Assignment of function to a domain of unknown function: DUF1537 is a new kinase family in catabolic pathways for acid sugars

Xinshuai Zhang[a], Michael S. Carter[a], Matthew W. Vetting[b], Brian San Francisco[a], Suwen Zhao[c], Nawar F. Al-Obaidi[b], Jose O. Solbiati[a], Jennifer J. Thiaville[d], Valérie de Crécy-Lagard[d], Matthew P. Jacobson[c], Steven C. Almo[b], and John A. Gerlt[a,e,f,1]

[a]Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [b]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461; [c]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94158; [d]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611; [e]Department of Biochemistry, University of Illinois at Urbana–Champaign, Urbana, IL 61801; and [f]Department of Chemistry, University of Illinois at Urbana–Champaign, Urbana, IL 61801

Edited by Gregory A. Petsko, Weill Cornell Medical College, New York, NY, and approved June 7, 2016 (received for review April 5, 2016)

Using a large-scale "genomic enzymology" approach, we (*i*) assigned novel ATP-dependent four-carbon acid sugar kinase functions to members of the DUF1537 protein family (domain of unknown function; Pfam families PF07005 and PF17042) and (*ii*) discovered novel catabolic pathways for D-threonate, L-threonate, and D-erythronate. The experimentally determined ligand specificities of several solute binding proteins (SBPs) for TRAP (tripartite ATP-independent permease) transporters for four-carbon acids, including D-erythronate and L-erythronate, were used to constrain the substrates for the catabolic pathways that degrade the SBP ligands to intermediates in central carbon metabolism. Sequence similarity networks and genome neighborhood networks were used to identify the enzyme components of the pathways. Conserved genome neighborhoods encoded SBPs as well as permease components of the TRAP transporters, members of the DUF1537 family, and a member of the 4-hydroxy-L-threonine 4-phosphate dehydrogenase (PdxA) oxidative decarboxylase, class II aldolase, or ribulose 1,5-bisphosphate carboxylase/oxygenase, large subunit (RuBisCO) superfamily. Because the characterized substrates of members of the PdxA, class II aldolase, and RuBisCO superfamilies are phosphorylated, we postulated that the members of the DUF1537 family are novel ATP-dependent kinases that participate in catabolic pathways for four-carbon acid sugars. We determined that (*i*) the DUF1537/PdxA pair participates in a pathway for the conversion of D-threonate to dihydroxyacetone phosphate and CO_2 and (*ii*) the DUF1537/class II aldolase pair participates in pathways for the conversion of D-erythronate and L-threonate (epimers at carbon-3) to dihydroxyacetone phosphate and CO_2. The physiological importance of these pathways was demonstrated in vivo by phenotypic and genetic analyses.

DUF1537 | kinase | four-carbon acid sugars | conserved genome neighborhoods | genomic enzymology

As a result of advances in genome sequencing, the number of sequences in the protein databases is rapidly increasing: for example, >60 million sequences in release 2016_02 of the UniProt database that is increasing in size at the rate of ∼2% per month (1). Approximately two-thirds of the proteins identified in genome projects have annotations based on sequence homology to previously annotated proteins; the remaining proteins are classified as "hypothetical proteins" or "uncharacterized proteins" (2, 3). However, this automated process results in propagation of misleading or incorrect annotations (4–7). Correction of annotations by experimental characterization is essential for realizing the value of genome sequences.

The Pfam database organizes the "protein universe" into homologous families [16,295 protein families in Pfam release 29.0 (8)]. In release 29.0, 3,892 families are annotated as a "domain of unknown function" (DUF) because no member has an experimentally characterized function (9). Given their wide taxonomic distribution, including bacterial pathogens, many DUFs are likely biologically essential (10). Thus, reliable experimental identification of the in vitro enzymatic activities and in vivo physiological functions for the DUF families is important. However, the assignment of functions to uncharacterized proteins is challenging (11–13). The methods now available for discovering the functions of uncharacterized proteins, including DUFs, are inefficient and often depend on inference from the functions of characterized homologs (11). Therefore, new strategies are required to confront and solve the functional assignment challenge.

The DUF1537 family (Pfam families PF07005 and PF17042) contains 4,610 sequences in release 2016_02 of the UniProt database; members of DUF1537 are found in diverse bacterial phyla, including *Proteobacteria*, *Firmicutes*, *Cyanobacteria*, *Actinobacteria*, and *Bacteroidetes*. Some organisms contain multiple members of the family, suggesting diverse biological functions. In the Uniprot and National Center for Biotechnology Information databases, DUF1537 proteins often are annotated as "Hrp (hypersensitive reaction and pathogenicity) type III effectors" on the basis of remote functional relationships and low sequence homology. "Hrp" proteins commonly are secreted via a type III secretion system and often are involved in plant tissue infections

Significance

Domain of unknown function (DUF) families constitute 3,892 of the 16,295 families in the Pfam database (release 29.0). Given their biological importance, large-scale strategies are required to accomplish their functional assignments. Here, we illustrate an integrated "genomic enzymology" strategy to identify diverse functions within the DUF1537 family (PF07005). We combined high-throughput ligand screening results for transport system solute binding proteins with the synergetic analysis of sequence similarity networks and genome neighborhood networks to establish that the members of the DUF1537 family are novel ATP-dependent four-carbon sugar kinases. This study illustrates the utility of this strategy and enhances our knowledge of bacterial carbohydrate catabolism.

Author contributions: X.Z., M.S.C., M.W.V., S.Z., V.d.C.-L., M.P.J., S.C.A., and J.A.G. designed research; X.Z., M.S.C., M.W.V., B.S.F., N.F.A.-O., J.O.S., and J.J.T. performed research; X.Z., M.S.C., and N.F.A.-O. contributed new reagents/analytic tools; X.Z., M.S.C., M.W.V., B.S.F., J.O.S., J.J.T., and J.A.G. analyzed data; and X.Z., M.S.C., M.W.V., M.P.J., S.C.A., and J.A.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data Bank, www.pdb.org (PDB ID codes 4XGJ, 4XFM, 4XFR, 4XG0, and 5DMH).

[1]To whom correspondence should be addressed. Email: j-gerlt@illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1605546113/-/DCSupplemental.

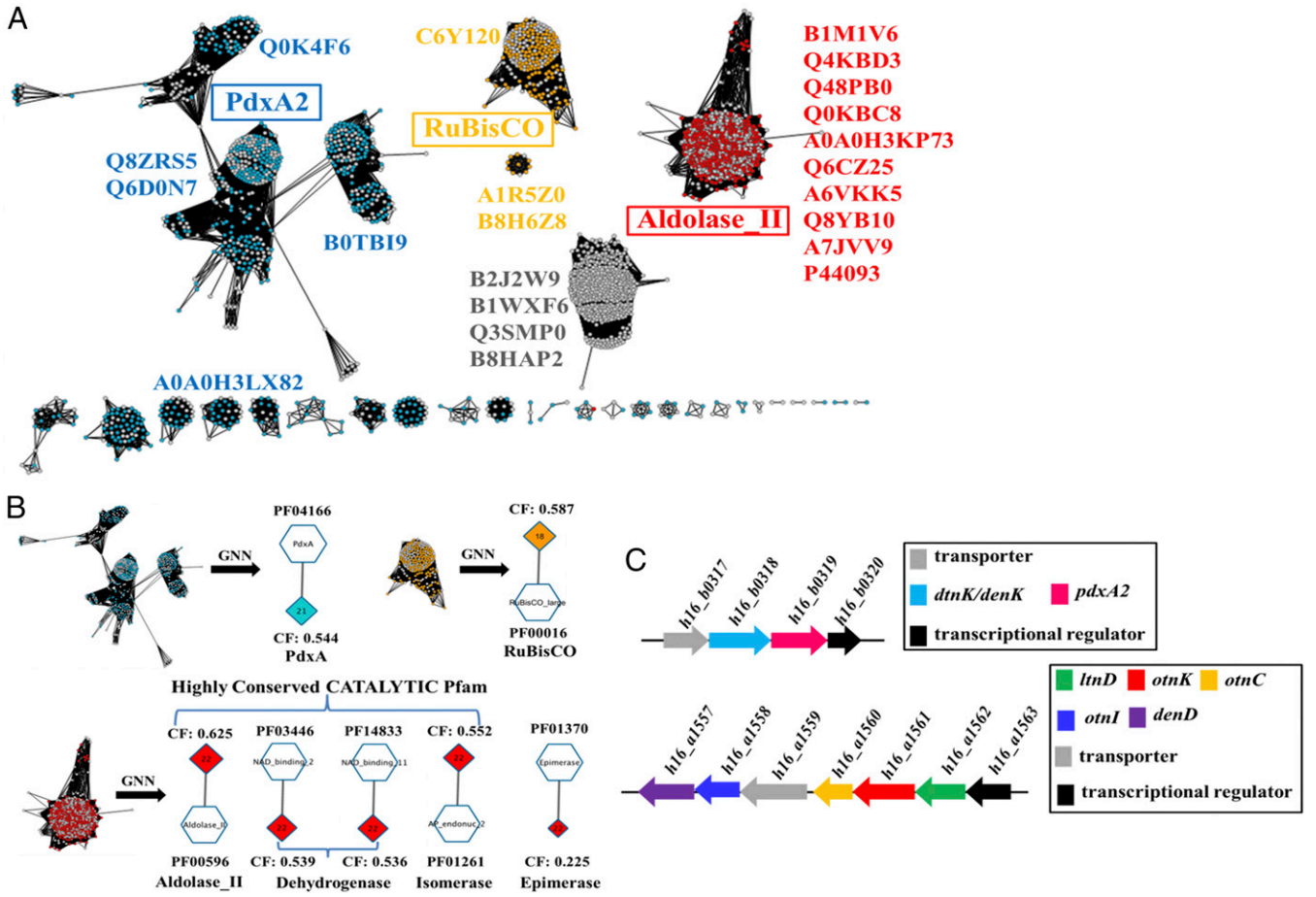

Fig. 1. SSN/GNN analysis of DUF1537 family. (*A*) SSN for the DUF1537 family (PF07005) displayed at an alignment score of 60 (~35% sequence identity). The nodes in the network are colored based on conserved genome neighborhoods that encode the members of respective nodes. PdxA2–, aldolase–, and RLP–DUF1537 proteins are denoted by the sea blue, red, and orange nodes, respectively. The UniProt IDs for 22 DUF1537 proteins are also indicated (*SI Appendix*, Table S1). (*B*) The cluster-specific GNNs for PdxA2–, aldolase– and RLP–DUF1537 clusters and only the highly conserved [indicated by cluster fraction (CF)], catalytic genome proximal enzyme families specific to these clusters were shown here. (*C*) Representative PdxA2–DUF1537 and aldolase–DUF1537 gene clusters encoded in *R. eutropha* H16.

(14). Although evidence of secretion has not been reported for DUF1537 proteins, one bioinformatic analysis predicted a secretory signal on the N terminus of the *Pseudomonas syringae* DUF1537 protein HopAN1 (Hrp outer protein effector protein) (Uniprot: Q87V79) (15). The gene encoding VguB (virulence gluconate metabolism), a member of the DUF1537 family, was associated with plant virulence and D-gluconate metabolism in *Pectobacterium carotovorum* WPP14 (16). In addition, members of the family are elevated in *Burkholderia pseudomallei* in a human infection model (17). Both the phylogenetic diversity of species with members of the DUF1537 family and the correlation of the proteins with virulence make them productive targets for functional characterization.

As described in this article, we used in vitro and in vivo functional assignment of members of the DUF1537 family to further develop our large-scale genomic enzymology-based strategy for functional assignment of novel enzymes in novel microbial catabolic pathways (18). The strategy takes advantage of the frequently observed genomic colocation of microbial genes that encode the transport systems and enzymes for catabolism of an extracellular solute.

The first step in our strategy is the experimental screening of the ligand specificity for a transport system solute binding protein (SBP) (18). This step identifies the starting metabolite for the pathway and helps locate the genes that encode the enzyme components of the pathway. We focus on the SBPs of bacterial TRAP (tripartite ATP-independent permease) and ATP-binding cassette transport systems; both have an extracellular SBP that binds and delivers its ligand to the integral membrane permease components for transport into the cell (19, 20). We then synergistically use protein family sequence similarity networks (SSNs) and genome neighborhood networks (GNNs) to discover the enzyme components of the pathway and infer their functions (21). A SSN is a readily accessible method (constructed with the Enzyme Function Initiative-Enzyme Similarity Tool web tool) for segregating protein families, including those of SBPs as well as enzymes, into isofunctional groups (22). A GNN (constructed with the Enzyme Function Initiative-Genome Neighborhood Tool web tool) then enables identification of conserved genome neighborhoods for isofunctional groups in the SSN and partitions the enzymes encoded by these neighborhoods into Pfam families, aiding inference of the reactions in the pathway given the expected identity of the substrate for the pathway (the ligand for the SBP).

In this study, we used the experimentally determined specificities of four orthologous TRAP SBPs for four-carbon acid sugars to predict and then experimentally assign kinase functions to members of the DUF1537 family (Pfam families PF07005 and PF17042). With our integrated SSN and GNN analysis, we identified the enzymes and inferred the reactions in three novel catabolic pathways for D-erythronate, D-threonate, and L-threonate; we then biochemically

Table 1. Selected kinetic parameters for DtnK, PdxA2, LtnD, and DenD

Uniprot (annotation)	Substrate	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M (M^{-1} s^{-1})
Q0K4F6 (ReDtnK)	D-Threonate	45 ± 4	0.12 ± 0.03	3.8×10^5
	4-Hydroxy-L-threonine	27 ± 2	86 ± 9	3.1×10^2
Q8ZRS5 (SeDtnK)	D-Threonate	22 ± 1.3	0.29 ± 0.05	7.5×10^4
	4-Hydroxy-L-threonine	4.0 ± 0.1	3.8 ± 0.5	1.1×10^3
Q0K4F5 (RePdxA2)	D-Threonate 4-phosphate	3.4 ± 0.09	0.12 ± 0.01	2.9×10^4
	4-Hydroxy-L-threonine 4-phosphate	0.35 ± 0.005	0.20 ± 0.02	1.8×10^3
P58718 (SePdxA2)	D-Threonate 4-phosphate	8.9 ± 0.2	0.054 ± 0.006	1.6×10^5
	4-Hydroxy-L-threonine 4-phosphate	0.37 ± 0.02	0.19 ± 0.04	1.9×10^3
Q0KBC7 (ReLtnD)	L-Threonate, NAD^+	29 ± 0.7	0.35 ± 0.04	8.5×10^4
	L-Threonate, $NADP^+$	2.2 ± 0.07	0.30 ± 0.04	7.5×10^3
Q6CZ26 (PaLtnD)	L-Threonate, NAD^+	54 ± 0.7	0.13 ± 0.007	4.1×10^5
	L-Threonate, $NADP^+$	4.7 ± 0.1	0.97 ± 0.1	4.8×10^3
Q0KBD2 (ReDenD)	D-Erythronate, NAD^+	19 ± 0.3	0.59 ± 0.04	3.3×10^4
	D-Erythronate, $NADP^+$	3.0 ± 0.08	2.4 ± 0.3	1.2×10^3

Error is the SD.

and physiologically verified those predictions. We expect that this strategy will be useful for the discovery of other novel metabolic pathways as well as assigning functions to other DUF families.

Results

Synergistic Analysis of SSNs and GNNs Enables the Prediction That Members of DUF1537 Are Novel Four-Carbon Acid Sugar Kinases. We previously identified four SBPs that bind four-carbon acid sugars, including D-erythronate and L-erythronate (18). When the SSN for the TRAP SBP family (IPR018389) is filtered at an alignment score threshold of 80 (~45% sequence identity) (*SI Appendix*, Fig. S1*A*), the SBPs (Uniprot: Q12HD7, Q12CD8, A1WPV4, and Q1QSK0) are located in two clusters. A GNN generated using the SSN for the TRAP SBP family revealed that many of the genes encoding the members of the SSN clusters are proximal to those encoding members of the DUF1537 family (PF07005) (*SI Appendix*, Fig. S1*B*). This colocalization allows the hypothesis that members of DUF1537 are involved in the catabolism of four-carbon acid sugars.

SSNs for the DUF1537 family (PF07005 in Pfam 29.0) were constructed with differing alignment scores (*SI Appendix*, Fig. S2) and used to generate GNNs (22). Using an SSN filtered at an alignment score of 60 (~35% sequence identity), the GNN identified proteins encoded by three distinct genome neighborhoods that appear to encode distinct metabolic pathways (Fig. 1); these are distinguished by the presence of genes for members of the 4-hydroxy-L-threonine 4-phosphate dehydrogenase (PdxA) (PF04166), class II aldolase (PF00596), or ribulose 1,5-bisphosphate carboxylase/oxygenase, large subunit (RuBisCO) (PF00016) families (Fig. 1*B*). In addition, some members of the DUF1537 family are fusion proteins with a member of either the PdxA family (32 sequences from *Ralstonia*, *Arthrobacter*, *Rhodococcus*, and *Burkholderia*) or the class II aldolase family (five sequences primarily from *Clostridium*), thereby confirming a functional relationship (23).

We hypothesized that members of the DUF1537 family catalyze the ATP-dependent phosphorylation of four-carbon acid sugars because: (*i*) the substrates for members of the PdxA, class II aldolase, and RuBisCO superfamilies are phosphorylated; (*ii*) the genome neighborhoods lacked known kinase genes; and (*iii*) the ligands for the SBPs that were used to target the DUF1537 family are four-carbon acid sugars. In the analyses that follow, the DUF1537 proteins, genome neighborhoods, and metabolic pathways are designated as PdxA2–, aldolase–, or RLP–DUF1537 proteins, neighborhoods, and pathways based on the genome neighbors identified in the GNN for the DUF1537 clusters [the RuBisCO homologs belong to the RuBisCO-like protein (RLP) clade of the RuBisCO family; known RLP proteins do not catalyze the carboxylation or oxygenation reactions catalyzed by autotrophic RuBisCOs (24)].

Biochemical Characterization of the PdxA2–DUF1537 Pathway. High-throughput protein production for members of the DUF1537 family was performed (25). We produced 22 of 122 target proteins (cloned from an in-house collection of ~400 gDNAs); at least one purified protein was obtained for most of the major clusters in the DUF1537 SSN shown in Fig. 1*A*. A four-carbon sugar library with 22 potential substrates was assembled, including four-carbon mono- and diacid sugars, aldotetroses, ketotetroses, alditols, and modified sugars/analogs (*SI Appendix*, Fig. S3*A*). The purified proteins were screened for ATP-dependent kinase activity using the library. Activity was detected for 20 proteins; the substrates were D-threonate or D-erythronate (*SI Appendix*, Table S1).

We first focused on members of the DUF1537 family that are encoded genome proximal to members of the PdxA family (Fig. 1 and *SI Appendix*, Fig. S4; see *SI Appendix*, Table S11 for a full list of proteins from this study). The SSN for the PdxA family filtered with an alignment score of 80 (~47% sequence identity) revealed that the members that are encoded genome proximal to DUF1537 proteins (designated a PdxA2 subgroup) are separated from the proteins (PdxA) that catalyze the $NAD(P)^+$-dependent oxidative decarboxylation of 4-hydroxy-L-threonine 4-phosphate (4HT-4P) to form 3-amino-1-hydroxyacetone 1-phosphate in the biosynthetic pathway for pyridoxal 5′-phosphate (PLP) (*SI Appendix*, Fig. S5) (26, 27), suggesting a distinct function.

Because 4-HT is structurally similar to D-threonate (*SI Appendix*, Fig. S3*B*), a hit in our screen, we chemically synthesized 4-HT (28) and compared the in vitro activities of the DUF1537 proteins with 4-HT, D-threonate, and D-erythronate (Table 1 and *SI Appendix*, Table S2). For five PdxA2–DUF1537 proteins (Uniprot: Q0K4F6, Q8ZRS5, Q6D0N7, B0TBI9, and A0A0H3LX82) (Fig. 1*A* and *SI Appendix*, Table S2), D-threonate or D-erythronate was the preferred substrate; although the values of k_{cat} are similar, the values of K_M for D-threonate/D-erythronate are ~100-fold less than those for 4-HT. The PdxA2–DUF1537 enzymes (DtnK, D-threonate kinase and DenK, D-erythronate kinase) catalyze the ATP-dependent phosphorylation of D-threonate or D-erythronate to generate D-threonate 4-phosphate or D-erythronate 4-phosphate.

The NAD^+-dependent oxidation activities of members of the PdxA2 group for D-threonate 4-phosphate and D-erythronate 4-phosphate (products of DtnK and DenK, respectively) were evaluated by quantitating the reduction of NAD^+. The values of k_{cat}/K_M for D-threonate 4-phosphate or D-erythronate 4-phosphate were 10- to 100-fold greater than those for 4-HT-4P (Table 1 and *SI Appendix*, Table S3). Using both coupled-enzyme spectrophotometric

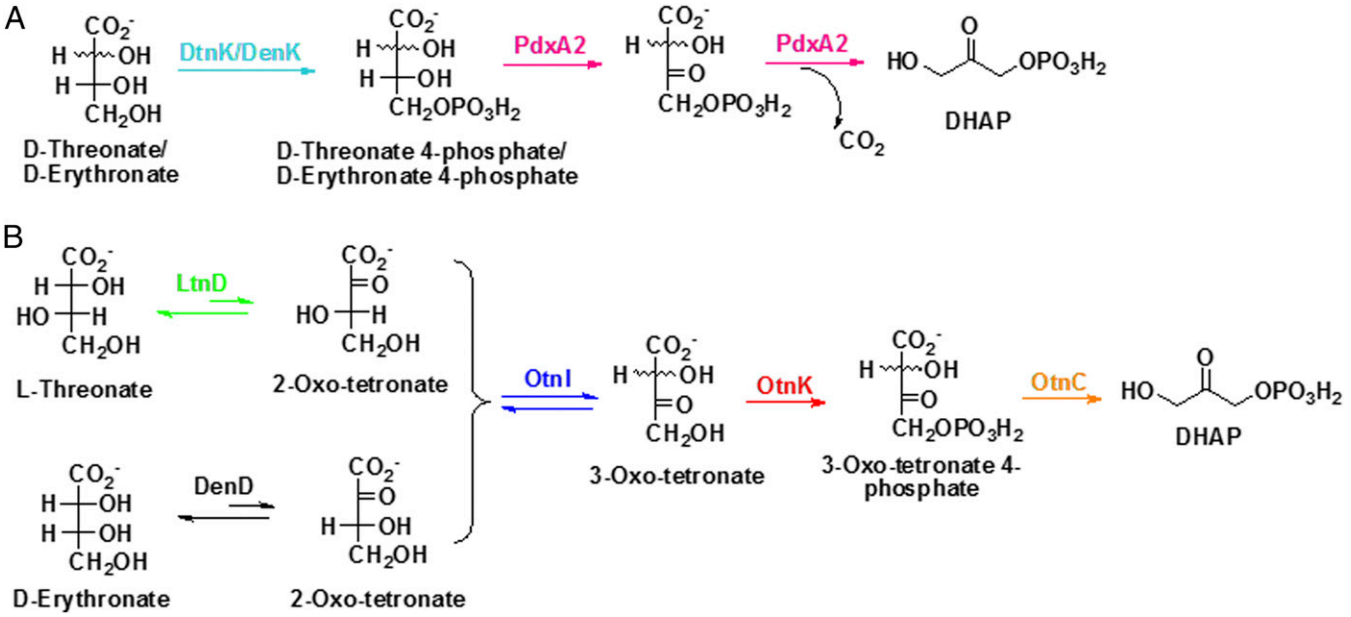

Fig. 2. Pathways for degradation of D/L-threonate and D-erythronate: (*A*) PdxA2–DUF1537 pathway; (*B*) aldolase–DUF1537 pathway.

(*SI Appendix*, Fig. S6) and ^{1}H NMR assays (*SI Appendix*, Fig. S7), we observed that the PdxA2 proteins catalyze the decarboxylation reaction to generate dihydroxyacetone phosphate (DHAP) and CO_2 (analogous to the reaction catalyzed by PdxA). When glycerol 3-phosphate dehydrogenase (G3PDH) was added after the PdxA2-catalyzed reaction was complete, we observed immediate oxidation of NADH (reduction of DHAP) (*SI Appendix*, Fig. S6), providing further support that PdxA2 catalyzes both the oxidation and subsequent decarboxylation of D-threonate 4-phosphate to produce DHAP. Therefore, the in vitro activities of DtnK or DenK and the PdxA2 proteins suggest catabolic pathways in which DtnK or DenK phosphorylates D-threonate or D-erythronate and the members of the PdxA2 group oxidatively decarboxylate the products to generate DHAP and CO_2 (Fig. 2*A*). The weak promiscuities of DtnK and DenK for 4HT and of PdxA2 for 4-HT-4P provide the potential for evolution of a new catabolic pathway.

Biological Characterization of the PdxA2–DUF1537 Pathway. The physiological roles of DtnK and its associated PdxA2 were studied by deleting the gene encoding each protein in *Salmonella enterica* ser. Typhimurium LT2. In contrast to the wild-type strain, the SeΔ*dtnK* and SeΔ*pdxA2* strains were unable to use D-threonate as a carbon source (Fig. 3 and *SI Appendix*, Fig. S13). When the full-length version of the deleted coding region was provided *in trans* in each mutant strain, the resulting growth mirrored that of the wild-type strain carrying the same empty plasmid (Fig. 3 and *SI Appendix*, Fig. S13); therefore, both genes are required for D-threonate growth.

To investigate whether the physiological role of the *dtnK* gene cluster is conserved despite its phylogenetic source (different phylogenetic class), *dtnK* (UniProt: Q0K4F6) was deleted in *Ralstonia eutropha* H16. The resulting strain also demonstrated impaired D-threonate growth (*SI Appendix*, Fig. S16).

Biochemical Characterization of the Aldolase–DUF1537 Pathway. For the members of the DUF1537 family encoded by gene clusters that encode members of the class II aldolase family (PF00596) (Fig. 1 and *SI Appendix*, Fig. S8), the kinase screen detected activities with D-threonate (*SI Appendix*, Table S1), although the value of k_{cat}/K_M was low (29 M^{-1} s^{-1} for A0A0H2VCE6). Therefore, we expected that a structural analog of D-threonate is the physiological substrate. The GNN revealed that this genome context also encodes both an isomerase (PF01261) and a dehydrogenase (PF03446 and PF14833) (Fig. 1*B* and *SI Appendix*, Fig. S8). Four dehydrogenases encoded by the aldolase–DUF1537 gene clusters from different organisms (Uniprot: Q6CZ26, A0A0H2VA68, P44979, and Q0KBC7) were screened using a library of 53 sugars (*SI Appendix*, Table S6). This library included D-gluconate (a possible substrate for a member of the DUF1537 family implicated by studies with *P. carotovorum* (16); and see *Biological Characterization of the Aldolase–DUF1537 Pathway*); however, activity was detected only with L-threonate. The dehydrogenases can use either NAD^+ or $NADP^+$ as cosubstrate, with a preference for NAD^+; the value of k_{cat}/K_M for oxidation of L-threonate with NAD^+ was ~10^5 M^{-1} s^{-1} (Table 1 and *SI Appendix*, Table S4).

Assuming that oxidation of L-threonate is the first reaction in the aldolase–DUF1537 pathway, we conducted a series of coupled-enzyme assays to delineate the sequence of reactions catalyzed by the members of the dehydrogenase, isomerase, and DUF1537 families (*SI Appendix*, Fig. S9 *D–F*). L-Threonate is oxidized (LtnD, L-threonate dehydrogenase), isomerized (OtnI, 2-oxo-tetronate isomerase), and phosphorylated (OtnK, 3-oxo-tetronate kinase) before conversion to dihydroxyacetone phosphate and CO_2 by the member of the class II aldolase family (OtnC, 3-oxo-tetronate 4-phosphate decarboxylase). The value of k_{cat}/K_M for the decarboxylase-catalyzed reaction was estimated (*SI Appendix*, Fig. S10 and Table S5) and is consistent with the predicted physiological role of the reaction. The same order of reactions was established for the enzymes from the gene clusters in *Pectobacterium atrosepticum* SCRI 1043, *Haemophilus influenza* KW20, and *R. eutropha* H16 (*SI Appendix*, Fig. S11 *A*, *B*, and *D*).

The identities of the pathway intermediates were established by considering that the final products were possible only if the substrate for the decarboxylase (OtnC) was 3-oxo-tetronate 4-phosphate. We propose that the β-ketoacid decarboxylation reaction involves the formation of an enolate anion stabilized by a divalent metal, by virtue of its membership in the class II aldolase family (29). Therefore, OtnK, the member of the DUF1537 family, must have phosphorylated 3-oxo-tetronate, the isomerization product of OtnI. Our results, however, could not distinguish whether the substrate for OtnI/product of LtnD was 2-oxo- or 4-oxo-tetronate until we demonstrated in vivo that hydroxypyruvate

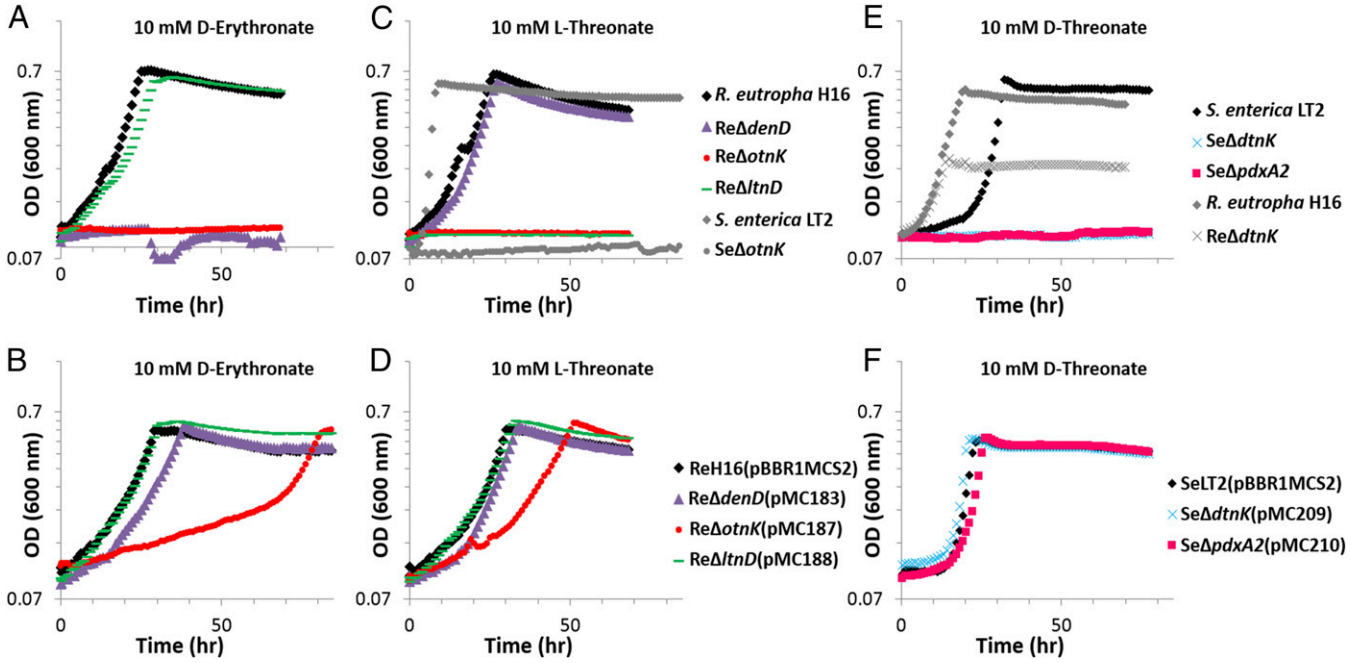

Fig. 3. Growth of selected wild-type, mutant strains (*A*, *C*, *E*), and complemented strains (*B*, *D*, *F*) with 10 mM D-erythronate (*A*, *B*), 10 mM L-threonate (*C*, *D*), or 10 mM D-threonate (*E*, *F*). The same color (similar to *SI Appendix*, Figs. S4 and S8) and marker represent growth of a strain (*A*, *C*, *E*) and the corresponding complemented strain (*B*, *D*, *F*) for each growth condition. Gray markers represent strains for which no complementation data were collected. Strain identities are presented in the above legends. Plasmid identities are detailed in *SI Appendix*, Table S13 and strain identities are detailed in *SI Appendix*, Table S12. For complete growth data, see *SI Appendix*, Figs. S13–S17.

isomerase (Hyi, UniProt: Q0K5R4) compensated for the loss of OtnI (UniProt: Q0KBD1) (see *Biological Characterization of the Aldolase–DUF1537 Pathway*), implying that OtnI catalyzes the isomerization of the carbonyl group between C2 and C3 similar to that catalyzed by Hyi. Therefore, we concluded that the substrate for OtnI was 2-oxo-tetronate, completing the pathway: L-threonate is oxidized to 2-oxo-tetronate (LtnD), 2-oxo-tetronate is isomerized (OtnI) and phosphorylated (OtnK) to 3-oxo-tetronate 4-phosphate, and 3-oxo-tetronate 4-phosphate is decarboxylated (OtnC) to DHAP and CO_2 (Fig. 2*B*).

Approximately 25% of the gene clusters encoding the aldolase–DUF1537 pathway include an additional gene for a protein annotated as "NAD^+-dependent epimerase" (PF01370) (Fig. 1 and *SI Appendix*, Fig. S8). We hypothesized that this protein participates in an alternate version of the pathway, allowing utilization of a different four-carbon acid substrate. Although the protein is annotated as an "epimerase" [homologs catalyze the conversion of UDP-galactose to UDP-glucose in galactose metabolism (30, 31)], the inversion of configuration is accomplished by two successive NAD^+-dependent oxidoreductase reactions. We used our sugar library (*SI Appendix*, Table S6) to screen two orthologous dehydrogenases (Unitprot: P44094 and Q0KBD2) and determined that both catalyze the efficient ($k_{cat}/K_M > 10^3$ M^{-1} s^{-1}) oxidation of D-erythronate with either NAD^+ or $NADP^+$ as cosubstrate, indicating that they are D-erythronate dehydrogenases (DenD) (Table 1 and *SI Appendix*, Table S4). Therefore, we hypothesized that DenD participates in a pathway for the utilization of D-erythronate analogous to that for utilization of L-threonate: DenD catalyzes oxidation of D-erythronate to 2-oxo-tetronate that is further catabolized to DHAP and CO_2 by OtnI, OtnK, and OtnC in the pathway for L-threonate degradation (Figs. 2*B* and *SI Appendix*, Figs. S9 *A–C* and S11 *C* and *E*).

On the basis of these in vitro activities, we conclude that the members of the aldolase (decarboxylase) and DUF1537 (kinase) families participate in convergent pathways for utilization of L-threonate and D-erythronate, with different dehydrogenases generating the common 2-oxo-tetronate (LtnD for L-threonate and DenD for D-erythronate) that is isomerized to 3-oxo-tetronate (OtnI), phosphorylated to 3-oxo-tetronate 4-phosphate (OtnK), and decarboxylated to DHAP and CO_2 (OtnC) (Fig. 2*B*). In this pathway, OtnK catalyzes the ATP-dependent phosphorylation of 3-oxo-tetronate (*SI Appendix*, Fig. S12), a structural analog of the D-threonate substrate for the PdxA2-DUF1537 protein in the pathway for D-threonate utilization, thereby explaining the low value of k_{cat}/K_M observed for D-threonate (3-oxo-tetronate is an unstable β-ketoacid so cannot be added to our screening library).

Biological Characterization of the Aldolase–DUF1537 Pathway. To investigate whether the aldolase–DUF1537 pathway assembled in vitro is functional in vivo, we determined the effects of deleting the genes in the *R. eutropha* H16 aldolase–DUF1537 gene cluster that includes *denD* (Fig. 1*C*). Wild-type cells grow with either L-threonate or D-erythronate as sole carbon source. With the exception of *otnI*, deletion of the genes in the neighborhood abolished growth with L-threonate or D-erythronate (Figs. 3 and *SI Appendix*, Fig. S14). The growth observed for the *otnI* deletion strain, albeit impaired, is explained by a redundant/promiscuous activity from a hydroxypyruvate isomerase [Hyi, Uniprot: Q0K5R4, 58% identical to and encoded in the same genomic context as the authentic hydroxypyruvate isomerase in *Escherichia coli K12* (UniProt: P30147) (32)] that shares 47% sequence identity with OtnI. When *hyi* was deleted, growth with L-threonate and D-erythronate was unaffected. When *hyi* and *otnI* both were deleted in a single strain, growth with L-threonate or D-erythronate was abolished. Because *hyi* is not essential for optimal L-threonate or D-erythronate growth, OtnI is likely the exclusive isomerase for L-threonate and D-erythronate assimilation in the wild-type strain. Deletion of *ltnD* did not affect D-erythronate growth, and deletion of *denD* did not affect L-threonate growth. When each gene was expressed *in trans* in the corresponding mutant strain, L-threonate or D-erythronate growth was restored, confirming that the product of each gene was required for either L-threonate or D-erythronate catabolism (Fig. 3 and *SI Appendix*, Fig. S14).

We observed similar phenotypic results when the gene for OtnK (Uniprot: Q8ZMG5) was deleted in *S. enterica*. Unlike the wild-type strain, Se*ΔotnK* was unable to grow with L-threonate or D-erythronate (*SI Appendix*, Fig. S17), indicating that the physiological role of the aldolase–DUF1537 protein is conserved among species despite a previous report by Mole et al. that an *otnK* gene (designated *vguB* in the study) in *P. carotovorum* WPP14 was required for virulence in plant leaves and for utilization of D-gluconate as a carbon source (16). With *P. carotovorum* wild-type, Pc*ΔotnK*, and complemented strains provided by Mole et al., we observed no connection between the *otnK* operon and D-gluconate growth; instead, our results indicate that the *otnK* operon is required for L-threonate growth (*SI Appendix*, Fig. S15).

Structural Characterization of the DUF1537 Family. The structures of several members of the DUF1537 family (~400 amino acids) were determined by X-ray crystallography. These members included the citrate-bound structure of DtnK from *Bordetella bronchiseptica* RB50 (BbDtnK, Uniprot: A0A0H3LX82), the unliganded and D-threonate-bound structures of DtnK from *P. atrosepticum* SCRI 1043 (PaDtnK, Uniprot: Q6D0N7), and the ADP-bound structure of OtnK from *R. eutrophia* H16 (ReOtnK, Uniprot: Q0KBC8) (*SI Appendix*, Tables S7–S9). In addition, unliganded structures of two OtnK orthologs (*H. influenza* KW20, HiOtnK, Uniprot: P44093, PDB ID code 1YZY; and *S. enterica* ser. Typhimurium LT2, SeOtnK, Uniprot: Q8ZMG5, PDB ID code 3DQQ) had been reported previously (*SI Appendix*, Fig. S18). Together, these structures represent a diverse set of sequences in the DUF1537 family, with ReOtnK, HiOtnK, and SeOtnK sharing ~40–50% sequence identity and ReOtnK, BbDtnK, and PaDtnK sharing ~20% sequence identity (*SI Appendix*, Table S10).

Members of the DUF1537 family are composed of two domains: an N-terminal domain (residues 1–247, ReOtnK numbering) and a C-terminal domain (residues 266–429) connected by a variable linker sequence (Fig. 3*A* and *SI Appendix*, Fig. S19). The N-terminal domain exhibits an α/β-fold composed of an eight-stranded parallel β-sheet (strand order S2N, S3N, S1N, S4N, S11N, S5N, S10N, and S9N) flanked by helices H1N, H5N, H6N, H7N, and H8N on one face and helices H2N, H3N, and H4N on the opposing face of the central β-sheet. The C-terminal domain also exhibits an α/β-fold composed of a seven-stranded mixed β-sheet (strand order S2C, S3C, S1C, S4C, S7C, S6C, and S5C; strands S6C and S7C antiparallel) flanked by helices H1C and H7C on one face and helices H2C, H3C, H4C, H5C, and H6C on the other face of the β-sheet (*SI Appendix*, Fig. S19). A search for structural homologs for either domain, using PDBeFold (33), yielded low quality hits with RMSDs of >3.0 Å over less than 70% of the structural elements. These distant structural homologs yielded no additional information on the function or the location of active site features.

The central β-sheet of the N-terminal domain abuts the central β-sheet of the C-terminal domain in an antiparallel fashion creating a 15-strand continuous β-sheet. In the case of ReOtnK, HiOtnK, and SeOtnK, the continuous β-sheet is formed within the same subunit (i.e., the protein is monomeric), whereas in BbDtnK and PaDtnK the protein crystallized as a domain swapped dimer with the N-terminal domain of one subunit forming an intermolecular β-sheet with the C-terminal domain of a second subunit (*SI Appendix* Fig. S18). Regardless of its composition (inter- or intramolecular), every structure has the conserved N- to C-terminal domain β-sheet/β-sheet interaction and, except for the ADP-bound ReOtnK structure (see below), the relative orientation of these two domains is highly conserved (Fig. 4*A*, open confirmation).

The D-threonate-bound structure of PaDtnK (cocrystallization based on activity screening) demonstrated that the acid-sugar binding site is located in the N-terminal domain adjacent to the domain-domain interface (Fig. 4*B*). Binding of D-threonate produced no large-scale rearrangements compared with the apo structure (RMSD = 0.23 Å). D-Threonate is bound in a solvent-accessible depression and forms nine direct hydrogen bonds with the protein (Fig. 4*C*). Two strictly conserved consecutive Asp residues (Asp16 and Asp17 in PaDtnK) and a conserved Arg (Arg60 in PaDtnK) are responsible for coordination of the 4-OH of D-threonate. Asp16 orientates Arg60, which hydrogen bonds to the 4-OH of D-threonate, whereas Asp17 makes a direct hydrogen bond. The position and strict conservation of Asp17 among all DUF1537 family members suggests that Asp17 is the active site base responsible for activation of the alcohol for nucleophilic attack on the γ-phosphate of ATP.

The structure of ReOtnK, determined from crystals formed in the presence of ATP, yielded unambiguous density for ADP (Fig. 4*D*). ADP is bound only in one of the two molecules (i.e., monomers) per asymmetric unit; the ADP free monomer is in the "open" conformation, whereas in the ADP-bound structure, the C-terminal domain has rotated 30° [as calculated by DynDom (34)] toward the N-terminal domain to form a "closed" structure with a solvent-excluded catalytic site (*SI Appendix*, Fig. S20). ADP is bound predominately by residues donated by the C-terminal domain, again adjacent to the domain–domain interface. The diphosphate moiety is coordinated by three backbone amides donated from the N-terminal end of H6C, along with a fourth backbone amide from Gly414 (ReOtnK numbering), whereas the adenine moiety is bound within a greasy pocket formed by strand 3, helix 2, helix 4, and helix 5 of the C-terminal domain (Fig. 4*E*). The only residues from the N-terminal domain that contact ADP are donated by a loop between strand 8 and helix 4, with the side-chains of Pro156 and Leu157 resting against the nonpolar face of adenosine and His155 hydrogen bonding to the β-phosphate. In general, residues of the nucleotide-binding pocket are not strictly conserved, presumably in part due to the large number of mainchain atoms that are used to coordinate the ligand (six of the nine direct hydrogen bonds).

A molecular model of the ternary complex was constructed by superimposing the N-terminal domain of the D-threonate complex with N-terminal domain of the ADP complex (Fig. 4*F*). The model suggests that a loop between S7C and H7C (i.e., Ser413, Phe416) could additionally be involved in recognition of the sugar substrate when the protein is in the closed ATP bound state. The model highlights the close proximity of the two binding sites when the proteins are in the closed state, where the 4-OH of D-threonate is 5.3 Å from the β-phosphate of ADP and again is consistent with Asp17 being the putative active site base. Structure-based sequence alignments suggest that all family members are kinases with the majority catalyzing the phosphorylation of four-carbon sugars.

Discussion

DUF families are a large set of uncharacterized protein families that are found in the Pfam database, and the number of DUFs is substantially increasing with the rapidly accumulating genome sequencing data (9). Many of the widespread DUFs in bacteria are biologically essential, and the importance of prioritizing these DUFs has been recognized in recent years (10). Before this study, only one DUF family, the DUF849 family of 922 proteins, had been evaluated in large scale by a single report (11), which heavily relied on previously published enzymatic activity and catalytic mechanism/liganded structure of one of its members (35, 36). Recognizing the difficulties associated with DUF functional assignment, we attempt to apply an integrated "genomic enzymology" strategy that had only been used for functional characterization of members of families with known functions (37, 38).

Here, we illustrate the power of the approach to accurately predict and verify functions across the DUF1537 family (PF07005 and PF17042, containing 4,610 sequences). An integrated SSN/GNN analysis of DUF1537 proteins identified three predominant conserved genome neighborhoods and permitted

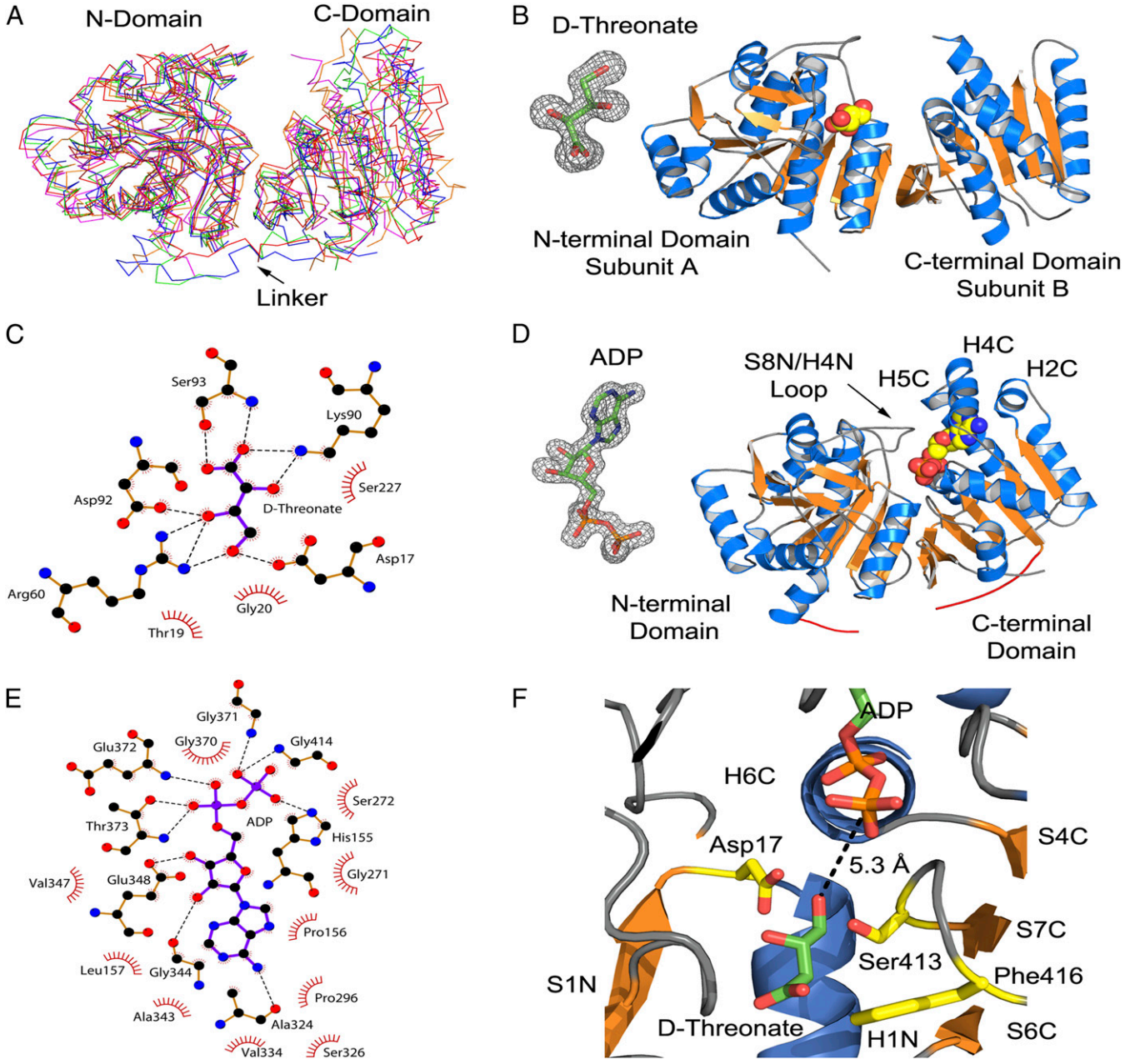

Fig. 4. Crystallography of DUF1537 proteins. (*A*) Superposition of DUF1537 structures in the "open" form. Shown are the Cα traces of P44093 (red), Q8ZMG5 (green), Q0KBC8 (blue), Q6D0N7 (magenta), and A0A0H3LX82 (orange). (*B*) Ribbon diagram of Q6D0N7 in complex with D-threonate (CPK model). The N-terminal domain of subunit A and the C-terminal domain of subunit B are uniquely shown as the potential catalytic unit. (*Inset*) *Fo–Fc* omit electron density map for D-threonate contoured at 3σ. (*C*) LIGPLOT diagram of the interactions of Q6D0N7 with D-threonate. Asp17 is strictly conserved in all DUF1537 proteins, whereas Asp92, conserved among PdxA2–DUF1537 proteins, is a cysteine in Aldolase–DUF1537and RLP–DUF1537 proteins. (*D*) Ribbon diagram of Q0KBC8 in complex with ADP (CPK model). (*Inset*) *Fo–Fc* omit electron density map for ADP contoured at 3σ. (*E*) LIGPLOT diagram of the interactions of Q0KBC8 with ADP. (*F*) Molecular model of a ternary D-threonate/ADP complex.

the prediction that DUF1537 proteins catalyze ATP-dependent phosphorylation. We constrained the possible kinase substrates upon recognizing that some of the neighborhoods also encoded SBPs that bound four-carbon monoacid sugars (18). Directed by the prediction, we confirmed functions for DUF1537 proteins from two of the three predominant genome neighborhoods: the PdxA2–DUF1537 and the aldolase–DUF1537 neighborhoods.

By characterizing multiple phylogenetically distinct DUF1537 proteins from each neighborhood, we demonstrate that the OtnK and DtnK annotations can be extrapolated to the entire corresponding SSN clusters (1,729 OtnK sequences and 724 DtnK sequences out of 3,673 sequences in the DUF1537 SSN) (Dataset S1). Characterization of the remaining genes and gene products from each neighborhood further permitted identification of novel functions in five additional Pfam families (PF04166, PF03446/PF14833, PF01370, PF01261, and PF00596) (*SI Appendix*, Table S11) in metabolic pathways for D-threonate, D-erythronate, and L-threonate.

To our knowledge, the pathways provide the first precedence for D-threonate and D-erythronate as biological substrates and offer the first complete catabolic pathway for L-threonate, a

degradation product of the abundant plant metabolite L-ascorbate (39–42). The presence of a fusion protein (Uniprot: B3H739) in *Arabidopsis thaliana* that includes fused domains similar to LtnD, OtnK, and OtnC suggests that the L-threonate catabolic pathway also might be used by plants, potentially answering the 25-y-old question of the biological and environmental fate of L-ascorbate. Together with the previously identified tetritol metabolism in our laboratory (37), we have filled many of the gaps in four-carbon sugar metabolism in the Kyoto Encyclopedia of Genes and Genomes pathway database.

Members of the DUF1537 family with characterized kinase activities were also structurally characterized. Our liganded structures demonstrate that the acid sugar is bound by the N-terminal domain (PF07005) and nucleotide by the C-terminal domain (PF17042). To our knowledge, the studies also yielded the first "closed" DUF1537 structure where a domain rotation of 30° brings the acid-sugar and nucleotide binding site in proximity for catalysis; this structure provides a reliable model for in silico docking studies of kinases with unknown substrate specificities. For example, the functions of the RLP–DUF1537 proteins from the third conserved genome neighborhood remain unknown and are the subject of ongoing studies. From a structure/function comparison of OtnK and DtnK proteins, we observe that the functional groups at the substrate C3 correlate with genome context related conserved residues [Cys for OtnK, Asp for DtnK (Asp92) (Fig. 4*C*)] that are positioned to coordinate the C3 moiety (carbonyl for OtnK, hydroxyl for DtnK). Because the homologous residue in RLP–DUF1537 proteins is a Cys, we posit that the substrates of RLP–DUF1537 proteins are four-carbon backbone carboxylic acids with a carbonyl group at C3.

Our integrated genomic enzymology strategy, as demonstrated here, is a powerful tool for assigning enzymatic functions across entire protein families and elucidating metabolic pathways, especially when the generic reaction types of families are unknown and other methods, such as computational modeling, are inconclusive. We anticipate we will refine our strategy and experimentally characterize more DUF families.

UniProt Accession IDs. This manuscript describes functional characterization of proteins with the following UniProt accession IDs: Q6D0N7, A0A0H3LX82, Q8ZRS5, B0TBI9, Q0K4F6, Q6D0N8, A0A0H3LQK8, P58718, B0TBI8, Q0K4F5, Q6CZ26, A0A0H2VA68, P44979, Q0KBC7, P44094, Q0KBD2, Q6CZ24, Q57199, Q0KBC9, A0A0H2VA12, Q0KBD1, Q6CZ23, Q57151, Q0KBC8, Q6CZ25, P44093, Q4KBD3, B1M1V6, A0A0H3KP73, A6VKK5, Q8YB10, A7JVV9, and Q48PB0 (*SI Appendix*, Table S11).

Materials and Methods

Enzyme assays were performed with a UV-Visible spectrophotometer (Varian CARY 300 Bio). The consumption or formation of NADH was monitored as the decrease or increase in the absorbance at 340 nm with an extinction coefficient (ε) of 6,220 $M^{-1}{\cdot}cm^{-1}$. Locus tags and Uniprot identifiers for each gene and protein in this study can be found in *SI Appendix*, Table S11.

ATP-Dependent Kinase Activity Screening with a Focused Sugar Library. ATP-Dependent kinase activity was determined spectrophotometrically by following consumption of NADH. The formation of ADP was coupled to the oxidation of NADH via pyruvate kinase (PK) and lactate dehydrogenase (LDH) in the presence of ATP, phosphoenolpyruvate (PEP), and NADH. The assay (200 μL) contained 1 μM of purified enzyme, 0.1 M Tris·HCl buffer (pH 8.0), 10 mM KCl, 10 mM $MgCl_2$, 2.5 mM ATP, 2.5 mM PEP, 0.3 mM NADH, 5 U PK/LDH from rabbit muscle (Sigma), and 1 mM substrate. The reactions were performed in Corning 96-well clear flat-bottom UV-transparent microplates; the absorbance readings at 340 nm were measured with a TECAN microplate reader after incubating the reaction solution at 25 °C for 10 min, 2 h, and 16 h. The results of different time points allow estimations of relative activities (*SI Appendix*, Table S1).

Kinetic Assay for PdxA2–DUF1537 Protein Kinase Activity with D-Threonate, D-Erythronate, or 4-HT. ATP-dependent kinase activities of PdxA2–DUF1537 proteins were assayed by measuring the consumption of NADH (see above for details). The reaction mixture (25 °C) contained variable concentrations of substrate, 100 mM Tris·HCl buffer (pH 8.0), 10 mM KCl, 10 mM $MgCl_2$, 2.5 mM ATP, 2.5 mM PEP, 0.16 mM NADH, 5 U PK/LDH from rabbit muscle (Sigma), and enzyme in a final volume of 200 μL. Data were fit to the Michaelis–Menten equation (Table 1 and *SI Appendix*, Table S2).

Kinetic Assay for PdxA2 Oxidative Activity with D-Threonate 4-Phosphate, D-Erythronate 4-Phosphate, and 4-HT 4-Phosphate. Oxidation activities of PdxA2 proteins were assayed by measuring the formation of NADH. The reaction mixture (25 °C) contained variable concentrations of substrate, 50 mM Tris·HCl buffer (pH 8.0), 1.5 mM NAD^+, and enzyme in a final volume of 200 μL. Data were fit to the Michaelis–Menten equation (Table 1 and *SI Appendix*, Table S3).

Dehydrogenase Activity Screening with a Focused Sugar Library. Purified dehydrogenases were screened for oxidation activity using a library of 53 sugars (*SI Appendix*, Table S6). Activity was assayed by measuring the formation of NADH. The assay (200 μL) contained 1.5 μM of purified enzyme, 50 mM Tris·HCl buffer (pH 8.0), 1.5 mM NAD^+, and 1 mM substrate. The reactions were performed in Corning 96-well clear flat-bottom UV-transparent microplates, and the absorbance readings at 340 nm were measured with a TECAN microplate reader after incubating the reaction solution at 25 °C for 10 min, 2 h, and 16 h.

Kinetic Assay for Dehydrogenase Activity with L-Threonate or D-Erythronate. Dehydrogenases were assayed by measuring the formation of formazan by the increase in absorbance at 500 nm. Briefly, NAD(P)H was generated in the oxidation of L-threonate or D-erythronate. This reaction was coupled to diaphorase, which used the NAD(P)H to catalyze reduction of *p*-iodonitrotetrazolium violet (INT) into formazan (molar extinction coefficient 12,990 $M^{-1}{\cdot}cm^{-1}$). The reaction mixture (25 °C) contained variable concentrations of substrate, 50 mM Tris·HCl buffer (pH 8.5), 1.5 mM $NAD(P)^+$, 0.64 mM INT (Sigma), 2 unit diaphorase from *Clostridium kluyveri* (Sigma) and enzyme in a final volume of 200 μL. Data were fit to the Michaelis–Menten equation (*SI Appendix*, Table S4).

Bacterial Strains and Growth Conditions. *R. eutropha* H16 (DSM-428) and its derived strains were grown shaking aerobically in nutrient broth (Difco) or in defined media [per liter: 50 mL 20× salts (20 g NH_4Cl, 6 g $MgSO_4{\cdot}7H_2O$, 3 g KCl, 0.1 g $CaCl_2{\cdot}2H_2O$, 0.05 g $FeSO_4{\cdot}7H_2O$), 50 mL 20× phosphate buffer pH 7.0 (500 mM KH_2PO_4 mixed with 500 mM Na_2HPO_4 until pH 7.0), 1 mL 1,000× trace element solution (per liter: 0.5 g NaEDTA, 0.3 g $FeSO_4{\cdot}7H_2O$, 3 mg $MnCl_2{\cdot}4H_2O$, 5 mg $CoCl_2{\cdot}6H_2O$, 1 mg $CuCl_2{\cdot}2H_2O$, 2 mg $NiCl_2{\cdot}6H_2O$, 3 mg $Na_2MoO_4{\cdot}2H_2O$, 5 mg $ZnSO_4{\cdot}7H_2O$, 2 mg H_3BO_3), 1 mL 1,000× vitamin solution (per liter: 0.1 g cyanocobalamin, 0.3 g pyridoxamine-2 HCL, 0.1 g Ca-D-pantothenate, 0.2 g thiamine dichloride, 0.2 g nicotinic acid, 0.08 g 4-aminobenzoic acid, 0.02 g D-biotin), and 900 mL H_2O] with 140 μg/mL kanamycin as necessary. *S. enterica* serovar Typhimurium LT2 and its derivatives were grown shaking in LB or in defined media at 37 °C with 50 μg/mL kanamycin as necessary. *E. coli* strains DH5α,S17-1 (43) and BL21(DE3) were grown shaking aerobically in Luria-Bertani (LB) broth at 37 °C with 100 μg/mL ampicillin or 50 μg/mL kanamycin as necessary. *P. carotovorum* WPP14 and its derived strains (16) were grown at 30 °C in LB or defined media with 25 μg/mL kanamycin and 34 μg/mL chloramphenicol as necessary. Plates were prepared with 1.8% (wt/vol) agar for rich medium and 2.5% (wt/vol) agar for defined medium. Defined medium was supplemented with 10 mM carbon source. Growth studies were performed as 300 μL cultures in a Bioscreen C instrument based on optical density (OD) at 600 nm. The inoculum was derived from mid exponential cultures in rich media; cells were first collected by centrifugation and then rinsed with defined medium once before inoculation.

Isolation and Complementation of Markerless Deletion Strains of *R. eutropha* and *S. enterica*. Deletion strains of *R. eutropha* were isolated as previously described (44) with modifications described below. Deletions removed in-frame portions of the coding region without introducing any exogenous sequence in the final strain. Briefly, 750–1,000 bp of regions flanking the targeted coding region were amplified by PCR (for primers, see *SI Appendix*, Table S13). Each fragment contained ≥27 bp of the coding region and preserved the reading frame of the coding region. Using Gibson cloning (New England Biolabs), the fragments were assembled and inserted into pK18mobsacB (45) digested with XbaI/BamHI. The plasmid was transferred into *R. eutropha* via conjugation with *E. coli* S17-1 transformed with the plasmid. Single cross-over strains were isolated as kanamycin-resistant on nutrient plates. Colonies were additionally streaked for isolation on selective

nutrient plates. Resulting single colonies were incubated in nonselective nutrient broth until OD ~0.8. A 100-μL aliquot of a 100-fold dilution of the culture was plated on nutrient plates containing 10% sucrose. Double crossover strains were identified as kanamycin-sensitive after colonies were patched on nutrient plates with and without kanamycin. Among the double crossover strains, wild-type revertants were separated from mutants by colony PCR using primers that directed amplification across the deleted region. Mutant genotypes were confirmed by sequencing the entire region of the genome that was available for recombination with the fragments contained on the corresponding plasmid (see *SI Appendix*, Table S13 for plasmids).

A modified version of the Datsenko and Wanner (46) protocol was used to delete regions of the *S. enterica* genome. Genes were inactivated by replacing all but the first and last 27 bp of the coding region with an 85-bp scar sequence. Briefly, each of three PCR products were generated: (*i*) arm1, ~1,000 bp upstream of the target gene; (*ii*) the 1.6-kb kanamycin-resistance cassette from pKD4 (46); and (*iii*) arm2, ~1,000 bp downstream of the target gene. Primer sequences are provided in *SI Appendix*, Table S13. The fragments were assembled by overlapping extension PCR (Ho Gene 1989) based on ~50 bp of shared sequence between adjacent fragments. Assembly primers for arm1 and the kanamycin-resistance cassette were appropriately chosen among those used to amplify the individual fragments. The final assembled ~3.6-kb fragment was assembled using appropriate primers and consisted of the kanamycin-resistance cassette flanked by arm1 and arm2. Electrocompetent *S. enterica* carrying the λ-Red helper plasmid pKD46 was transformed with 100 ng of the final assembled product according to the protocol of Datsenko and Wanner (46). Double cross-over strains were isolated as kanamycin-resistant and confirmed by genomic PCR. The kanamycin-resistance genes were eliminated using helper plasmid pCP20 encoding the FLP recombinase. Final strains were cured of all plasmids as described previously (46).

Complementation of the mutant strains relied on expressing the respective deleted gene from the *h16_A1563* promoter (279 bp) for *h16_A1557 – h16_A1562* and the *stm0161* promoter (234 bp) for *stm0162* and *stm0163*. Promoters and coding regions were amplified separately by PCR (see *SI Appendix*, Table S13 for primers). The products were fused and ligated into pBBR1MCS2 (47) (digested with EcoRI/XbaI) by Gibson ligation (New England Biolabs). Plasmids were transferred into *R. eutropha* strains by conjugation with S17-1 transformed with the corresponding plasmid. Plasmids were transferred into *S. enterica* strains by electroporation.

ACKNOWLEDGMENTS. This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract DE-AC02-06CH11357. Use of the Lilly Research Laboratories Collaborative Access Team (LRL-CAT) beamline at Sector 31 of the Advanced Photon Source was provided by Eli Lilly Company, which operates the facility. This research was supported by NIH U54GM093342.

1. Galperin MY, Koonin EV (2010) From complete genome sequence to 'complete' understanding? *Trends Biotechnol* 28(8):398–406.
2. Galperin MY (2001) Conserved 'hypothetical' proteins: New hints and new puzzles. *Comp Funct Genomics* 2(1):14–18.
3. Galperin MY, Koonin EV (2004) 'Conserved hypothetical' proteins: Prioritization of targets for experimental study. *Nucleic Acids Res* 32(18):5452–5463.
4. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* 333(4):863–882.
5. Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLOS Comput Biol* 5(12):e1000605.
6. Hsiao TL, Revelles O, Chen L, Sauer U, Vitkup D (2010) Automatic policing of biochemical annotations using genomic correlations. *Nat Chem Biol* 6(1):34–40.
7. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* 8(12):995–1005.
8. Finn RD, et al. (2014) Pfam: The protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230.
9. Bateman A, Coggill P, Finn RD (2010) DUFs: Families in search of function. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 66(Pt 10):1148–1152.
10. Goodacre NF, Gerloff DL, Uetz P (2013) Protein domains of unknown function are essential in bacteria. *MBio* 5(1):e00744-13.
11. Bastard K, et al. (2014) Revealing the hidden functional diversity of an enzyme family. *Nat Chem Biol* 10(1):42–49.
12. Zhang H, et al. (2014) The highly conserved domain of unknown function 1792 has a distinct glycosyltransferase fold. *Nat Commun* 5:4339.
13. Prakash A, Yogeeshwari S, Sircar S, Agrawal S (2011) Protein domain of unknown function 3233 is a translocation domain of autotransporter secretory mechanism in gamma proteobacteria. *PLoS One* 6(11):e25570.
14. Arnold R, Jehl A, Rattei T (2010) Targeting effectors: The molecular recognition of type III secreted proteins. *Microbes Infect* 12(5):346–358.
15. Xiao Y, Heu S, Yi J, Lu Y, Hutcheson SW (1994) Identification of a putative alternate sigma factor and characterization of a multicomponent regulatory cascade controlling the expression of *Pseudomonas syringae* pv. syringae Pss61 hrp and hrmA genes. *J Bacteriol* 176(4):1025–1036.
16. Mole B, Habibi S, Dangl JL, Grant SR (2010) Gluconate metabolism is required for virulence of the soft-rot pathogen *Pectobacterium carotovorum*. *Mol Plant Microbe Interact* 23(10):1335–1344.
17. Su YC, Wan KL, Mohamed R, Nathan S (2008) A genome level survey of *Burkholderia pseudomallei* immunome expressed during human infection. *Microbes Infect* 10(12-13):1335–1345.
18. Vetting MW, et al. (2015) Experimental strategies for functional annotation and metabolism discovery: Targeted screening of solute binding proteins and unbiased panning of metabolomes. *Biochemistry* 54(3):909–931.
19. Mulligan C, Fischer M, Thomas GH (2011) Tripartite ATP-independent periplasmic (TRAP) transporters in bacteria and archaea. *FEMS Microbiol Rev* 35(1):68–86.
20. Higgins CF (2001) ABC transporters: Physiology, structure and mechanism–an overview. *Res Microbiol* 152(3-4):205–210.
21. Zhao S, et al. (2014) Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife* 3:1–32.
22. Gerlt JA, et al. (2015) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim Biophys Acta* 1854(8):1019–1037.
23. Yanai I, Derti A, DeLisi C (2001) Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci USA* 98(14):7940–7945.
24. Tabita FR, et al. (2007) Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. *Microbiol Mol Biol Rev* 71(4):576–599.
25. Sauder MJ, et al. (2008) High throughput protein production and crystallization at NYSGXRC. *Methods Mol Biol* 426:561–575.
26. Cane DE, Hsiung YJ, Cornish JA, Robinson JK, Spenser ID (1998) Biosynthesis of vitamin B-6: The oxidation of 4-(phosphohydroxy)-L-threonine by PdxA. *J Am Chem Soc* 120(8):1936–1937.
27. Sivaraman J, et al. (2003) Crystal structure of *Escherichia coli* PdxA, an enzyme involved in the pyridoxal phosphate biosynthesis pathway. *J Biol Chem* 278(44):43682–43690.
28. Wolf E, Spenser ID (1995) [2,3-C-13(2)]-4-hydroxy-L-threonine. *J Org Chem* 60(21):6937–6940.
29. Joerger AC, Gosse C, Fessner WD, Schulz GE (2000) Catalytic action of fuculose 1-phosphate aldolase (class II) as derived from structure-directed mutagenesis. *Biochemistry* 39(20):6033–6041.
30. Thoden JB, Wohlers TM, Fridovich-Keil JL, Holden HM (2001) Human UDP-galactose 4-epimerase. Accommodation of UDP-N-acetylglucosamine within the active site. *J Biol Chem* 276(18):15131–15136.
31. Thoden JB, et al. (1997) Structural analysis of UDP-sugar binding to UDP-galactose 4-epimerase from *Escherichia coli*. *Biochemistry* 36(21):6294–6304.
32. Ashiuchi M, Misono H (1999) Biochemical evidence that *Escherichia coli* hyi (orf b0508, gip) gene encodes hydroxypyruvate isomerase. *Biochim Biophys Acta* 1435(1-2):153–159.
33. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2256–2268.
34. Hayward S, Berendsen HJ (1998) Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. *Proteins* 30(2):144–154.
35. Kreimeyer A, et al. (2007) Identification of the last unknown genes in the fermentation pathway of lysine. *J Biol Chem* 282(10):7191–7197.
36. Bellinzoni M, et al. (2011) 3-Keto-5-aminohexanoate cleavage enzyme: a common fold for an uncommon Claisen-type condensation. *J Biol Chem* 286(31):27399–27405.
37. Huang H, et al. (2015) A general strategy for the discovery of metabolic pathways: d-threitol, l-threitol, and erythritol utilization in *Mycobacterium smegmatis*. *J Am Chem Soc* 137(46):14570–14573.
38. Wichelecki DJ, et al. (2015) ATP-binding cassette (ABC) transport system solute-binding protein-guided identification of novel d-altritol and galactitol catabolic pathways in *Agrobacterium tumefaciens* C58. *J Biol Chem* 290(48):28963–28976.
39. Smirnoff N (2011) Vitamin C: The metabolism and functions of ascorbic acid in plants. *Adv Bot Res* 59:107–177.
40. Green MA, Fry SC (2005) Vitamin C degradation in plant cells via enzymatic hydrolysis of 4-O-oxalyl-L-threonate. *Nature* 433(7021):83–87.
41. Loewus FA (1999) Biosynthesis and metabolism of ascorbic acid in plants and of analogs of ascorbic acid in fungi. *Phytochemistry* 52(2):193–210.
42. Williams M, Loewus FA (1978) Biosynthesis of (+)-tartaric acid from l-[4-C]ascorbic acid in grape and geranium. *Plant Physiol* 61(4):672–674.
43. Au AC, et al. (1983) Study of acute trichinosis in Ghurkas: Specificity and sensitivity of enzyme-linked immunosorbent assays for IgM and IgE antibodies to *Trichinella* larval antigens in diagnosis. *Trans R Soc Trop Med Hyg* 77(3):412–415.
44. Carter MS, Alber BE (2015) Transcriptional regulation by the short-chain fatty acyl coenzyme A regulator (ScfR) PccR controls propionyl coenzyme A assimilation by *Rhodobacter sphaeroides*. *J Bacteriol* 197(19):3048–3056.
45. Schäfer A, et al. (1994) Small mobilizable multi-purpose cloning vectors derived from the *Escherichia coli* plasmids pK18 and pK19: Selection of defined deletions in the chromosome of *Corynebacterium glutamicum*. *Gene* 145(1):69–73.
46. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA* 97(12):6640–6645.
47. Kovach ME, et al. (1995) Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes. *Gene* 166(1):175–176.