. Therefore, rTSβ likely contributes to the observed resistance phenotypes via mechanisms independent of dehydratase activity, such as protein-protein interactions. Evidence for the association of rTSβ with thymidylate synthase and dihydrofolate reductase has been observed elsewhere, and is a promising hypothesis for the rTSβ phenotypes seen [75, 103].

6.3 The Enzyme Function Initiative

6.3.1. ManD-related advances in the EFI

Since the work described in this dissertation was completed, many advances in the EFI have been validated by the successes of the work in the ManD subgroup. First and foremost, the generation of SSNs has become trivial with the EFI-Enzyme Similarity Tool (EFI-EST; http://efi.igb.illinois.edu/efi-est/). The initial SSN generated for functional discovery in the ManD subgroup (Chapter 2) was quite labor intensive (constructed by Dr. Shoshana Brown, Babbitt lab, UCSF). Through the characterization of the ManD subgroup, along with other EFI projects, it was found that SSNs were essential for visualizing protein similarity in large datasets, which in turn is used to select targets in large-scale functional discovery projects and aide in the accurate transfer of function between homologous proteins. EFI-EST uses a web interface for automated creation of SSNs for any protein family (Pfam or InterPro) or for individual protein sequence BLASTs (2000 returned sequences). In a matter of hours a quartile plot and SSN are generated with minimal input from the user. This tool has made the generation of SSNs accessible to all researchers, not just those proficient in bioinformatics.

Additionally, the manual analysis of the genome neighborhoods of every ManD proved useful in determining targets as well as for predicting functions/pathways. Naturally, manually determining the genome neighborhoods of a large group of proteins can be a bottleneck in the functional discovery process. Genome neighbors are extremely useful in functional discovery and a more efficient method was needed. The EFI has since developed the EFI-Genome Neighborhood Tool (EFI-GNT) tool (http://efi.igb.illinois.edu/efi-gnt/). Similar to EFI-EST, EFI-GNT is an automated, web-based tool that requires little input from the user (a SSN is all

that is required). The program generates a wheel-and-spoke genome neighborhood network (GNN) that connects SSn clusters with commonly occurring Pfam protein families, in an effort to summarize the enzymatic capacity of the genome neighbors and shed light on the function of the SSN query sequences. A colored-coded version of the user-submitted SSN allows the user to visually recognize which clusters are present in each wheel-and-spoke in the GNN. EFI-GNT is another researcher-friendly tool developed by the EFI to aid in the quick and accurate characterization of proteins of unknown function.

Furthermore, the work done in *Chromohalobacter salexigens* DSM3043 (Chapter 4) verified that ThermoFluor of transcriptional regulators could be a useful strategy in determining the substrate for a particular genome neighborhood. The initial idea for ThermoFluor was to screen the solute-binding domains of ATP-binding cassette (ABC) and Tripartite ATP-independent periplasmic (TRAP) transporters. This was done on a large scale and many ligands were found (used in Chapter 3 for a TRAP DctP, but no hits were seen in that case). The information is being formatted for publication at the time of this dissertation's preparation. The "*Chromohalobacter* story" verified that this strategy could also be applied to GntR transcriptional regulators (generally inhibitors). GntRs contain an N-terminal DNA-binding domain and a C-terminal solute-binding domain. It was unknown how the DNA-binding domain effects the screening hits. The results indicate that useful information can be gleaned from screening the entire protein without having to engineer truncations, and that this is a useful strategy for gaining insights into the potential substrate of a particular genome neighborhood.

6.3.2. Future of the EFI

The EFI has taken great strides during the work described in this dissertation, and will continue to do so in the future. The current future goals of the EFI are to correlate physiological data to ThermoFluor screening of solute binding proteins. This will likely result in the deorphanization of many proteins of unknown function, and hopefully elucidate novel metabolic pathways. A recent collaboration with Dr. Andrei Osterman from the Sanford-Burnham Institute has given the EFI a new lens to visualize potential metabolic pathways. The Osterman group takes advantage of the fact that many genes organize into genetic motifs that share common regulators (determined by proximal conserved transcription factor binding sites) and have co-occurrence across many related organisms to identify novel metabolic pathways. This approach is being combined with ThermoFluor data of solute binding proteins to fill in the "blanks" of incomplete metabolic networks.

The EFI is also moving into the realm of the human microbiome. Preparations are underway to accommodate the large-scale purification of proteins in an anaerobic environment, and also to facilitate anaerobic cultivation/genetic manipulation of these source bacteria. In this manner, the EFI hopes to apply its functional discovery prowess to the human microbiota and discover new metabolic/syntrohic pathways, enzymatic reactions, and drug targets which could be important for human health.

APPENDIX: A

Crystallization conditions

The crystals of high-activity ManD/EFI target 502214 from *Caulobacter* sp. K31 liganded with Mg^{2+}, Cl^-, and glycerol (4FI4) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (17.4 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl, 10% glycerol, 5 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.16 M $MgCl_2$, 0.08 M Tris-HCl, pH 8.5, 24% PEG4K, 20% glycerol.

The crystals of high-activity ManD/Uniprot ID Q9AAR4 from *Caulobacter crescentus* CB15 liganded with Mg^{2+}, Cl^-, D-mannonic acid, CO_3^{2-}, and glycerol (4GME) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (17 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl, 10% glycerol, 5 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.1 M ammonium sulfate, 0.1 M Tris-HCl, pH 8.5, 25% PEG3350, and 10% glycerol.

The crystals of high-activity ManD/Uniprot ID Q9AAR4 from *Caulobacter crescentus* CB15 liganded with Mg^{2+}, Cl^-, CO_3^{2-}, and glycerol (3VCN) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (17 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.1M bicine, pH 9.0, and 20% PEG6000.

The crystals of low-activity ManD/Uniprot ID B3PDB1from *Cellvibrio japonicus* Ueda107 liganded with Mg^{2+}, Cl^-, 2-[N-cyclohexylamino]ethane sulfonic acid, and glycerol (3V3W) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (10.1 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl,

10% glycerol, 5 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.1 M CHES/NAOH pH 9.5, 30% PEG3000, and 15% glycerol.

The crystals of D-mannonate dehydratase from *Cellvibrio japonicus* Ueda107 liganded with Mg^{2+}, Cl^-, and L-tartaric acid (3V4B) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (10.1 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl, 10% glycerol, 5 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.2M KNaTartrate, 20% Peg3350, and 20% glycerol.

The crystals of the mutant P317A of promiscuous ManD/GlcD/Uniprot ID Q1QT89 from *Chromohalobacter salexigens* liganded with Mg^{2+} and D-gluconate (3QKF) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (12 mg/mL) in 10 mM Tris (pH 8.0), 100 mM NaCl, 5 mM $MgCl_2$, and 40 mM D-gluconate; the precipitant contained 15% PEG 3350, 0.1 M succinic acid (pH 7.0), and 5 mM $MgCl_2$.

The crystals of promiscuous ManD/GlcD/ Uniprot ID Q1QT89 from *Chromohalobacter salexigens* DSM 3043 liganded with Na^+, Cl^-, and glycerol (4F4R) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (13.3 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.2 M sodium acetate, 0.1 M Tris-HCl pH 8.5, 30% Peg4000, and 20% glycerol.

The crystals of promiscuous ManD/GlcD/ Uniprot ID Q1QT89 from *Chromohalobacter salexigens* DSM 3043 liganded with Mg^{2+}, SO_4^{2-}, and glycerol (3OW1) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (26 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM

$MgCl_2$; the precipitant contained 1.5 M ammonium sulfate, 0.1 M Bis-Tris, 0.1 M sodium chloride, pH 6.5.

The crystals of promiscuous ManD/GlcD/ Uniprot ID Q1QT89 from *Chromohalobacter salexigens* DSM 3043 liganded with Mg^{2+} and glycerol (3PK7) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (26 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 15% PEG3350, 0.1 M succinic acid, pH 7.0.

The crystals of promiscuous ManD/GlcD/ Uniprot ID Q1QT89 from *Chromohalobacter salexigens* DSM 3043 liganded with Mg^{2+}, D-mannonic acid, and 2-keto-3-deoxygluconate (3P93) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (26 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.3 M magnesium formate, 0.1M bis-tris, pH 5.5.

The crystals of promiscuous ManD/GlcD/ Uniprot ID Q1QT89 from *Chromohalobacter salexigens* liganded with Mg^{2+}, and D-gluconate (3QKE) were grown by sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (26 mg/mL) in 10 mM Tris (pH 8.0), 5 mM $MgCl_2$; and 40 mM D-gluconate; the precipitant contained 20% PEG3350, 0.2 M sodium malonate (pH 7.0), and 5 mM $MgCl_2$.

The crystals of promiscuous ManD/GlcD/ Uniprot ID Q1QT89 from *Chromohalobacter salexigens* liganded with Co^{2+} and D-arabinohydroxamate (3RGT) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (26 mg/mL) in 10 mM Tris (pH 8.0), and 40 mM D-arabinohydroxamate; the precipitant contained 20% polyvinylpyrrolidone K15, 0.1 M Tris (pH 8.5), and 10 mM $CoCl_2$.

The crystals of no-activity protein/Uniprot ID A6M2W4 from *Clostridium beijerinckii* liganded with Mg^{2+} (3S47) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (37 mg/mL) 10 mM Tris (pH 8.0), 100 mM NaCl, and 5 mM $MgCl_2$; the precipitant contained 45% 2-methyl-2,4-pentanediol, 0.1 M Bis-Tris (pH 5.5), 0.2 M ammonium acetate, and 5 mM $MgCl_2$.

The crystals of no-activity protein/Uniprot ID A4W7D6 from *Dickeya dadantii* Ech703 liganded with Mg^{2+}, formic acid, I^-, Cl^-, and glycerol (4IHC) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (16.4 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.2 M ammonium iodide, pH 7.5, 20% PEG3350.

The crystals of low-activity ManD/Uniprot ID Q8FHC7 from *Enterobacter* sp. 638 liganded with Mg^{2+}, Cl^-, and glycerol (3TJI) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (16.2 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl, 10% glycerol, 5 mM DTT, 5 mM $MgCl_2$; the precipitant contained 2.5 M NaCl, 100 mM Tris pH 7.5, 200 mM $MgCl_2$, 20% glycerol.

The crystals of no-activity protein/Uniprot ID C8ZZN2 from *Enterococcus gallinarum* EG2 liganded with Mg^{2+}, Cl^-, and glycerol (4HNL) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (19.7 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl, 10% glycerol, 5 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.2 M $MgCl_2$, pH 8.5, 0.1 M Tris-HCl, and 16% PEG MME4000.

The crystals of low-activity ManD/Uniprot ID Q8FHC7 from *Escherichia coli* CFT073 complexed with Mg^{2+} (4IL2) were grown by the sitting drop vapor diffusion method at 9° C. The crystallization drop contained protein solution (10 mg/ml) in 20 mM Tris (pH 7.5), 100 mM

NaCl, 50 mM imidazole, and 10 mM D-mannonate; precipitant contained 25% PEG 1,500 and 20% v/v glycerol.

The crystals of the loop mutant (V161A, R163A, K165G, L166A, Y167G, Y168A, E169G) of high-activity D-mannonate dehydratase (ManD) from *Novosphingobium aromaticivorans* liganded with Mg^{2+} and glycerol (4K8G) were grown by the sitting drop vapor diffusion method at 4° C. The crystallization drop contained protein solution (20 mg/ml) in 50 mM Tris (pH 7.5), 100 mM NaCl, 5 mM $MgCl_2$, and 10 mM D-mannonate; the precipitant contained 28% PEG 400, 100 mM HEPES-Na^+, and 200 mM $CaCl_2$.

The crystals of the mutant A314P of high-activity D-mannonate dehydratase from *Novosphingoium aromaticivorans* liganded with Mg^{2+} and D-mannonate (3R4E) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (12 mg/mL) in 100 mM imidazole, 20 mM Tris (pH 8.0), 150 mM NaCl, 5 mM $MgCl_2$, and 40 mM D-mannonate; the precipitant contained 30% PEG 400, 0.1 M HEPES (pH 7.5), and 5 mM $MgCl_2$.

The crystals of low-activity GlcD/Uniprot ID D4GJ14 from *Pantoea ananatis* LMG 20103 liganded with Mg^{2+}, Cl^-, 1,2-ethanediol, and gluconic acid (3T6C) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (13.1 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl, 10% glycerol, 5 mM DTT, 5 mM $MgCl_2$; the precipitant contained 1 M K/Na tartrate and 100 mM MES pH 6.0; the soak contained 1 M K/Na tartrate, 100 mM MES pH 6.0, 20% ethylene glycol, and 70 mM sodium gluconate.

The crystals of D-mannonate dehydratase from *Pectobacterium carotovorum subsp. carotovorum* PC1 liganded with Mg^{2+}, formic acid, Cl^-, and glycerol (4E4F) were grown by the

sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (17.9 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.17 M sodium acetate, pH 8.5, 0.085 M Tris-HCl, 25.5% PEG4000, 15% glycerol.

The crystals of low-activity GlcD/Uniprot ID B5R541 from *Salmonella enterica subsp. enterica serovar Enteritidis str.* P125109 liganded with Cl^- and glycerol (3TW9) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (15.2 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.2 M sodium chloride, 100 mM sodium acetate, pH 4.6, and 30% MPD.

The crystals of low-activity GlcD/Uniprot ID B5R541 from *Salmonella enterica subsp. enterica serovar Enteritidis str.* P125109 liganded with Mg^{2+}, Cl^-, and glycerol (3TWA) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (15.2 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.2 M ammonium acetate, 100 mM Tris-HCl pH 8.5, and 25% PEG3350.

The crystals of low-activity GlcD/Uniprot ID B5R541 from *Salmonella enterica subsp. enterica serovar Enteritidis str.* P125109 liganded with Cl^- and glycerol (3TWB) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (15.2 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 0.2 M sodium chloride, 100 mM sodium acetate, pH 4.6, and 30% MPD.

The crystals of high-activity ManD/Uniprot ID Q1NAJ2 from *Sphingomonas sp.* SKA58 liganded with Mg^{2+}, Cl^-, and glycerol (3THU) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (9.0 mg/mL) in 10 mM HEPES pH 7.8, 150 mM NaCl, 10% glycerol, 5 mM DTT, 5 mM $MgCl_2$; the precipitant contained 25% PEG3350, 100 mM TRIS pH 8.5, 200 mM $MgCl_2$, and 20% glycerol.

The crystals of no-activity protein/Uniprot ID A5KUH4 from *Vibrionales* bacterium liganded with Mg^{2+}, and glycerol (3R25) were grown by sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (44 mg/mL) in 10 mM Tris (pH 8.0), and 5 mM $MgCl_2$; the precipitant contained 0.5 M magnesium formate, and 0.1 M Bis-Tris (pH 6.5).

The crystals of the no-activity protein/ Uniprot ID A5KUH4 from *Vibrionales* bacterium liganded with Mg^{2+} and D-arabinonate (3SBF) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (116 mg/mL) in 10 mM Tris (pH 7.8), 5 mM $MgCl_2$, and 40 mM D-arabinonate; the precipitant contained 10% PEG3350, 0.1 M HEPES (pH 7.5), 0.2 M L-proline, and 5 mM $MgCl_2$.

The crystals of no-activity protein/Uniprot ID D0X4R4 from *Vibrio harveyi* liganded with Mg^{2+}, Cl^-, glycerol, maleic acid, and malonic acid (4GIS) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (13.56 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 35% Tascimate pH 7.0, 20% glycerol and 50 mM $MgCl_2$.

The crystals of no-activity protein/Uniprot ID D0X4R4 from *Vibrio harveyi* liganded with Mg^{2+}, Cl^-, 1,2-ethanediol, Na^+, and SO_4^{2-} (4GIR) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (13.56 mg/mL)

in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 2 M ammonium sulfate, 20% ethylene glycol, and 50 mM $MgCl_2$.

The crystals of no-activity protein/Uniprot ID D0X4R4 from *Vibrio harveyi* liganded with Mg^{2+}, Cl^-, 1,2-ethanediol, glycerol, 4-(2-hydroxyethyl)-1-piperazine ethanesulfonic acid (4GGH) were grown by the sitting drop vapor diffusion method at room temperature. The protein solution contained the protein (13.56 mg/mL) in 10 mM HEPES pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM DTT, 5 mM $MgCl_2$; the precipitant contained 200 mM NaCl, 0.1 M HEPES, 25% PEG3350, 20% ethylene glycol, 50 mM $MgCl_2$, and pH 7.5.

L-glycerate	L-xylonate	D-glucarate/L-gularate	L-galactonate	L-alturonate
D-glycerate	D-xylonate	L-glucarate/D-gularate	D-galactonate	D-alturonate
D-glycerate 3-P	L-ribonate	L-talonate	L-fuconate	L-alluronate
meso-tartrate	D-ribonate	D-talonate	D-fuconate	D-alluronate
L-tartrate	L-lyxonate	L-mannonate	6-deoxy-L-talonate	L-taluronate
D-tartrate	D-lyxonate	D-mannonate	L-rhamnonate	D-taluronate
L-malate	L-arabinonate	L-idonate	L-galacturonate	L-glucuronate
D-malate	D-arabinonate	D-idonate	D-galacturonate	D-glucuronate
L-threonate	D/L allarate	L-gluconate	L-galactonate 6-P	
D-threonate	D/L galactarate	D-gluconate	D-gluconate 6-P	
D-erythronate	L-altrarate/D-talarate	L-allonate	L-mannuronate	
L-erythronate	D-altrarate/L-talarate	D-allonate	D-mannuronate	
D/L xylarate	L-mannarate	L-gulonate	L-iduronate	
D/L ribarate	D-mannarate	D-gulonate	D-iduronate	
L-arabinarate/D-lyxarate	L-idarate	L-altronate	L-guluronate	
D-arabinarate/L-lyxarate	D-idarate	D-altronate	D-guluronate	

Figure A.1. The 72 acid sugars in the acid sugar library which was used for detection of dehydration activity.

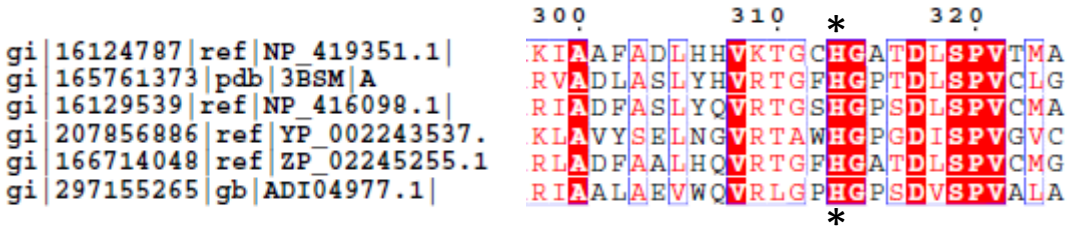


Figure A.2. The sequence alignment of one ManD from each catalytically active cluster in the ManD log(BLAST E-value) of -190 SSN (cluster 2, 3, 4, 5, 7, and 8) shows that His315 of *Cs*ManD is conserved. This alignment was generated using ESPript (www. espript.ibcp.fr).

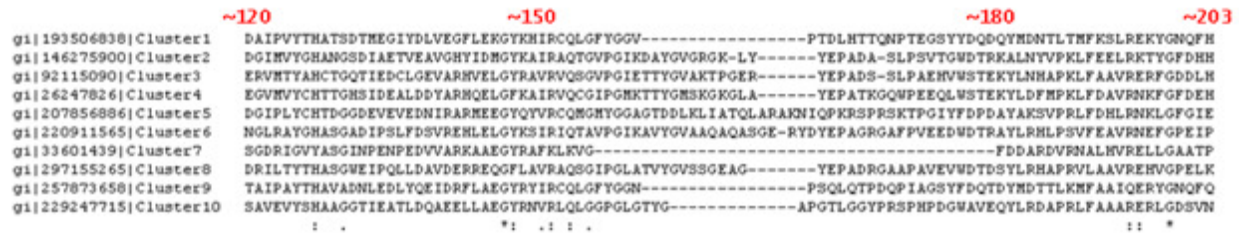


Figure A.3. The sequence alignment of one ManD from each cluster of screened ManDs, highlighting the "150-180s" loop region, shows a high degree of divergence. Most notably we can see a variance in loop length given by the gaps in alignment present. This alignment was generated using clustalW (http://embnet.vital-it.ch/software/ClustalW.html).

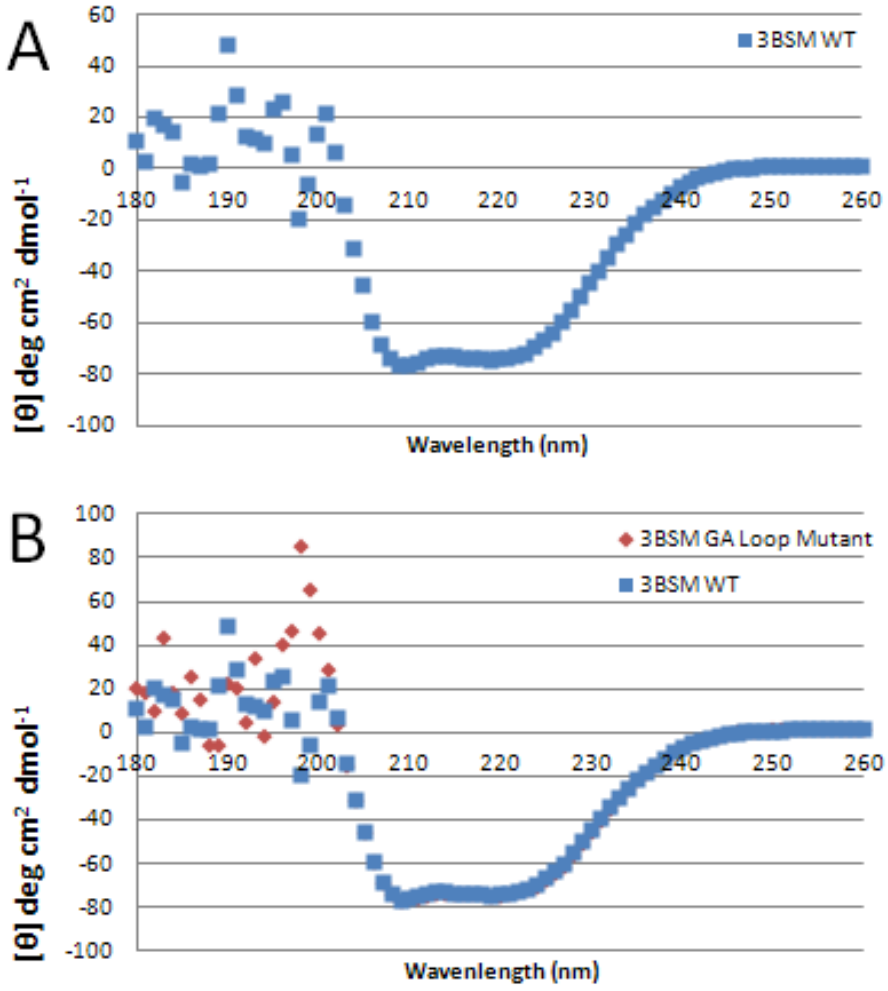


Figure A.4. Circular dichroism data shows that the secondary structure of 3BSM WT (blue squares) matches the secondary structure of 3BSM loop mutant (red diamonds. The low UV range is disrupted due to buffer conditions.

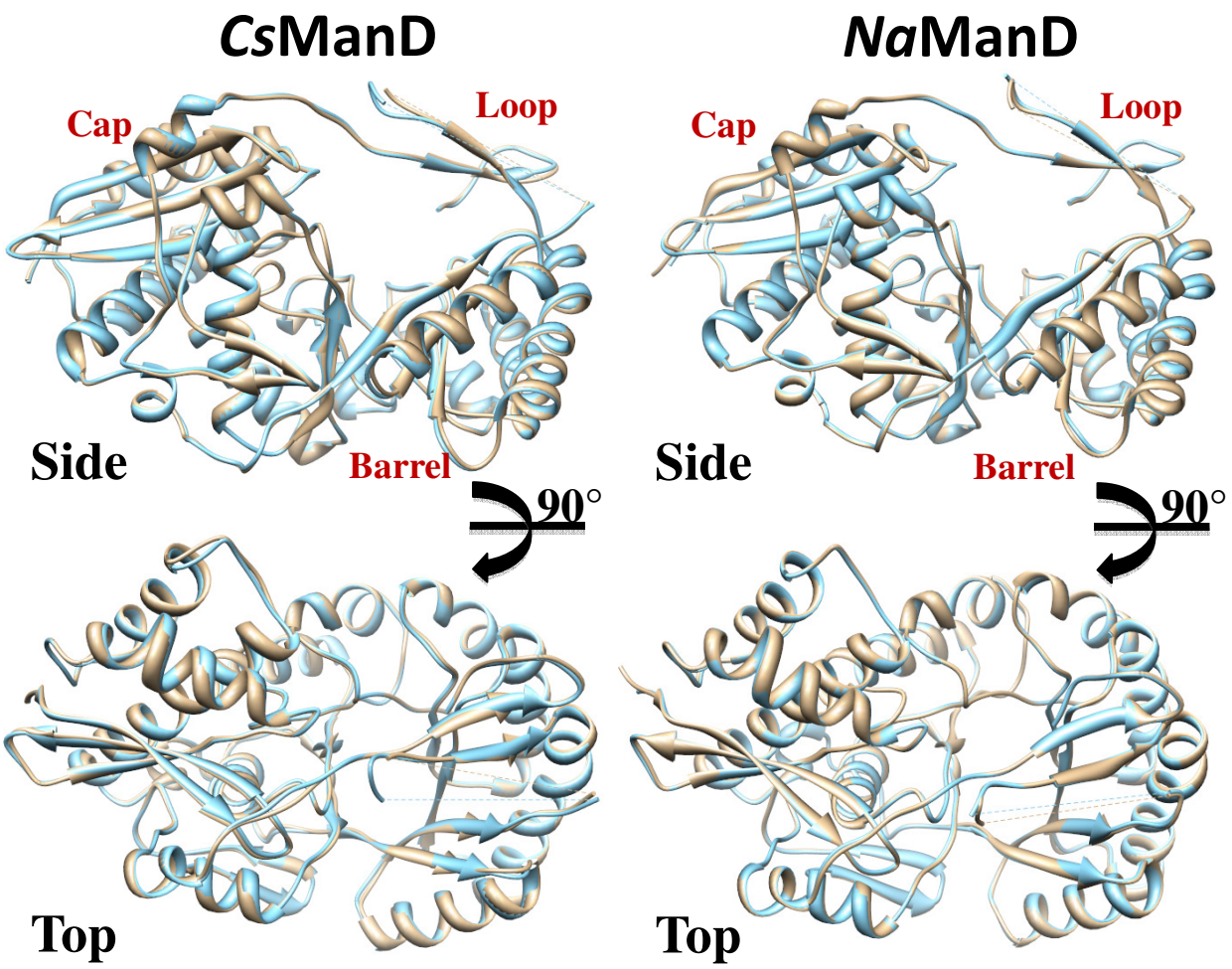


Figure A.5. The high degree of superposition seen in these structural overlays show that the GA loop mutants (tan) have the same secondary structure as wild type *Cs*ManD (3PK9, blue) or *Na*ManD (2QJM, blue). The "150-180s" loop is not ordered in any structure.

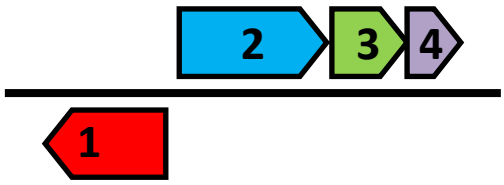


Figure A.6. The above genome neighborhood is an example of the genomic context in which no activity ManDs can be found. In this case, the downstream genes of ManD in glucuronate metabolism are present (KdgK and KdgpA). The annotations are as follows: (1) ManD, (2) phosphotransferase system protein, (3) KdgK, and (4) KdgpA.

Table A.1. SDM primers used in this study

Primer name	Sequence
*Na*ManD_GA_Loop_FOR	5'-CATCAAGGACGCCTATGGCGCGGGTGCCGGCGGGGCCGGCGGCGGACCGGCCGATGCCAGCCTCC -3'
*Na*ManD_GA_Loop_REV	5'-GGAGGCTGGCATCGGCCGGTCCGCCGCCGGCCCCGCCGGCACCCGCGCCATAGGCGTCCTTGATG -3'
*Cs*ManD_GA_Loop_FOR	5'-CGGGCATCGAGACAACGTATGGTGTGGCAGGAGCCGGAGGGGCAGGAGGTGGACCTGCGGACTCATC -3'
*Cs*ManD_GA_Loop_REV	5'-GATGAGTCCGCAGGTCCACCTCCTGCCCCTCCGGCTCCTGCCACACCATACGTTGTCTCGATGCCCG -3'
*Na*ManD_A314P_FOR	5'-GCTGCCACGGCCCGACCGACCTT -3’
*Na*ManD_A314P_REV	5’- AAGGTCGGTCGGGCCGTGGCAGC -3’
*Cs*ManD_P317A_FOR	5'-GGGATTCCATGGCGCGACCGATCTGTC -3’
*Cs*ManD_P317A_REV	5’- GACAGATCGGTCGCGCCATGGAATCCC -3’
*Cs*ManD_H315N_FOR	5'-CCGTACGGGATTCAATGGCCCGACCGA -3'
*Cs*ManD_H315N_REV	5'- GCATGCCCAAGTTACCGGGCTGGCT -3'
*Cs*ManD_H315Q_FOR	5'-GTCCGTACGGGATTCCAGGGCCCGACC -3'
*Cs*ManD_H315Q_REV	5'-CAGGCATGCCCTAAGGTCC CGGGCTGG -3'

Table A.2. Data collection and refinement statistics for D-mannonate dehydratases

	Caulobacter sp. K31 WT	*Caulobacter crescentus* CB15 WT	*Caulobacter crescentus* CB15 WT	*Cellvibrio japonicus* WT	*Cellvibrio japonicus* WT	*Chromohalobacter salexigens* WT
Data collection						
Space group	I422	I4	I4	I422	I422	$P2_1$
No. of molecules	3	2	2	1	1	8
in asym. unit						
Cell dimensions						
a (Å)	137.62	115.80	116.3	124.3	124.71	86.58
b (Å)	137.62	115.80	116.3	124.3	124.71	177.79
c (Å)	264.70	118.13	118.08	111.3	110.87	109.99
β°						102.62
Resolution (Å)	2.00	2.00	1.37	1.4	1.4	1.8
No. of unique	85719	53061	163715	85189	85453	295193
reflections						
R_{merge}	0.091	0.082	0.090	0.076	0.071	0.081
Completeness (%)	100.0	91.0	92.60	98.60	100.0	98.5
Refinement						
Resolution (Å)	47.7-2.0	50.0-2.0	50.0-1.45	32.1-1.4	32.14-1.4	25.0-1.8
R_{cryst}	0.151	0.157	0.130	0.137	0.148	0.161
R_{free}	0.204	0.196	0.188	0.152	0.161	0.194
No. atoms						
Protein	9479	6312	6312	3172	3076	24501
Waters	878	507	848	427	475	2195
Bound ligands	GOL, CL, MG, UNL	GOL, CL, MG, CO3, D-mannonate	MG, GOL, CL, CO3	MG, GOL, CL, NHE	TARTRATE, MG, CL	MG,GOL, SO4
Ligand atoms	84	27	32	40	12	149
R.m.s deviations						
Bond lengths (Å)	0.007	0.008	0.012	0.017	0.017	0.007
Bond angles (°)	1.222	1.223	1.454	1.732	1.628	1.1
PDB entry	4FI4	4GME	3VCN	3V3W	3V4B	3OW1

	Chromohalobacter salexigens WT	*Chromohalobacter salexigens* WT	*Chromohalobacter salexigens* WT	*Chromohalobacter salexigens* WT	*Chromohalobacter salexigens* WT	*Chromohalobacter salexigens* Mutant P317A
Data collection						
Space group	C2	P21212	P422	C2	$P2_12_12_1$	C2
No. of molecules	8	4	1	8	8	8
in asym. unit						
Cell dimensions						
a (Å)	195.54	110.71	111.39	195.27	110.24	196.33
b (Å)	85.78	179.66	111.39	85.76	167.26	85.79
c (Å)	195.14	85.42	84.58	195.10	168.86	195.64
β°	110.33			110.31		110.47
Resolution (Å)	1.55	1.9	1.8	1.64	1.8	1.7
No. of unique	398136	133334	49800	366762	286929	332505
reflections						
Rmerge	0.093	0.106	0.120	0.078	0.109	0.095
Completeness (%)	91.0	99.2	100.0	99.8	99.8	99.4
Refinement						
Resolution (Å)	25.0-1.55	25.0-1.9	30.9-1.4	25.0-1.64	25.0-1.8	25.0-1.7
Rcryst	0.179	0.178	0.146	0.158	0.199	0.168
Rfree	0.211	0.222	0.171	0.184	0.241	0.202
No. atoms						
Protein	25315	12349	3236	24894	24392	25043
Waters	2513	857	343	2889	1871	3144
Bound ligands	Mg, D-Gluconate	Co, D-arabinohydroxamate	MG, CL, NA	MG, GOL	D-mannonate MG, KDG	D-gluconate, MG, GOL, CL
Ligand atoms	128	52	14	57	108	124
R.m.s deviations						
Bond lengths (Å)	0.006	0.007	0.008	0.006	0.007	0.007
Bond angles (°)	1.06	1.06	1.052	1.0	1.1	1.0
PDB entry	3QKE	3RGT	4F4R	3PK7	3P93	3QKF

	Clostridium beijerincki WT	*Dickeya dadantii* Ech703 WT	*Enterobacter* SP. 638 WT	*Enterococcus gallinarum* EG2 WT	*Escherichia coli* CFT073 WT	*Novosphingobium aromaticivorans* Mutant A314P
Data collection						
Space group	$P42_12$	C2	P42212	I422	P4	$C222_1$
No. of molecules	2	8	4	1	4	4
in asym. unit						
Cell dimensions						
a (Å)	111.26	224.33	182.15	115.94	123.78	116.22
b (Å)	111.26	224.35	182.15	115.94	123.78	165.36
c (Å)	139.80	87.08	104.80	119.97	112.21	167.20
. β°					90.0, 90.0, 90.0	
Resolution (Å)	1.7	1.94	1.8	1.48	1.95 (1.95-2.0)	1.65
No. of unique	85180	277432	113519	67893	122958	191348
reflections						
Rmerge	0.096	0.19	0.129	0.056	0.116 (0.663)	0.088
Completeness (%)	88.1	88.2	99.2	99.9	99.9 (100.0)	98.9
Refinement						
Resolution (Å)	25.0-1.7	40.0-2.0	128.8-1.8	47.6-1.48	29.83-1.95	25.0-1.65
Rcryst	0.243	0.175	0.252	0.124	0.166	0.158
Rfree	0.285	0.228	0.282	0.162	0.198	0.183
No. atoms						
Protein	6183	24589	12620	3174	12364	12435
Waters	326	1978	1329	405	472	1003
Bound ligands	Mg	GOL, CL, MG, IOD, FMT	CL, MG, GOL	GOL, CL, MG	MG	D-mannonate MG, PGE, PDO
Ligand atoms	2	147	46	9	4	116
R.m.s deviations						
Bond lengths (Å)	0.007	0.01	0.011	0.008	0.01	0.006
Bond angles (°)	1.07	1.437	1.274	1.248	1.5	1.1
PDB entry	3S47	4IHC	3TJI	4HNL	4IL2	3R4E

	Novosphingobium aromaticivorans Loop mutant	*Pantoea ananatis* WT	*Pectobacterium carotovorum* PC1 WT	*Salmonella enterica* P125109 WT	*Salmonella enterica* P125109 WT	*Salmonella enterica* P125109 WT
Data collection						
Space group	I422	P4212	C2	P42212	I422	I422
No. of molecules	1	2	4	4	5	5
in asym. unit						
Cell dimensions						
a (Å)	125.77	130.79	204.28	177.53	144.68	144.73
b (Å)	124.77	130.79	86.34	177.53	144.68	144.73
c (Å)	119.74	104.68	114.46	110.15	446.80	447.00
β°	90.0, 90.0, 90.0		123.07			
Resolution (Å)	1.25 (1.25-1.28)	1.6	2.0	1.65	1.76	1.76
No. of unique	131204	110754	113519	209416	233074	232167
reflections						
Rmerge	0.093 (0.766)	0.1	0.077	0.129	0.177	0.18
Completeness (%)	100.0 (100.0)	99.0	99.2	100.0	100.0	99.8
Refinement						
Resolution (Å)	29.9-1.25	21.2-1.6	50.0-2.0	50.0-1.7	50.0-1.8	40.0-1.76
Rcryst	0.141	0.139	0.190	0.212	0.169	0.171
Rfree	0.148	0.168	0.245	0.252	0.206	0.206
No. atoms						
Protein	3072	6454	12295	12522	16120	16139
Waters	534	954	730	1171	2002	2016
Bound ligands	GLYCEROL, MG	D-GLUCONATE MG, CL, EDO	FORMATE, CL, MG, GOL	GLYCEROL, CL	GLYCEROL, CL, MG	GLUCONATE, CL, MG, GOL
Ligand atoms	7	67	90	35	61	115
R.m.s deviations						
Bond lengths (Å)	0.006	0.016	0.007	0.013	0.009	0.013
Bond angles (°)	1.1	1.771	1.034	1.448	1.274	1.497
PDB entry	4K8G	3T6C	4E4F	3TW9	3TWA	3TWB

	Sphingomonas SP. SKA58 WT	*Vibrionales bacterium* WT	*Vibrionales bacterium* P311A	*Vibrio harveyi* WT	*Vibrio harveyi* WT	*Vibrio harveyi* WT
Data collection						
Space group	I422	I222	P42212	I4122	P41212	C2221
No. of molecules	3	8	4	2	4	4
in asym. unit						
Cell dimensions						
a (Å)	124.8	167.99	163.37	181.20	146.84	92.95
b (Å)	124.8	211.52	163.37	181.20	146.84	209.55
c (Å)	343.09	211.67	110.41	115.83	169.52	208.47
. β°						
Resolution (Å)	1.8	1.6	1.5	1.8	2.0	1.9
No. of unique	123065	419783	232178	88533	113621	158060
reflections						
Rmerge	0.077	0.089	0.083	0.133	0.133	0.113
Completeness (%)	98.60	86.2	98.3	100.0	91.7	99.2
Refinement						
Resolution (Å)	171.55-1.8	25.0-1.6	25.0-1.5	29.0-1.8	38.2-1.8	93.6-1.8
Rcryst	0.138	0.161	0.180	0.152	0.149	0.169
Rfree	0.165	0.186	0.206	0.183	0.193	0.203
No. atoms						
Protein	9540	25388	12660	6489	12973	12760
Waters	1142	1827	1397	788	1357	1694
Bound ligands	MG, GOL, CL, UNX	Mg, GOL	Mg, D-Arabinonate	CL, MG, GOL, MAE, MLA	CL, MG, GOL, EDO, SO4,NA	CL, MG, GOL, EDO, EPE
Ligand atoms	59	58	79	39	54	141
R.m.s deviations						
Bond lengths (Å)	0.007	0.006	0.006	0.007	0.009	0.008
Bond angles (°)	1.057	0.98	1.06	1.118	1.088	1.096
PDB entry	3THU	3R25	3SBF	4GIS	4GIR	4GGH

Table A.3. List of all structures solved in the ManD subgroup.

PDB	Bound	Uniprot ID	Activity
4FI4	Glycerol, Mg^{2+}	B0T0B1	D-Mannonate (12000 $M^{-1}s^{-1}$)
3THU	Glycerol, Mg^{2+}	Q1NAJ2	D-Mannonate (4200 $M^{-1}s^{-1}$)
4GME	D-Mannonate, Mg^{2+}	Q9AAR4	D-Mannonate (12300 $M^{-1}s^{-1}$)
3VCN	Glycerol, Mg^{2+}	Q9AAR4	D-Mannonate (12000 $M^{-1}s^{-1}$)
2QJJ*	Mg^{2+}	A4XF23	D-Mannonate (3200 $M^{-1}s^{-1}$)
2QJM*,**	D-Mannonate, Mg^{2+}	A4XF23	D-Mannonate (3200 $M^{-1}s^{-1}$)
2QJN*	2-keto-3-deoxy-D-Gluconate, Mg^{2+}	A4XF23	D-Mannonate (3200 $M^{-1}s^{-1}$)
3R4E**	D-Mannonate, Mg^{2+}	A4XF23	D-Mannonate (3200 $M^{-1}s^{-1}$)
4K8G**	Glycerol, Mg^{2+}	A4XF23	D-Mannonate (3200 $M^{-1}s^{-1}$)
3V3W	Glycerol, Mg^{2+}	B3PDB1	D-Mannonate (100 $M^{-1}s^{-1}$)
3V4B	Tartrate, Mg^{2+}	B3PDB1	D-Mannonate (100 $M^{-1}s^{-1}$)
3BSM*	Apo	*Cs*ManD/Q1QT89	D-Mannonate (5 $M^{-1}s^{-1}$) D-Gluconate (40 $M^{-1}s^{-1}$)
3OW1	Sulfate, Mg^{2+}	*Cs*ManD/Q1QT89	D-Mannonate (5 $M^{-1}s^{-1}$) D-Gluconate (40 $M^{-1}s^{-1}$)
3PK7	Glycerol, Mg^{2+}	*Cs*ManD/Q1QT89	D-Mannonate (5 $M^{-1}s^{-1}$) D-Gluconate (40 $M^{-1}s^{-1}$)
4F4R	Glycerol, Mg^{2+}	*Cs*ManD/Q1QT89	D-Mannonate (5 $M^{-1}s^{-1}$) D-Gluconate (40 $M^{-1}s^{-1}$)
3P93	2-keto-3-deoxy-D-Gluconate, Mg^{2+}	*Cs*ManD/Q1QT89	D-Mannonate (5 $M^{-1}s^{-1}$) D-Gluconate (40 $M^{-1}s^{-1}$)
3QKE	D-Gluconate, Mg^{2+}	*Cs*ManD/Q1QT89	D-Mannonate (5 $M^{-1}s^{-1}$) D-Gluconate (40 $M^{-1}s^{-1}$)
3QKF**	D-Gluconate, Mg^{2+}	*Cs*ManD/Q1QT89	D-Mannonate (5 $M^{-1}s^{-1}$) D-Gluconate (40 $M^{-1}s^{-1}$)
3RGT	D-Arabinohydroxamate, Co^{2+}	*Cs*ManD/Q1QT89	D-Mannonate (5 $M^{-1}s^{-1}$) D-Gluconate (40 $M^{-1}s^{-1}$)
3T6C	D-Gluconate, Mg^{2+}	D4GJ14	D-Gluconate (120 $M^{-1}s^{-1}$)
3TW9	Glycerol	B5R541	D-Gluconate (80 $M^{-1}s^{-1}$)
3TWA	Glycerol, Mg^{2+}	B5R541	D-Gluconate (80 $M^{-1}s^{-1}$)
3TWB	D-Gluconate, Mg^{2+}	B5R541	D-Gluconate (80 $M^{-1}s^{-1}$)
4IHC	Glycerol, Mg^{2+}	C6CBG9	D-Mannonate (60 $M^{-1}s^{-1}$) D-Gluconate (50 $M^{-1}s^{-1}$)
4E4F	Glycerol, Mg^{2+}	C6D9S0	D-Mannonate (N/A $M^{-1}s^{-1}$)
4IL2	Mg^{2+}	Q8FHC7	D-Mannonate (10 $M^{-1}s^{-1}$)
3DFH*	Apo	A5KUH4	None
3R25	Glycerol, Mg^{2+}	A5KUH4	None
3SBF**	D-Arabinonate, Mg^{2+}	A5KUH4	None
3TJI	Glycerol, Mg^{2+}	A4W7D6	None
3GY1*	Mg^{2+}	A6M2W4	None
3S47	Mg^{2+}	A6M2W4	None
4GIS	Glycerol, Mg^{2+}, maleic acid, malonic acid	D0X4R4	None
4GIR	Mg^{2+}, HEPES, ethylene glycol	D0X4R4	None
4GGH	Mg^{2+}, HEPES, ethylene glycol	D0X4R4	None
4HNL	Glycerol, Mg^{2+}	C8ZZN2	None

*denote structures solved in previous studies
**mutant proteins

APPENDIX: B

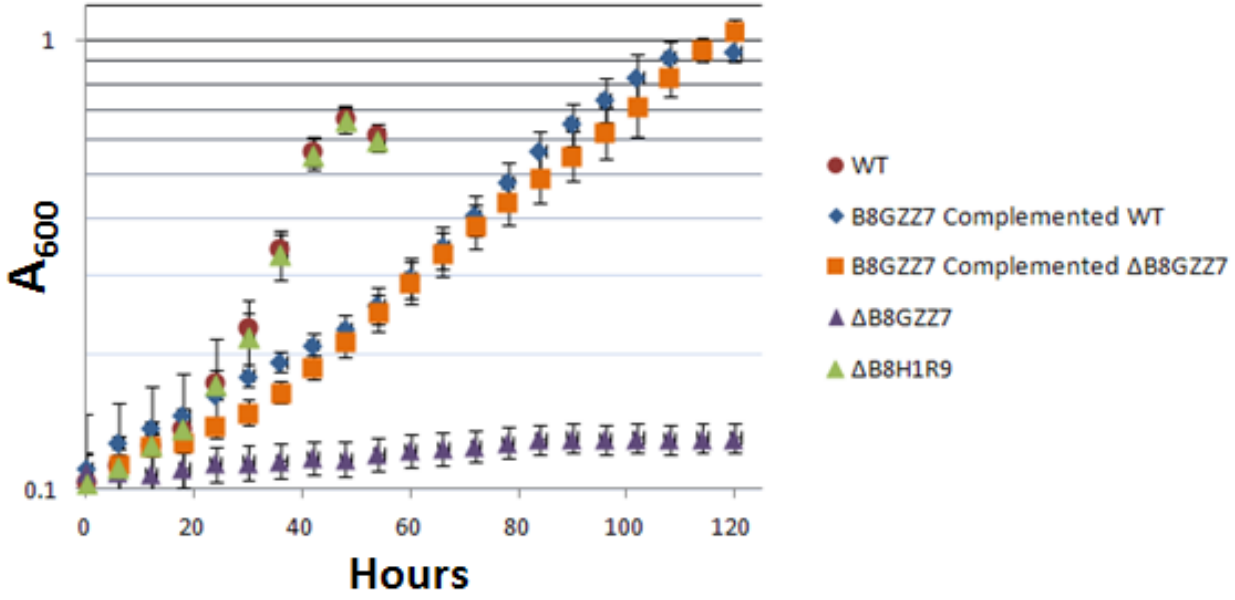


Figure B.1. Growth of *Caulobacter crescentus* NA1000 wild type and the indicated strains grown on M2 minimal salts media with D-glucuronate as the sole source of carbon. WT (purple circles), WT complemented with B8GZZ7 in pSRK-Kan (red squares), ΔB8GZZ7 complemented with B8GZZ7 in pSRK-Kan (blue diamonds), and ΔB8GZZ7 (orange triangles).

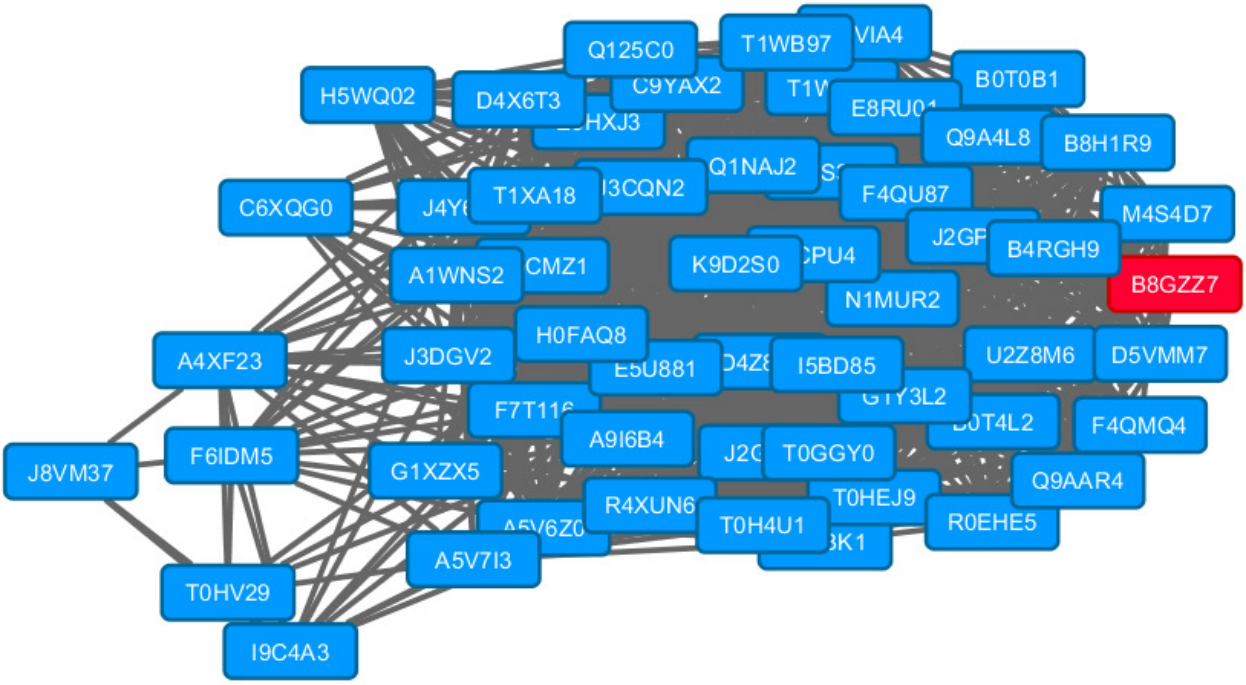


Figure B.2. A SSN of the ManDs that cluster with B8GZZ7 (highlighted in red) at a log(BLAST E_value) of -185 (~70% identity).

Table B.1. Compounds screened via ThermoFluor

D-altrarate/L-talarate	D-threose	D-altrose	trans-ferulate	D-tryptophan	suberate
L-altrarate/D-talarate	L-threose	L-altrose	trans-cinnamate	L-tyrosine	tetradecandiate
L-mannarate	D-erythrose	D-allose	hydrocinnamate	D-tyrosine	N-caproate
D-mannarate	L-erythrose	L-allose	gallate	L-asparagine	malonate
L-idarate	D-arabinose	L-fucose	syringic	D-asparagine	meleate
D-idarate	L-arabinose	D-fucose	sinapate	L-glutamine	D/L-malate
D-glucarate/L-gularate	D-lyxose	D-arabitol	L-alanine	D-glutamine	fumarate
L-glucarate/D-gularate	L-lyxose	D-mannitol	D-alanine	L-histidine	succinate
L-galacturonate	D-xylose	ribitol	L-serine	D-histidine	pyruvate
D-galacturonate	L-xylose	galacitol	D-serine	L-aspartate	2-oxobutyrate
L-glucuronate	D-ribose	D-sorbitol	L-threonine	D-aspartate	2-oxovalerate
D-glucuronate	L-ribose	xylitol	D-threonine	L-glutamate	D/L-lactate
L-mannuronate	D-galactose	myo-inositol	L-valine	D-glutamate	mandelate
D-mannuronate	L-galactose	benzoate	D-valine	L-arginine	fructose/lactose
L-iduronate	D-glucose	3-hydroxybenzoate	L-leucine	D-arginine	maltose/sucrose
D-iduronate	L-glucose	4-hydroxybenzoate	D-leucine	L-lysine	trehalose/D-(+)raffinose
L-guluronate	D-mannose	3,4-dihydroxybenzoate	L-isoleucine	L-ALA-D-GLU	glucosamine/N-AcGlucosamine
D-guluronate	L-mannose	3,4-dihydroxyhydrocinnamate	L-phenylalanine	D-ALA-D-ALA	N-acetylneuraminate/ N-acetylmuramate
L-alluronate	D-idose	phenylacetate	D-phenylalanine	3OH-proline	ectoine/5-hydroxyectoine
D-alluronate	L-idose	3,4-dihydroxylphenylacetate	L-proline	4OH-proline	glycinebetaine/prolinebetaine/ 4OH-prolinebetaine
L-taluronate	D-gulose	benzoylformate	D-proline	2,6-diaminopimelate	
D-taluronate	L-gulose	vanillate	L-methionine	5-aminolevulinate	
L-alturonate	D-talose	caffeate	D-methionine	pyroglutamate	
D-alturonate	L-talose	p-coumarate	L-tryptophan	adipate	

Table B.2. Uniprot IDs of ManDs that cluster with B8GZZ7 at a log(BLAST E-value) of -185

Uniprot ID	Organism
H0FAQ8	Achromobacter arsenitoxydans SY8.
D4X6T3	Achromobacter piechaudii ATCC 43553.
J4Y6C9	Achromobacter piechaudii HLE.
E3HXJ3	Achromobacter xylosoxidans (strain A8).
F7T116	Achromobacter xylosoxidans AXX-A.
E5U881	Achromobacter xylosoxidans C54.
R4XUN6	Achromobacter xylosoxidans NH44784-1996.
F4QU87	Asticcacaulis biprosthecum C19.
F4QMQ4	Asticcacaulis biprosthecum C19.
E8RU01	Asticcacaulis excentricus (strain ATCC 15261 / DSM 4724 / VKM B-1370 /CB 48).
G1XZX5	Azospirillum amazonense Y2.
G1Y3L2	Azospirillum amazonense Y2.
A9I6B4	Bordetella petrii (strain ATCC BAA-461 / DSM 12804 / CCUG 43448).
U2Z8M6	Brevundimonas abyssalis TAR-001.
H5WQ02	Burkholderiales bacterium JOSHI_001.
B8GZZ7	Caulobacter crescentus (strain ATCC 19089 / CB15).
Q9A4L8	Caulobacter crescentus (strain ATCC 19089 / CB15).
B8GZZ7	Caulobacter crescentus (strain NA1000 / CB15N).
B8H1R9	Caulobacter crescentus (strain NA1000 / CB15N).
R0EHE5	Caulobacter crescentus OR37.
D5VMM7	Caulobacter segnis (strain ATCC 21756 / DSM 7131 / JCM 7823 / NBRC15250 / LMG 17158 / TK0059)
D5VIA4	Caulobacter segnis (strain ATCC 21756 / DSM 7131 / JCM 7823 / NBRC15250 / LMG 17158 / TK0059)
B0T4L2	Caulobacter sp. (strain K31).
B0T0B1	Caulobacter sp. (strain K31).
J2GJ67	Caulobacter sp. AP07.
J2GPJ8	Caulobacter sp. AP07.
C9YAX2	Curvibacter putative symbiont of Hydra magnipapillata.
C6XQG0	Hirschia baltica (strain ATCC 49814 / DSM 5838 / IFAM 1418).
A4XF23	Novosphingobium aromaticivorans (strain DSM 12444).
T0H4U1	Novosphingobium lindaniclasticum LE124.
F6IDM5	Novosphingobium sp. PP1Y.
I9C4A3	Novosphingobium sp. Rr 2-17.
B4RGH9	Phenylobacterium zucineum (strain HLK1).
Q125C0	Polaromonas sp. (strain JS666 / ATCC BAA-500).
J3DGV2	Polaromonas sp. CF318.
I5BD85	Sphingobium indicum B90A.
D4Z888	Sphingobium japonicum (strain NBRC 101211 / UT26S).
N1MUR2	Sphingobium japonicum BiD32.
T0HEJ9	Sphingobium lactosutens DS20.
T0HV29	Sphingobium lactosutens DS20.
J2CPU4	Sphingobium sp. AP49.
T0GGY0	Sphingobium sp. HDIP04.
T0JBK1	Sphingobium ummariense RL-3.
K9D2S0	Sphingobium yanoikuyae ATCC 51230.
J8VM37	Sphingomonas sp. LH128.
M4S302	Sphingomonas sp. MM-1.
M4S4D7	Sphingomonas sp. MM-1.
Q1NAJ2	Sphingomonas sp. SKA58.
A5V6Z0	Sphingomonas wittichii (strain RW1 / DSM 6014 / JCM 10273).
A5V7I3	Sphingomonas wittichii (strain RW1 / DSM 6014 / JCM 10273).
C5CMZ1	Variovorax paradoxus (strain S110).
T1XA18	Variovorax paradoxus B4.
J3CQN2	Variovorax sp. CF313.
A1WNS2	Verminephrobacter eiseniae (strain EF01-2).
T1WB97	uncultured organism.
T1WB73	uncultured organism.

APPENDIX: C

L-glycerate	L-xylonate	D-glucarate/L-gularate	L-galactonate	L-alturonate
D-glycerate	D-xylonate	L-glucarate/D-gularate	D-galactonate	D-alturonate
D-glycerate 3-P	L-ribonate	L-talonate	L-fuconate	L-alluronate
meso-tartrate	D-ribonate	D-talonate	D-fuconate	D-alluronate
L-tartrate	L-lyxonate	L-mannonate	6-deoxy-L-talonate	L-taluronate
D-tartrate	D-lyxonate	D-mannonate	L-rhamnonate	D-taluronate
L-malate	L-arabinonate	L-idonate	L-galacturonate	L-glucuronate
D-malate	D-arabinonate	D-idonate	D-galacturonate	D-glucuronate
L-threonate	D/L allarate	L-gluconate	L-galactonate 6-P	
D-threonate	D/L galactarate	D-gluconate	D-gluconate 6-P	
D-erythronate	L-altrarate/D-talarate	L-allonate	L-mannuronate	
L-erythronate	D-altrarate/L-talarate	D-allonate	D-mannuronate	
D/L xylarate	L-mannarate	L-gulonate	L-iduronate	
D/L ribarate	D-mannarate	D-gulonate	D-iduronate	
L-arabinarate/D-lyxarate	L-idarate	L-altronate	L-guluronate	
D-arabinarate/L-lyxarate	D-idarate	D-altronate	D-guluronate	

Figure C.1. The 72 acid sugars in the acid sugar library used for screening *Cs*GulDH and *Cs*FR homologues for oxidation.

D-altrarate/L-talarate	D-threose	D-altrose	trans-ferulate	D-tryptophan	suberate
L-altrarate/D-talarate	L-threose	L-altrose	trans-cinnamate	L-tyrosine	tetradecandiate
L-mannarate	D-erythrose	D-allose	hydrocinnamate	D-tyrosine	N-caproate
D-mannarate	L-erythrose	L-allose	gallate	L-asparagine	malonate
L-idarate	D-arabinose	L-fucose	syringic	D-asparagine	meleate
D-idarate	L-arabinose	D-fucose	sinapate	L-glutamine	D/L-malate
D-glucarate/L-gularate	D-lyxose	D-arabitol	L-alanine	D-glutamine	fumarate
L-glucarate/D-gularate	L-lyxose	D-mannitol	D-alanine	L-histidine	succinate
L-galacturonate	D-xylose	ribitol	L-serine	D-histidine	pyruvate
D-galacturonate	L-xylose	galacitol	D-serine	L-aspartate	2-oxobutyrate
L-glucuronate	D-ribose	D-sorbitol	L-threonine	D-aspartate	2-oxovalerate
D-glucuronate	L-ribose	xylitol	D-threonine	L-glutamate	D/L-lactate
L-mannuronate	D-galactose	myo-inositol	L-valine	D-glutamate	mandelate
D-mannuronate	L-galactose	benzoate	D-valine	L-arginine	fructose/lactose
L-iduronate	D-glucose	3-hydroxybenzoate	L-leucine	D-arginine	maltose/sucrose
D-iduronate	L-glucose	4-hydroxybenzoate	D-leucine	L-lysine	trehalose/D-(+)raffinose
L-guluronate	D-mannose	3,4-dihydroxybenzoate	L-isoleucine	L-ALA-D-GLU	glucosamine/N-AcGlucosamine
D-guluronate	L-mannose	3,4-dihydroxyhydrocinnamate	L-phenylalanine	D-ALA-D-ALA	N-acetylneuraminate/ N-acetylmuramate
L-alluronate	D-idose	phenylacetate	D-phenylalanine	3OH-proline	ectoine/5-hydroxyectoine
D-alluronate	L-idose	3,4-dihydroxylphenylacetate	L-proline	4OH-proline	glycinebetaine/prolinebetaine/ 4OH-prolinebetaine
L-taluronate	D-gulose	benzoylformate	D-proline	2,6-diaminopimelate	
D-taluronate	L-gulose	vanillate	L-methionine	5-aminolevulinate	
L-alturonate	D-talose	caffeate	D-methionine	pyroglutamate	
D-alturonate	L-talose	p-coumarate	L-tryptophan	adipate	

Figure C.2. Compounds screened via ThermoFluor

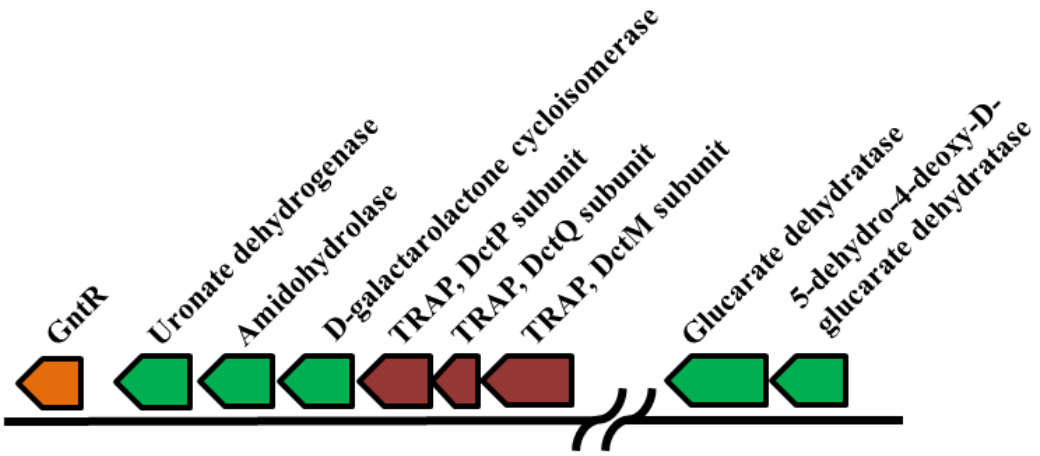


Figure C.3. Genome neighborhood of the genes encoding the D-glucuronate pathway in *Chromohalobacter salexigens* DSM3043. Enzymes are highlighted in green: uronate dehydrogenase (Uniprot ID Q1QUN7), D-galactarolactone cycloisomerase (Uniprot ID Q1QUN5; Gci), amidohydrolase (Uniprot ID Q1QUN6), glucarate dehydratase (Uniprot ID Q1QUM5: GlucD), and 5-dehydro-4-deoxy-D-glucarate dehydratase (Uniprot ID Q1QUM4). The TRAP transport system is highlight in red. A GntR regulator is highlighted in orange.

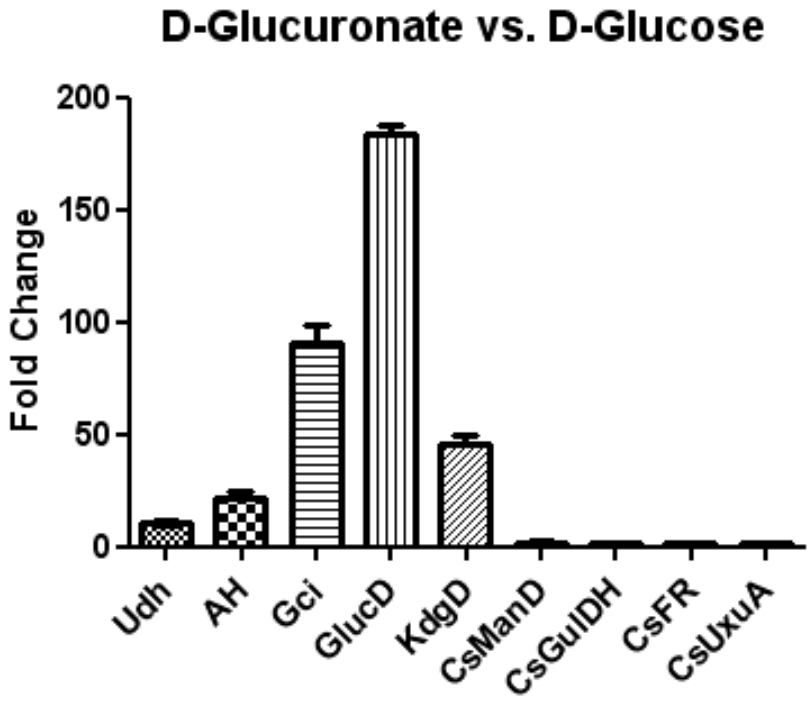


Figure C.4. qRT-PCR data for D-glucuronate metabolism in *C. salexigens* DSM3043 (D-glucuronate vs. D-glucose).

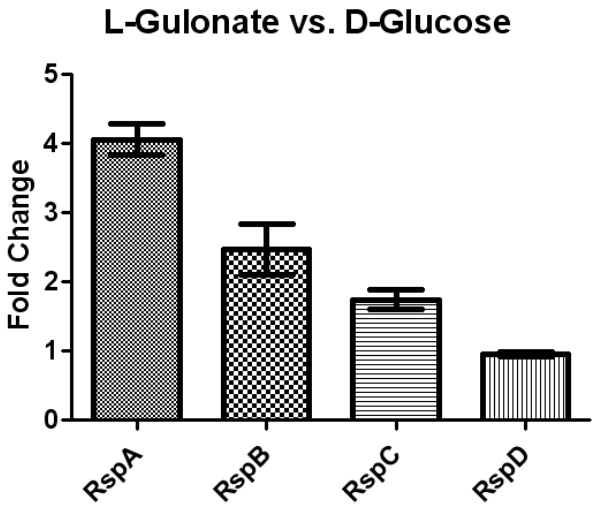


Figure C.5. qRT-PCR data of the *RspABCD* genes in *Salmonella enterica serovar Enteritidis* str. P125109, comparing growth on L-gulonate to D-glucose. Cells were grown as described in the Material and Methods section to an optical density of 0.4 - 0.5 at 600 nm (early log phase).

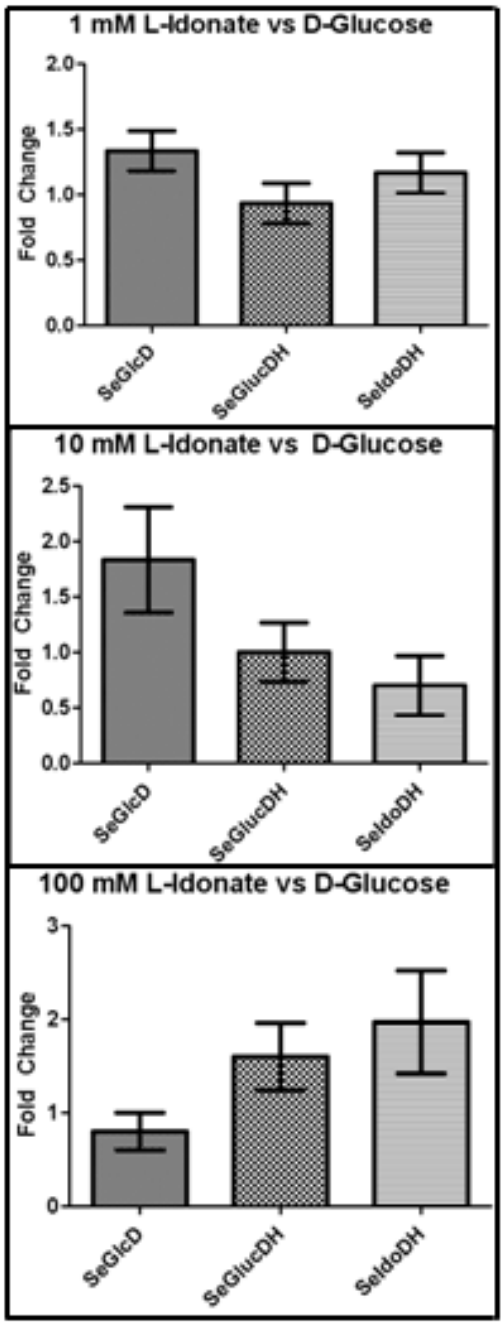


Figure C.6. qRT-PCR data of the *SeGlcD* genome neighborhood in *Salmonella enterica serovar Enteritidis* str. P125109, comparing growth on L-Idonate to D-glucose at varying concentrations. Cells were grown as described in the Material and Methods section to an optical density of 0.4 - 0.5 at 600 nm (early log phase).

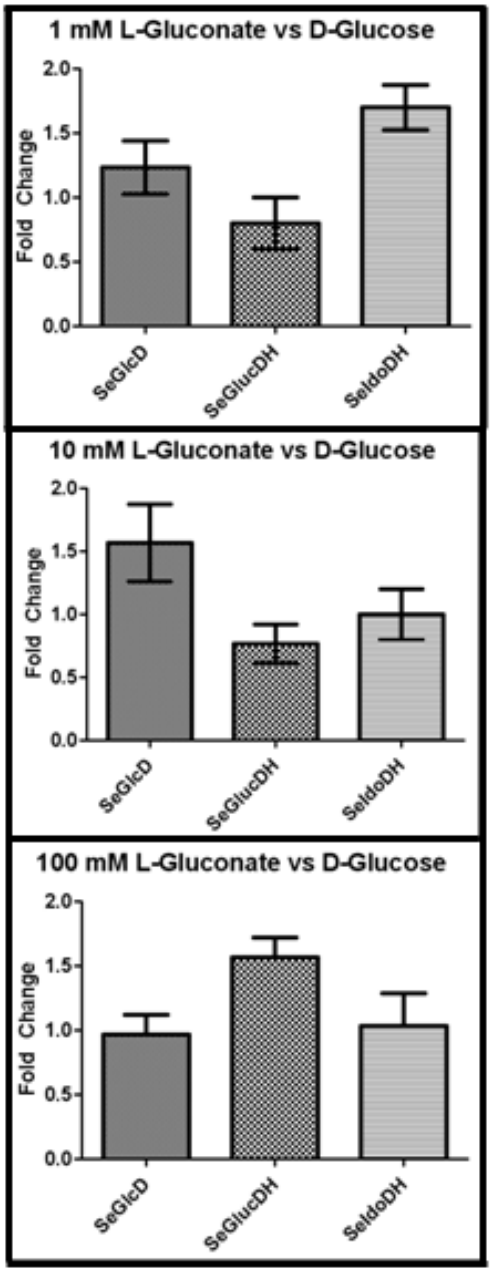


Figure C.7. qRT-PCR data of the *SeGlcD* genome neighborhood in *Salmonella enterica serovar Enteritidis* str. P125109, comparing growth on D-gluconate to D-glucose at varying concentrations. Cells were grown as described in the Material and Methods section to an optical density of 0.4 - 0.5 at 600 nm (early log phase).

Table C.1. PCR primers used in this study

Purpose	Primer name	Sequence
Cloning	RspB_FOR	5’-AATGGTTATCTCTAAAAAAGAGGTAGGATCCGTACTACTTACTCG -3’
Cloning	RspB_REV	5’- CGGCGAAATATCCATATGAACCTGCTTATCTGCCAGT -3’
Cloning	RspD_FOR	5’- GATCAACCACATGTTGTAAGAGCTCCGCAACCGTCC -3’
Cloning	RspD_REV	5’- GTCTGACCGCTGAGCGGATCCGCCGC -3’
Cloning	CsGntR_FOR	5’- GCAGGGTTTTCATATGACGCCATCGC -3’
Cloning	CsGntR_REV	5’- GTCGCGAATCCTCATGCACTGACTC -3’
RT-PCR	CsGntR_RTPCR_FOR	5’- GCGAAGCCTTCATTCGTCTC -3’
RT-PCR	CsGntR_RTPCR_REV	5’- GAGTTCATGCAGCTCGTCGT -3’
RT-PCR	CsManD_RTPCR_FOR	5’- CGACTATCGCTTCGAGGATG -3’
RT-PCR	CsManD_RTPCR_REV	5’- TACTTGGCAGCGAGTTCCTC -3’
RT-PCR	CsAD_RTPCR_FOR	5'- ACCGTCTGCGACGTCATC -3'
RT-PCR	CsAD_RTPCR_REV	5’- AGGTCCTGTCGTTGTCCATT-3’
RT-PCR	CsDctP_RTPCR_FOR	5'- ATGGCCTTCAGCGAGGTATTC -3'
RT-PCR	CsDctP_RTPCR_REV	5’- GGCATCGATTCGAGCTTCTG -3’
RT-PCR	CsDctQ_RTPCR_FOR	5'- CTGATCTGTCTCGCGGTAGG -3'
RT-PCR	CsDctQ_RTPCR_REV	5'- GCGGTCAAGCCAAAGAGAAC -3'
RT-PCR	CsDctM_RTPCR_FOR	5’- CATCGACACCGTCCTGACC -3’
RT-PCR	CsDctM_RTPCR_REV	5’- TCCATGAAGGTACCCACCAC -3’
RT-PCR	CsFR_RTPCR_FOR	5’- GAGCGAACCCGATGTACG -3’
RT-PCR	CsFR_RTPCR_REV	5’- GCATCGAGCCGTAGTCGT -3’
RT-PCR	CsUxuA_RTPCR_FOR	5'- TACCATCATCGCCGGACT -3'
RT-PCR	CsUxuA_RTPCR_REV	5’- GCATCGATCTCGGCATACTC-3’
RT-PCR	CsRpoD_RTPCR_FOR	5'- GTCTCGCACATCAAGGAAGTC -3'
RT-PCR	CsRpoD_RTPCR_REV	5’- AGGTTGGCCTCGACCATT-3’
RT-PCR	CsUdh_RTPCR_FOR	5'- GTTCGCCAGTTCCAATCACAC -3'
RT-PCR	CsUdh_RTPCR_REV	5'- GAACCGATGCGTATGTTGAGG -3'
RT-PCR	CsGci_RTPCR_FOR	5’- GCGGCAATCACTTTCACGTC -3’
RT-PCR	CsGci_RTPCR_REV	5’- ATATCAAGGGCAAGCATTACGG -3’
RT-PCR	CsAH_RTPCR_FOR	5’- GGGACTGGCATGACAAAGG -3’
RT-PCR	CsAH_RTPCR_REV	5’- GATCCGCGAGGTAATCGTTC -3’
RT-PCR	CsGlucD_RTPCR_FOR	5'- CCTTCGAGGTTGCCTTTCTC -3'
RT-PCR	CsGlucD_RTPCR_REV	5’- ATCGACTGGAAGCCGTATTGG -3’
RT-PCR	CsKdgD_RTPCR_FOR	5'- GTATTGACGCCAATGGAGAGG -3'
RT-PCR	CsKdgD_RTPCR_REV	5’- CAACCGTGGCAATGGTGTTCT -3’
RT-PCR	SeRspA_RTPCR_FOR	5’- TACCCATGCCGGTGGTATTAC -3’
RT-PCR	SeRspA_RTPCR_REV	5’-CCCTCGAACGTCCAGTTATG-3’
RT-PCR	SeRspB_RTPCR_FOR	5’- ATCGCATACCGGACAACATCG-3’
RT-PCR	SeRspB_RTPCR_REV	5’- GCCATCGCTAAACGCTCTTCA -3’
RT-PCR	SeRspC_RTPCR_FOR	5’- CGTTAACGCCTGCGTCAGAAA -3’
RT-PCR	SeRspC_RTPCR_REV	5’- ATTAACGCATCGGTTGGGATAG-3’
RT-PCR	SeRspD_RTPCR_FOR	5’- TCTATACCGTGGCGGAGATG -3’
RT-PCR	SeRspD_RTPCR_REV	5’- CATGGGATGTTCCAGCAGAAG-3’
RT-PCR	SeGlcD_RTPCR_FOR	5’- TTTCTGGTTGGCAAAGATCC-3’
RT-PCR	SeGlcD_RTPCR_REV	5’- CAACTTCATCACCACCATCG-3’
RT-PCR	SeIdoDH_RTPCR_FOR	5’- GCAATGCATTCCTTACGACA-3’
RT-PCR	SeIdoDH_RTPCR_REV	5’- AATGGGCAACCTGTTCTTCA-3’
RT-PCR	SeGlucDH_RTPCR_FOR	5’- GTGCACGGGTTATCCTGAAT-3’
RT-PCR	SeGlucDH_RTPCR_REV	5’- AATGCGCTGGTCAGGTTAGT-3’
Knockout	SeRspA_KO_FOR	5’-TTTCCTGCATCCCATAATGCGATGTAAGGAATAAAAATGTGTAGGCTGGAGCTGCTTCG -3’
Knockout	SeRspA_KO_REV	5’-GTTTTTCGATTACGATGCTTTTCATTTTTCGCTCCTTACATATGAATATCCTCCTTAG -3’
Knockout	SeRspB_KO_FOR	5’-

		AAGACGGCACACTGTGGAACTGGTAAGGAGCGAAAAATG TGTAGGCTGGAGCTGCTTCG -3’
Knockout	SeRspB_KO_REV	5’- GGTAAGATGCGTACCACTTACTCGCAATTATTTATATTATT TCGCCATATGAATATCCTCCTTAG-3’
Knockout	KO_CsAD_EcoRI_OE_FOR	5'- CGACCTGCACGTCCTGCACGAATTCCACCATCGCCTGACG CC -3'
Knockout	KO_CsAD_EcoRI_OExt_REV	5’- GGCGCATCGCGCGATGCGGACGGACTCCGCTCTGG -3’
Knockout	KO_CsAD_HindIII_OExt_FOR	5'- CCAGAGCGGAGTCCGTCCGCATCGCGCGATGCGCC -3'
Knockout	KO_CsAD_HindIII_OE_REV	5’- CGCCTGCCAGGTATCGACGAAAAGCTTCACGTTGGGTACG CGCAGC -3’
Knockout	KO_CsFR_EcoRI_OE_FOR	5'- GACCTTCATCATTCTCTACGGAATTCTCGGCGGCGTTTTCA CCCCCACCG -3'
Knockout	KO_CsFR_EcoRI_OExt_REV	5'- CCAGGTGTGTTCCATGGTCGTCTCGAACGATCACTCCCCGA GCAGC -3'
Knockout	KO_CsFR_HindIII_OExt_FOR	5’- GCTGCTCGGGGAGTGATCGTTCGAGACGACCATGGAACAC ACCTGG -3’
Knockout	KO_CsFR_HindIII_OE_REV	5’- GTGGCCGAGGTTGTCGCGCAAGCTTTCGGCATCGATCTCG GCATAC -3’
Knockout	KO_CsFRUxuA_EcoRI_OE_FOR	5’- GATGACCTTCATCATTCTCTACGGAATTCTCGGCGGCGTTT TCACCCCCACCG -3’
Knockout	KO_CsFRUxuA_EcoRI_OExt_REV	5’- CCTGGGAGCGGGTCGTCTCGAACGATCACTCCCCGAGCAG -3’
Knockout	KO_CsFRUxuA_HindIII_OExt_FOR	5'- CTGCTCGGGGAGTGATCGTTCGAGACGACCCGCTCCCAGG -3'
Knockout	KO_CsFRUxuA_HindIII_OE_REV	5’- GACCGGTGCCTCCGCAAGCTTGGCGAGCAGCGCCTCC -3’
Knockout	KO_CsUxuA_EcoRI_OE_FOR	5'- CGGCATCGAAACGGTGGCAGAAGCCGTCGAGCGCCCTGAA TTCG -3'
Knockout	KO_CsUxuA_EcoRI_OExt_REV	5’- CGTCCCCCTCCTGGGAGCGGGTCGTCTCCTATCGTGTGC -3’
Knockout	KO_CsUxuA_HindIII_OExt_FOR	5'- GCACACGATAGGAGACGACCCGCTCCCAGGAGGGGGACG -3'
Knockout	KO_CsUxuA_HindIII_OE_REV	5’- GACCGGTGCCTCCGCAAGCTTGGCGAGCAGCGCCTCC -3’
Knockout	KO_CsManD_BamHI_OE_FOR	5'- CGCCAGCCGGTCCGCGGATCCTTCATTCGTCTCTCGGAGG -3'
Knockout	KO_CsManD_BamHI_OExt_REV	5’- GCTCTGGATAAGGGAAGACGGGACTTCATGCACTGACTCC AGAGGGCAGC -3’
Knockout	KO_CsManD_HindIII_OExt_FOR	5'- GCTGCCCTCTGGAGTCAGTGCATGAAGTCCCGTCTTCCCTT ATCCAGAGC -3'
Knockout	KO_CsManD_HindIII_OE_REV	5’- GTCCATTGACGGTGAAGCTTGCGCCCAGCGAGCGGGCGG -3’

Table C.2. *Cs*GntR ThermoFluor hits

Substrate	ΔT_m
Mandelate	9.4
D-Mannonate	7.1
D/L-Lactate	6.2
L-Arabinarate/D-Lyxarate	4.9
D-Lyxonate	3.6
D-Tryptophan	2.7
Pyruvate	2.5
D-Arabinarate/L-Lyxarate	2.4
Gallate	2.2
L-Gulonate	2.1

APPENDIX: D

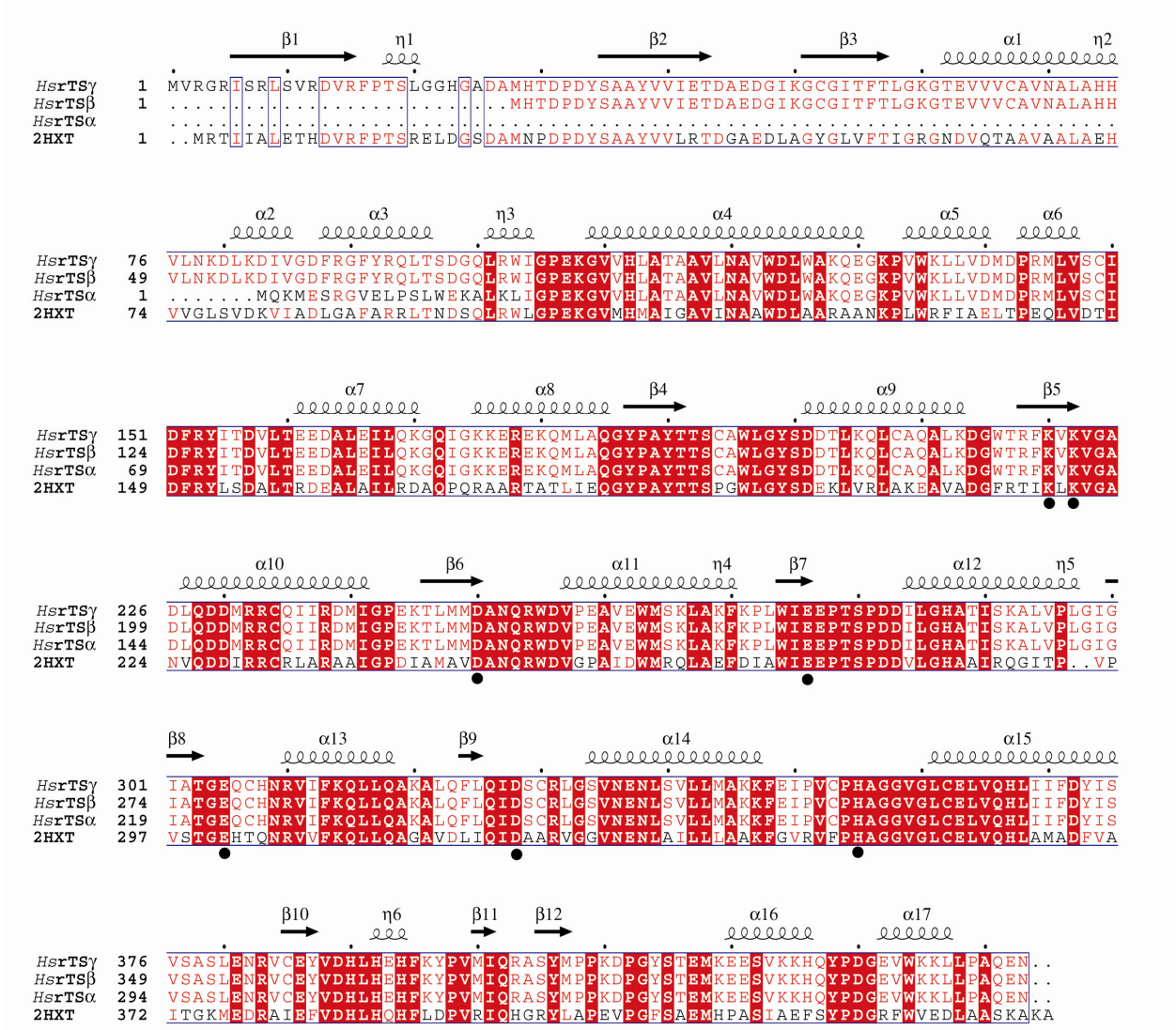


Figure D.1. Sequence alignment of human rTS isoforms and the structural homologue *Xc*FucD. Secondary structure elements from the *Hs*rTSγ structure are shown on top of the aligned sequences of *Hs*rTSα/β/γ isoforms and *Xc*FucD (PDB code 2HXT). Putative catalytic residues inferred from the *Xc*FucD structure are indicated as black dots.

L-glycerate	L-xylonate	D-glucarate/L-gularate	L-galactonate	L-alturonate
D-glycerate	D-xylonate	L-glucarate/D-gularate	D-galactonate	D-alturonate
D-glycerate 3-P	L-ribonate	L-talonate	L-fuconate	L-alluronate
meso-tartrate	D-ribonate	D-talonate	D-fuconate	D-alluronate
L-tartrate	L-lyxonate	L-mannonate	6-deoxy-L-talonate	L-taluronate
D-tartrate	D-lyxonate	D-mannonate	L-rhamnonate	D-taluronate
L-malate	L-arabinonate	L-idonate	L-galacturonate	L-glucuronate
D-malate	D-arabinonate	D-idonate	D-galacturonate	D-glucuronate
L-threonate	D/L allarate	L-gluconate	L-galactonate 6-P	
D-threonate	D/L galactarate	D-gluconate	D-gluconate 6-P	
D-erythronate	L-altrarate/D-talarate	L-allonate	L-mannuronate	
L-erythronate	D-altrarate/L-talarate	D-allonate	D-mannuronate	
D/L xylarate	L-mannarate	L-gulonate	L-iduronate	
D/L ribarate	D-mannarate	D-gulonate	D-iduronate	
L-arabinarate/D-lyxarate	L-idarate	L-altronate	L-guluronate	
D-arabinarate/L-lyxarate	D-idarate	D-altronate	D-guluronate	

Figure D.2. The list of 72 acid sugars in the acid sugar library, which were used for detection of dehydration activity.

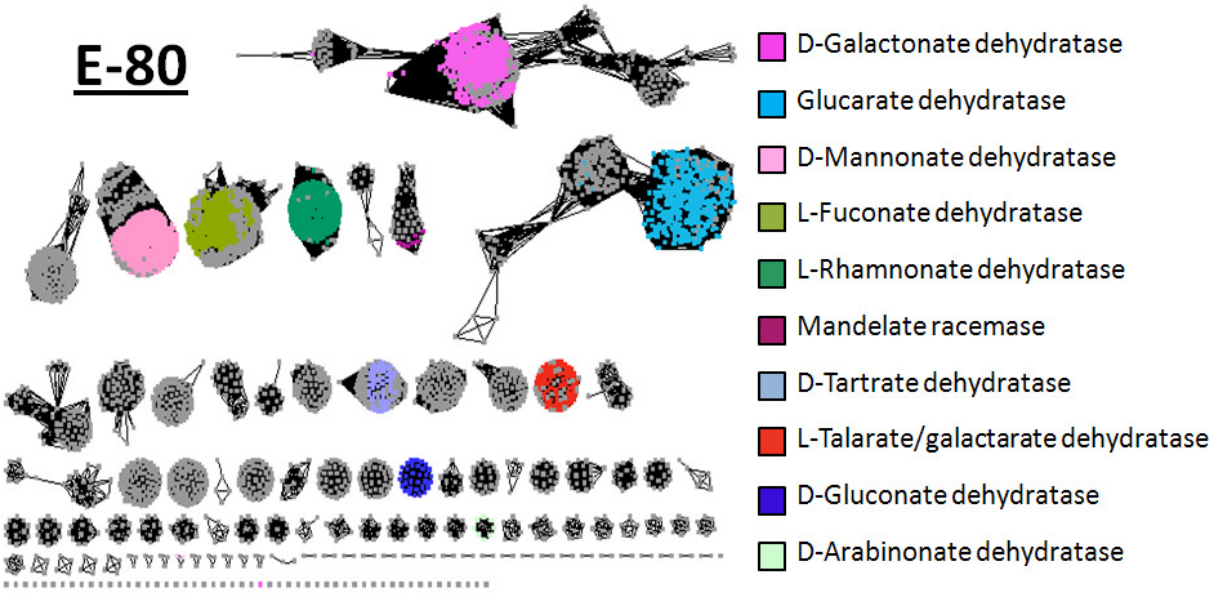


Figure D.3. A sequence similarity network for all acid sugar dehydratases and mandelate racemases in the enolase superfamily at an e-value threshold of 10^{-80} (~40% identity). rTSγ clusters with the L-fuconate dehydratases (olive green). All other nodes with known activities are colored and labeled in the legend. Grey denotes unknown activity.

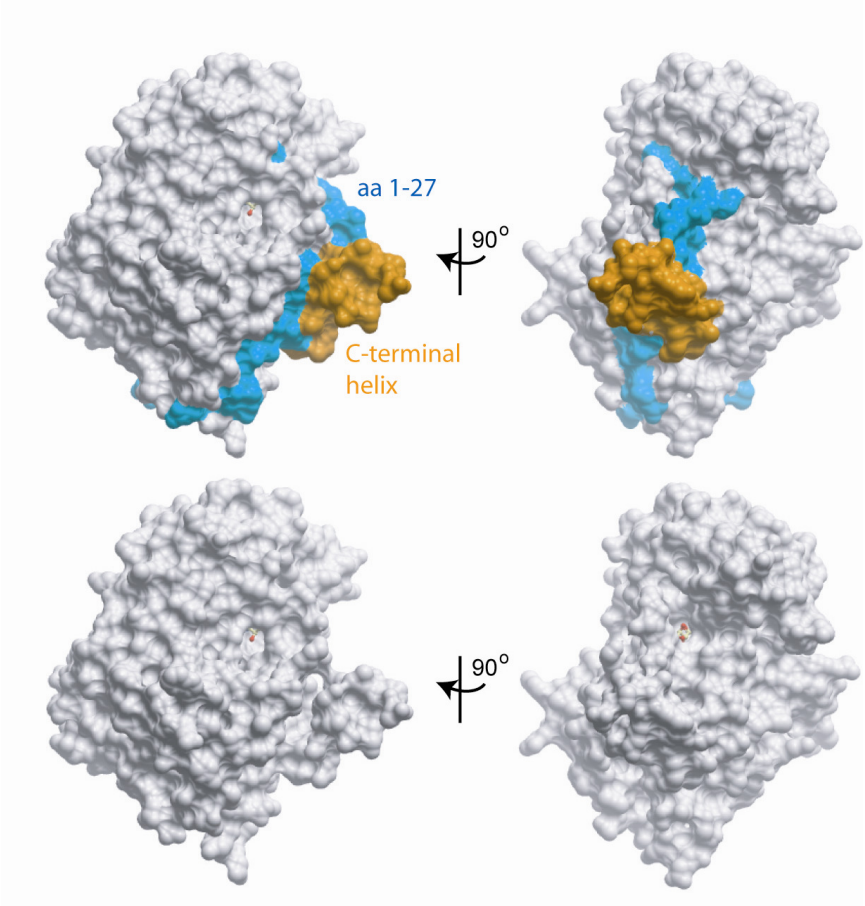


Figure D.4. Surface representation of *Hs*rTSγ structure in orthogonal views. Top, structure of *Hs*rTSγ highlighting the N-terminal 27 residues (cyan) and the juxtaposed C-terminal helix (orange). Bottom, structure of *Hs*rTSγ with the N-terminal 27 residues deleted from the model. In all views, the active site is indicated by the modeled ligand L-erythronohydroxamate (sticks) from the homologous *Xc*FucD structure (PDB 2HXT).

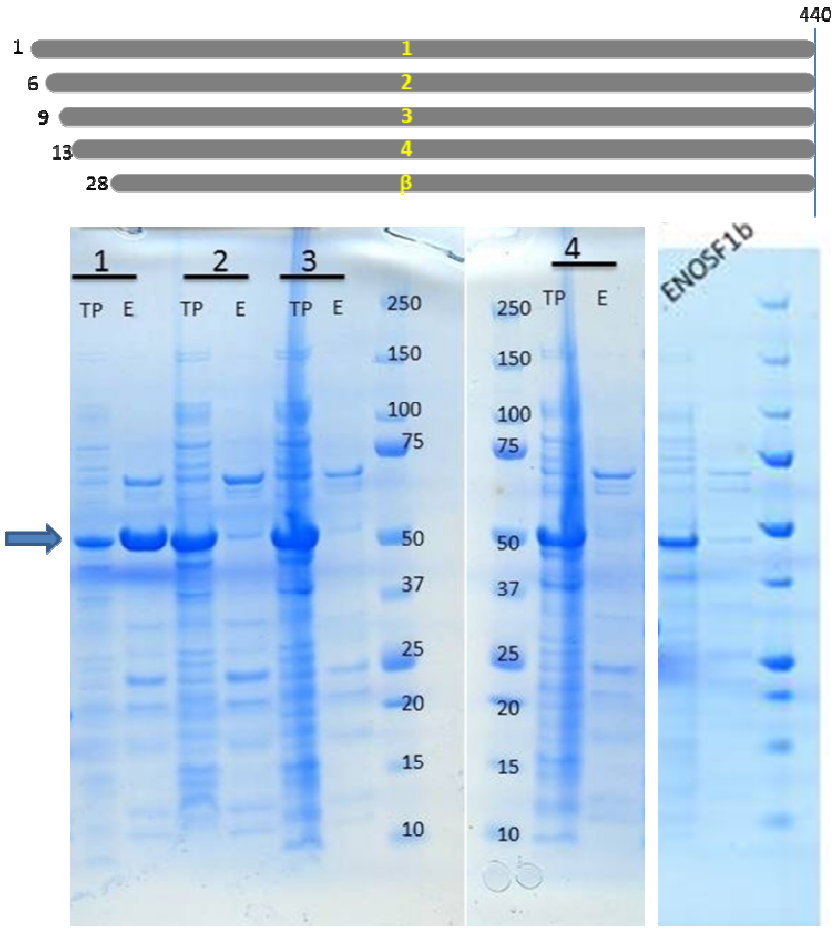


Figure D.5. Small scale expression and affinity purification of various *Hs*rTSγ constructs as well as *Hs*rTSβ. TP, total cell pellet; E, elution from nickel column purification. Arrow indicates position of rTS protein in SDS-PAGE.

Table D.1. Melting temperatures (T_m) of *Hs*rTSβ and *Hs*rTSγ as measured by differential scanning fluorimetry (n=4)

T_m (°C)	*Hs*rTSβ	*Hs*rTSγ
as purified	52.77 ± 0.1	57.17 ± 0.3
+$MgCl_2$	53.15 ± 0.1	57.51 ± 0.4
+$MgCl_2$, D-ertythronohydroxamate	53.38 ± 0.1	59.48 ± 0.2

REFERENCES

1. Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W., *Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein.* Nature, 1972. **237**: p. 82-8.
2. Fleischmann, R.D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995. **269**: p. 496-512.
3. Noble, I., *Human genome finally complete.* BBC News Online, 2003.
4. Consortium, I.H.G.S., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**: p. 931-45.
5. Bork, P., *Go hunting in sequence databases but watch out for the traps.* Genetwork, 1996. **12**: p. 425-27.
6. Karp, P.D., *What we do not know about sequence analysis and sequence databases.* Bioinformatics, 1998. **14**: p. 753-754.
7. Gilks, W.R., Audit, B., De Angelis, D., Tsoka, S., and Ouzounis, C.A., *Modeling the percolation of annotation errors in a database of protein sequences.* Bioinformatics, 2002. **18**(12): p. 1641-1649.
8. Devos, D.a.V., A., *Intrinsic errors in genome annotation.* TRENDS in Genetics, 2001. **117**(8): p. 429-431.
9. Jones, C.E., Brown, A.L., and Baumann U., *Estimating the annotation error rate of the curated GO database sequence annotations.* BMC Bioinformatics, 2007. **8**(170).

10. Gerlt, J.A., Allen, K.N., Almo, S.C., Armstrong, R.N., Babbitt, P.C., Cronan, J.E., Dunaway-Mariano, D., Imker, H.J., Jacobson, M.P., Minor, W., Poulter, C.D., Raushel, F.M., Sali, A., Shoichet, B.K., and Sweedler, J.V., *The Enzyme Function Initiative.* Biochemistry, 2011. **50**: p. 9950-9962.

11. Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D., *The Pfam protein families database.* Nucleic Acids Research, 2012(Database Issue 40): p. D290-D301.

12. Apweiler, R., Attwodd, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., DUrbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J., and Zdobnov, E. M., *The InterPro database, an integrated documentation resource for protein families, domains and functional sites.* Nucleic Acids Research, 2001. **29**: p. 37-40.

13. Atkinson, H.J., Morris J. H., Ferrin T. E., and Babbitt P. C., *Using sequence similarity networks for visualization of relationships across diverse protein superfamilies.* PLoS One, 2009. **4**: p. e4345.

14. Pantoliano, M.W., Petrella, E. C., Kwasnoski, J. D., Lobanov, V. S., Myslik, J., Graf, E., Carver, T., Asel, E., Springer, B. A., Lane, P., and Salemme, F. R., *High-density miniaturized thermal shift assays as a general strategy for drug discovery.* J. Biomol. Screen, 2001. **6**: p. 429-440.

15. Hermann, J.C., Marti-Arbona, R. Fedorov, A. A., Fedorov, E. V., Almo, S. C., Shoichet, B. K., Raushel, F. M., *Structure-based activity prediction for an enzyme of unknown function.* Nature 2007. **448**: p. 775-779.

16. Irwin, J.J., Schoichet, B. K., Mysinger, M. M., Huang, N., Colizzi, F., Wassam, P. and Cao, Y., *Automated dockign screens: a feasibility study.* J. Med. Chem., 2009. **52**: p. 5712-5720.

17. Lukk T., S.A., Kalyanaraman C., Brown S. D., Imker H. J., Song L., Fedorov A. A., Fedorov E. V., Toro R., Hillerich B., Seidel R., Patskovsky Y., Vetting M. W., Nair S. K., Babbitt P. C., Almo S. C., Gerlt J. A., Jacobson M. P., *Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily.* Proc. Natl. Acad. Sci., 2012. **109**: p. 4122-4127.

18. Freeman, W.M., Walker, S. J., and Vrana, K. E., *Quantitative RT-PCR: pitfalls and potential.* BioTechniques, 1999. **26**: p. 112-122, 124-125.

19. Morin, R.D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J. M., and Marra, M. A., *Profiling the HeLa S3 transciptome using randomly primed cDNA and massively parallel short-read sequencing.* BioTechniques 2008. **45**: p. 81-94.

20. Nicholson, J.K., Lindon, J. C., and Holmes, E., *'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.* Xenobiotica, 1999. **29**: p. 1181-1189.

21. Nicholson, J.K., Connelly, J., Lindon, J. C., and Holmes, E., *Metabonomics: a platform for studying drug toxicity and gene function.* Nat Rev Drug Discov, 2002. **1**: p. 153-161.

22. Robertson, D.G., *Metabolomics in toxicology: a review.* Toxicol Sci, 2005. **85**: p. 809-822.

23. Babbitt, P.C., Mrachko, G. T., Hasson, M. S., Huisman, G. W., Kolter, R., Ringe, D., Petsko, G. A., Kenyon, G. L., and Gerlt, J. A., *A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids.* Science, 1995. **267**: p. 1159 - 61.

24. Babbitt, P.C., Hasson, M.S., Wedekind, J.E., Palmer, D.R.J., Barret, W.C., Reed, G.H., Rayment, I., Ringe, D., Kenyon, G.L., and Gerlt, J.A., *The Enolase Superfamily: A General Strategy for Enzyme-Catalyzed Abstraction of the alpha-Protons of Carboxylic Acids.* Biochemistry, 1996. **35**: p. 16489-16501.

25. Gerlt, J.A., and Babbitt, P. C., *Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu. Rev. Biochem.*, 2001. **70**: p. 209-246.

26. Gerlt, J.A., Babbit P. C., and Rayment, I., *Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. Arch. Biochem. Biophys.*, 2005. **433**: p. 59-70.

27. Babbitt, P.C., and Gerlt, J. A., *Understanding enzyme superfamilies. Chemistry as the fundamental deeterminant in the evolution of new catalytic activities. J. Biol. Chem.*, 1997. **272**: p. 30591-4.

28. Rakus, J.F., Fedorov, A.A., Fedorov, E.V., Glasner, M.E., Vick, J.E., Babbitt, P.C., Almo, S.C., and Gerlt, J.A., *Evolution of enzymatic activities in the enolase superfamily: D-mannonate dehydratase from Novosphingobium aromaticivorans.* Biochemistry, 2007. **46**(45): p. 12896-12908.

29. Ashwel, G., *Enzymes of glucuronic and galacturonic acid metabolism in bacteria.* Methods Enzymol., 1962. **5**: p. 190-208.

30. Yew, W.S., Fedorov, A. A., Fedorov, E. V., Rakus, J. F., Pierce, R. W., Almo, S. C., and Gerlt, J. A., *Evolution of enzymatic activities in the enolase superfamily: L-fuconate dehydratase from Xanthomonas campestris.* Biochemistry, 2006. **45**: p. 14582-97.

31. Schnoes, A.M., Brown, S.D., Dodevski, I., and Babbitt, P.C., *Annotation error in public databases: misannotation of molecular function in enzyme superfamilies.* PLoS Comput Biol, 2009. **5**(12).

32. Tian, W.a.S., J., *How well is enzyme function conserved as a function of pairwise sequence identity?* J. Mol. Biol., 2003. **333**: p. 863-882.

33. Studier, F.W., *Protein production by suto-induction in high density shaking cultures.* Protein Expr. Purif., 2005. **41**: p. 207-34.

34. Gulick, A.M., Hubbard, B.K., Gerlt, J.A., and Rayment, I., *Evolution of enzymatic activities in the enolase superfamily: crystallographic and mutagenesis studies of the reaction catalyzed by D- glucarate dehydratase from Escherichia coli.* Biochemistry, 2000. **39**: p. 4590-4602.

35. Olson, J.A., *Spectrophotometric measurement of alpha-keto acid semicarbazones.* Arch. Biochem. Biophys., 1959. **85**: p. 225-233.

36. Evans, P., *Scaling and assessment of data quality.* Acta Cryst., 2005. **D62**: p. 62-72.

37. Szebenyi, D.M.E., Arval, A., Ealick, S., Laluppa, J. M., and Nielsen, C., *A system for integrated collection and analysis of crystallographic diffraction data.* J. Synchrotron Rad., 1997. **4**: p. 128-35.

38. McCoy, A.J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J., *Phaser crystallographic software.* J. Appl. Cryst., 2007. **40**: p. 658-74.

39. Murshudov, G.N., Vagin, A. A., and Dodson, E. J., *Refinement of macromolecular structures by the maximum-likelihood method.* Acta Cryst., 1997. **D53**: p. 240-55.

40. Adams, P.D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwillinger, T. C., and Zwart, P. H. , *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Cryst., 2010. **D66**: p. 213-21.

41. Emsley, P.L., B., Scott, W. G., and Cowtan, K., *Features and development of COOT.* Acta Cryst., 2010. **D66**: p. 486-501.

42. Barber A. E. 2nd, B.P.C., *Pythoscape: a framework for generation of large protein similarity networks.* Bioinformatics, 2012. **28**: p. 2845-6.

43. Shannon P., M.A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., Ideker T., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res., 2003. **13**: p. 2498-504.

44. Yew, W.S., Fedorov, A. A., Fedorov, E. V., Wood, B. M., Almo, S. C., and Gerlt, J. A., *Evolution of enzymatic activities in the enolase superfamily: D-tartrate dehydratase from Bradyrhizobium japonicum.* Biochemistry, 2006. **45**: p. 14598-608.

45. Yew, W.S., Fedorov, A. A., Fedorov, E. V., Almo, S. C., and Gerlt, J. A., *Evolution of enzymatic activities in the enolase superfamily: L-talarate/galactarate dehydratase from Salmonella typhimurium LT2.* Biochemistry, 2007. **46**: p. 9564-77.

46. Kilgore, W.W., and Starr, M. P., *Catabolism of galacturonic and glucuronic acids by Erwinia carotovora.* Biol. Chem., 1959. **234**: p. 2227-36.

47. Chang, Y.F., and Feingold, D. S., *Hexuronic acid dehydrogenase of Agrobacterium tumefaciens.* J. Bacteriol., 1969. **99**: p. 667-73.

48. Condemine, G., Robert-Baudouy, J., *Analysis of an Erwinia chrysanthemi gene cluster involved in pectin degradation.* Mol. Microbiol., 1991. **5**: p. 2191-202.

49. Vick, J.E., Schmidt, D. M., and Gerlt, J. A., *Evolutionary potential of (alpha/beta)$_8$-barrels: In vitro enhancement of a "new" reaction in teh enolase superfamily.* Biochemistry, 2005. **44**: p. 11722-29.

50. Neidhart, D.J., Kenyon, G. L., Gerlt, J. A., and Petsko, G. A., *Mandelate Racemase and Muconate Lactonizing Enzyme are Mechanistically Distinct and Structurally Homologous.* Nature, 1990. **347**: p. 692-4.

51. Schmidt, D.M.Z., Hubbard, B. K., and Gerlt, J. A., *Evolution of Enzymatic Activities in the Enolase Superfamily: Functional Assignment of Unknown Proteins in Bacillus subtilis and Escherichia coli as L-Ala-D/L-Glu Epimerases.* Biochemistry, 2001. **40**: p. 15707-15.

52. Wichelecki, D.J., Balthazor, B. M., Chau, A. A., Vetting, M. W., Fedorov, A. A., Fedorov, E. V., Lukk, T., Patskovsky, Y. V., Stead, M. B., Hillerich, B. S., Seidel, R. D., Almo, S. C., and Gerlt, J. A., *Discovery of function in the enolase superfamily: D-mannonate and D-gluconate dehydratases in the D-mannonate dehydratase subgroup.* Submitted for publication to Biochemistry, 2014.

53. Zhang, Q., Gao, F., Peng, H., Cheng, H., Liu, Y., Tang, J., Thompson, J., Wei, G., Zhang, J., Du, Y., Yan, J., Gao, G. F., *Crystal Structures of Streptococcus suis Mannonate*

Dehydratase (ManD) and Its Complex with Substrate: Genetic and Biochemical Evidence for a Catalytic Mechanism. Journal of Bacteriology, 2009. **191**: p. 5832-7.

54. Hottes, A.K., Meewan, M., Yang, D., Arana, N., Romero, P., McAdams, H. H., and Stephens, C., *Transcriptional Profiling of Caulobacter crescentus during Growth on Complex and Minimal Media.* Journal of Bacteriology, 2003. **186**: p. 1448-61.

55. Seifert, G.J., *Nucleotide Sugar Interconversions and Cell Wall Biosynthesis: How to Bring the Inside to the Outside.* Curr. Opin. Plant Biol., 2004. **7**: p. 277-84.

56. Toporowski, M.C., Nomellini, J. F., Awram, P., and Smit, J., *Two Outer Membrane Proteins are Required for Maximal Type I Secretion of the Caulobacter crescentus S-Layer Protein.* Journal of Bacteriology, 2004. **186**: p. 8000-9.

57. Gilchrist, A.a.S., J., *Transformation of Freshwater and Marine Caulobacters by Electroporation.* Journal of Bacteriology, 1991. **173**: p. 921-5.

58. Khan, S.R., Gaines, J., Roop II, R. M., and Farrand, S. K. , *Broad-Host-Range Expression Vectors with Tightly Regulated Promoters and Their Use to Examine the Influence of TraR and TraM Expression on Ti Plasmid Quorum Sensing.* Applied and Environmental Microbiology, 2008. **74**: p. 5053-62.

59. Wichelecki, D.J., Balthazor, B. M., Chau, A. C., Vetting, M. W., Fedorov, A. A., Fedorov, E. V., Lukk, T., Patskovsky, Y. V., Stead, M. B., Hillerich, B. S., Seidel, R. D., Almo, S. C., and Gerlt, J. A., *Discovery of function in the enolase superfamily: D-mannonate and D-gluconate dehydratases in the D-mannonate dehydratase subgroup.* Biochemistry, 2014. **53**: p. 2722-2731.

60. Wichelecki, D.J., Graff, D. C., Al-Obaidi, N., Almo, S. C., and Gerlt, J. A., *Identification of the physiological role of a high efficiency enolase superfamily member mannonate dehydratase in Caulobacter crescentus NA1000.* Biochemistry, 2014.

61. Arahal, D.R., Garcia, M. T., Vargas, C., Canovas, D., Nieto, J. J., and Ventosa, A., *Chromohalobacter salexigens sp. nov., a moderately halophilic species that includes Halomonas elongata DSM 3043 and ATCC 33174.* International Journal of Systematic and Evolutionary Microbiology, 2001. **51**: p. 1457-62.

62. Datsenko, K.A.a.W., B. L., *One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.* Proc Natl Acad Sci U S A, 2000. **97**: p. 6640-5.

63. Chang Y. F., a.F., D. S., *Hexuronic acid dehydrogenase of Agrobacterium tumefaciens.* J. Bacteriol., 1969. **99**: p. 667-73.

64. Wagner, G., and Hollmann, S., *Uronic acid dehydrogenase from Pseudomonas syringae. Purification and properties.* Eur. J. Biochem., 1976. **61**: p. 589-96.

65. Blumenthal, H.J., and Fish, D. C., *Bacterial conversion of D-glucarate to glycerate and pyruvate.* Biochem. Biophys. Res. Commun., 1963. **11**: p. 239-43.

66. Bouvier, J.T., Groninger-Poe, F. P., Vetting, M., Almo, S. C., and Gerlt, J. A., *Galactaro delta-lactone isomerase: lactone isomerization by a member of the amidohydrolase superfamily.* Biochemistry, 2014. **53**: p. 614-616.

67. Andberg, M., Maaheimo, H., Boer, H., Penttilä, M., Koivula, A., and Richard, P., *Characterization of a novel Agrobacterium tumefaciens galactarolactone cycloisomerase enzyme for direct conversion of d-galactarolactone to 3-deoxy-2-keto-l-threo-hexarate.* J Biol Chem, 2012. **287**: p. 17662-17671.

68. Cooper, R.A., *The pathway for L-gulonate catabolism in Escherichia coli K-12 and Salmonella Typhimurium LT-2.* FEBS Letters, 1980. **115**: p. 63-67.

69. Bausch, C., Peekhaus, N., Utz, C., Blais, T., Murray, E., Lowary, T., and Conway, T., *Sequence analysis of the GntII (subsidiary) system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in Eschericia coli.* Journal of Bacteriology, 1998. **180**: p. 3704-3710.

70. Jenson, R.A., *Enzyme recruitment in evolution of new function.* Annu. Rev. Microbiol., 1976. **30**: p. 409-25.

71. Peisajovich, S.G., Tawfik, D. S., *Protein engineers turned evolutionists.* Nat. Methods, 2007. **4**: p. 991-94.

72. Dolnick, B.J., *Cloning and Characterization of a Naturally Occurring Antisense RNA to Human Thymidylate Synthase mRNA.* Nucleic Acids Research, 1993. **21**(8): p. 1747-1752.

73. Dolnick, B.J.a.B., A.R., *Alternate Splicing of the rTS Gene Product and Its Overexpression in a 5-Fluorouracil-resistant Cell Lin.* Cancer Research, 1996(56): p. 3207-3210.

74. Dolnick, B.J., Lu, K., Yin, M., and Rustum, Y.M., *Recent Advances in the Study of rTS Proteins. rTS Expression During Growth and in Response to Thymidylate Synthase Inhibitors in Human Tumor Cells.* Advan. Enzyme Regul., 1997. **37**: p. 95-109.

75. Dolnick, B.J., Black, A.M., Winkler, P.M., Schindler, K., and Hsueh, C., *rTS Gene Expression is Associated with Altered Cell Sensitivity to Thymidylate Synthase Inhibitors.* Advan. Enzyme Regul. , 1996. **36**: p. 165-180.

76. Lin, Y.L.a.C., K.C., *rTS Beta as a Novel 5-Fluorouracil Resistance Marker of Colorectal Caner: A Preliminary Study.* Annals Academy of Medicine Singapore, 2010. **39**: p. 107-111.

77. Srimatkandada, S., Medina, W. D., Cashmore, A. R., Whyte, W., Engel, D., Moroson, B. A., Franco, C. T., Dube, S. K., and Bertino, J. R., *Amplification and organization of dihydrofolate reductase genes in a human leukemic cell line, K-562, resistant to methotrexate.* Biochemistry, 1983. **56**: p. 5774-81.

78. Black, A.R., and Dolnick, B. J., *Expression of rTS correlates with altered growth regulation of thymidylate synthase.* Cancer Research, 1996. **56**: p. 700-5.

79. Kuo, S.J., Wang, H.C., Chow, K.C., Chiou, S.H., Chiang, S.F., Lin T.Y., Chiang, I.P., and Chen D.R., *Expression of rTSβ as a 5-Fluorouracil Resistance marker in patients with Primary Breast Cancer.* Oncology Reports, 2008. **19**: p. 881-888.

80. Liang, P., Nair, J.R., Song, L., McGuire, J.J., and Dolnick, B.J., *Comparative Genomic Analysis Reveals a Novel Mitochondrial Isoform of Human rTS protein and Unusual Phylogenetic Distribution of the rTS Gene.* BMC Genomics, 2005. **6**(125).

81. Gerlt, J.A., Babbitt, P.C., and Rayment, I., *Divergent Evolution in the Enolase Superfamily: The Interplay of Mechanism and Specificity.* Archives of Biochemistry and Biophysics, 2005. **433**: p. 59-70.

82. Babbitt, P.C., Mrachko, G.T., Hasson, M.S., Huisman, G.W., Kolter, R., Ringe, D., Petsko, G.A., Kenyon, G.L., and Gerlt, J.A., *A Functionally Diverse Enzyme Superfamily that Abstracts the Alpha Protons of Carboxylic Acids.* Science, 1995. **267**: p. 1159-1161.

83. Glasner, M.E., Gerlt, J.A., and Babbitt, P.C., *Evolution of Enzyme Superfamilies.* Curr. Opin. Chem. Biol., 2006. **10**: p. 492-497.

84. Kabsch, W., *Integration, scaling, space-group assignment and post-refinement.* Acta Cryst., 2010. **D66**: p. 133-144.

85. CCP4, *The CCP4 suite: programs for protein crystallography.* Acta Cryst., 1994. **D50**: p. 760-763.

86. Sheldrick, G.M., *A short history of SHELX.* Acta Cryst., 2008. **A64**: p. 112-122.

87. Vonrhein, C., et al., *Automated structure solution with autoSHARP.* Methods Mol. Biol., 2007. **364**: p. 215-230.

88. Abrahams, J.P. and A.G.W. Leslie, *Methods used in the structure determination of bovine mitochondrial F1 ATPase.* Acta Cryst., 1996. **D52**: p. 30-42.

89. Cowtan, K., *The Buccaneer software for automated model building. 1. Tracing protein chains.* Acta Cryst., 2006. **D62**: p. 1002-1011.

90. Emsley, P. and K.D. Cowtan, *Coot: model-building tools for molecular graphics.* Acta Cryst., 2004. **D60**: p. 2126-2132.

91. Gulick, A.M., Hubbard, B.K., Gerlt, J.A., and Rayment, I., *Evolution of Enzymatic Activities in the Enolase Superfamily: Identification of the General Acid Catalyst in the Active Site of D-Glucarate Dehydratase from Escherichia coli.* Biochemistry, 2001. **40**: p. 10054 - 10062.

92. Yew, W.S., Fedorov, A.A., Fedorov, E.V., Rakus, J.F., Pierce, R.W., Almo, S.C., and Gerlt, J.A., *Evolution of Enzymatic Activities in the Enolase Superfamily: L-Fuconate Dehydratase from Xanthomonas campestris.* Biochemistry, 2006. **45**: p. 14582-14597.

93. Niesen, F.H., Berglund, H., and Vedadi, M., *The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability.* Nat Protoc., 2007. **2**: p. 2212-21.

94. Yuen R., a.S., H., *L-Fucose metabolism in mammals. I. Port liver L-fuconate hydro-lyase.* Can. J. Biochem., 1972. **50**: p. 798-806.

95. Becker, D.J., and Lowe, J. B., *Fucose: biosynthesis and biological function in mammals.* Glycobiology, 2003. **13**: p. 41R-53R.

96. Neidhard, D.J., Howell, P.L., Petsko, G.A., Powers, V.M., Li, R.S., Kenyon, G.L., and Gerlt, J.A., *Mechanism of the Reaction Catalyzed by Mandelate Racemase. 2. Crystal Structure of Mandelate Racemase at 2.5-Å resolution: Identification of the Active Site and Possible Catalytic Residues.* Biochemistry, 1991. **30**: p. 9264-9273.

97. Gerlt, J.A., and Raushel, F.M., *Evolution of function in (beta/alpha)(8)-barrel enzymes.* Curr. Opin. Chem. Biol., 2003. **7**: p. 252-264.

98. Yew, W.S., Fedorov, A.A., Fedorov, E.V., Wood, B.M., Almo, S.C., and Gerlt, J.A., *Evolution of Enzymatic Activities in the Enolase Superfamily: D-Tartrate Dehydratase from Bradyrhizobium japonicum.* Biochemistry, 2006. **45**: p. 14598-145608.

99. Yew, W.S., Fedorov, A.A., Fedorov, E.V., Almo, S.C., and Gerlt, J.A., *Evolution of Enzymatic Activities in the Enolase Superfamily: L-Talarate/Galactarate Dehydratase from Salmonella typhimurium LT2.* Biochemistry, 2007. **46**: p. 9564-9577.

100. Rakus J.F., F.A.A., Fedorov E.V., Glasner M.E., Hubbard B.K., Delli J.D., Babbitt P.C., Almo S.C., Gerlt J.A., *Evolution of enzymatic activities in the enolase superfamily: L-rhamnonate dehydratase.* Biochemistry, 2008. **47**: p. 9944 - 9954.

101. Dolnick, B.J., Angelino, N.J., Dolnick, R., and Sufrin, J.R., *A Novel Function for the rTS Gene.* Cancer Biology & Therapy, 2003. **2**(4): p. 364-369.

102. Dolnick, R., Wu, Q., Angelino, N.J., Stephanie, L.V., Chow, K., Sufrin, J.R., and Dolnick, B.J., *Enhancement of 5-Fluorouracil Sensitivity by an rTS Signaling Mimic in H630 Colon Cancer Cells.* Cancer Research, 2005. **65**(13): p. 5917-5924.

103. Dolnick, B.J., *Proteins encoded by the antisense strand gene (rTS) of thymidylate synthase interact with thymidylate synthase and dihydrofolate reductase.* Proc. Am. Assoc. Cancer Res., 1996. **37**: p. 652-653.

104. Vemuri, G.N., and Aristidou, A. A., *Metabolic engineering in the -omics era: elucidating and modulating regulatory networks.* Microbiology and Molecular Biology Reviews, 2005. **69**: p. 197-216.

105. Rabinovitch-Deere, C.A., Oliver, J. W. K., Rodriguez, G. M., and Atsumi, S., *Synthetic biology and metabolic engineering approaches to produce biofuels.* Chemical Reviews, 2013. **113**: p. 4611-4632.

106. Jones, D.T., and Woods, D. R., *Acetone-butanol fermentation revisited.* Microbiol Rev., 1986. **50**: p. 484-524.

107. Li, H., Cann, A. F., and Liao, J. C., *Biofuels: biomolecular engineering fundamentals and advances.* Annu Rev Chem Biomol Eng, 2010. **1**: p. 19-36.

108. Keasling, J.D., *Manufacturing molecules through metabolic engineering.* Science, 2010. **3**: p. 1355-1358.

109. Stephanopoulos, G., *Challenges in engineering microbes for biofuels production.* Science, 2007. **315**: p. 801-804.

110. Rodrigueze, M., and Atsumi, S., *Synthetic biology approaches to produce C3-C6 alcohols from microorganisms.* Curr. Chem. Biol., 2012. **6**: p. 32-41.

111. Wichelecki, D.J., Graff, D. C., Al-Obaidi, N., Almo, S. C., and Gerlt, J. A., *Identification of the physiological role of a high efficiency enolase superfamily member mannonate dehydratase in Caulobacter crescentus NA1000.* Submitted for publication to Biochemistry, 2014.

112. Huisman, G.W., and Kolter, R., *Sensing starvation: a homoserine lactone-dependent signaling pathway in Escherichia coli.* Science, 1994. **265**: p. 537-539.

113. Rost, B., *Enzyme function less conserved than anticipated.* J. Mol. Biol., 2002. **318**: p. 595-608.

114. Wichelecki, D.J., Froese, D. S., Kopec, J., Muniz, J. R. C., Yue, W. W., and Gerlt, J. A., *Enzymatic and structural characterization of rTSgamma provides insights into the function of rTSbeta.* Biochemistry, 2014. **53**: p. 2732-2738.