Published in final edited form as:

Nat Chem Biol. 2018 July ; 14(7): 696–705. doi:10.1038/s41589-018-0067-7.

Functional assignment of multiple catabolic pathways for D-apiose

Michael S. Carter^{§,1}, Xinshuai Zhang^{§,1}, Hua Huang^{§,1}, Jason T. Bouvier^{†,§}, Brian San Francisco[§], Matthew W. Vetting^{||}, Nawar Al-Obaidi^{||}, Jeffrey B. Bonanno^{||}, Agnidipta Ghosh^{||}, Rémi G. Zallot[§], Harvey M. Andersen[§], Steven C. Almo^{||}, and John A. Gerlt^{†,‡,§}

[†]Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States

[‡]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States

[§]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States

^{||}Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, United States

Abstract

Colocation of the genes encoding ABC, TRAP, and TCT transport systems and catabolic pathways for the transported ligand provides a strategy for discovering novel microbial enzymes and pathways. We screened solute binding proteins (SBPs) for ABC transport systems and identified three that bind D-apiose, a branched pentose in the cell walls of higher plants. Guided by sequence similarity networks (SSNs) and genome neighborhood networks (GNNs), the identities of the SBPs enabled the discovery of four catabolic pathways for D-apiose with eleven previously unknown reactions. The new enzymes include D-apionate oxidoisomerase that catalyzes hydroxymethyl group migration as well as 3-oxo-isoapionate 4-phosphate decarboxylase and 3-oxo-isoapionate 4-phosphate transcarboxylase/hydrolase that are RuBisCO-like proteins (RLPs). The web tools for generating SSNs and GNNs are publicly accessible (<http://efi.igb.illinois.edu/efi-est/>), so similar “genomic enzymology” strategies for discovering novel pathways can be used by the community.

Correspondence should be addressed to John A. Gerlt (j-gerlt@illinois.edu).

[†]Authors contributed equally to this work

Author Contributions

M.S.C., X.Z., H.H., M.W.V., S.C.A., and J.A.G. conceived the project. J.T.B., M.W.V., and J.A.G. developed the library for SBP ligand screening. M.W.V., N.A., A.G., J.B.B., and S.C.A. managed the protein purification pipeline for the SBPs and some pathway enzymes. X.Z. and H.H. contributed purification of the remaining pathway enzymes, biochemical characterization of all enzymes, and chemical validation of their substrates and products. M.W.V. contributed DSF screening of SBPs against the ligand library. M.S.C. identified D-apiose as the physiological ligand for the SBPs. M.W.V., J.B. B., and S.C.A. contributed crystallization data and analysis. M.S.C., X.Z., H.H., B.S.F. and J.A.G. evaluated SSNs, GNNs, biochemical, and biological data to hypothesize pathways. M.S.C., H.H., and R.Z. contributed biological validation of pathways. H.A. contributed molecular cloning. M.S.C., X.Z., and J.A.G. wrote the paper with contributions from all authors.

Financial Interests

No competing financial interests

Introduction

The UniProt database (<http://www.uniprot.org/>; 99,261,416 entries in Release 2017_11; increasing with a doubling time of ~2.5 years) is partitioned between proteins with functional annotations assigned by homology (TrEMBL; 98,705,220 entries) and human curation (SwissProt; 556,196 entries). Assignment of functions to uncharacterized proteins is difficult because the sequence boundaries between functionally distinct groups of orthologues are difficult to determine. Therefore, many proteins have incorrect, uncertain, or unknown functions. The challenge is to devise robust approaches for assignment of the *in vitro* activities and *in vivo* metabolic functions to uncharacterized proteins.

Multiple approaches have been used for assignment of functions to uncharacterized enzymes^{1–4}. Bioinformatic approaches, based only on sequence homology, are problematic⁵. Over-annotation, e.g., assignment of the function to homologues because they are members of a protein family, is a widely recognized problem in the automated annotation of proteomes encoded by sequenced genomes⁶. Analyses of sequence-function space in families is a valuable initial step even if uncharacterized homologues do not catalyze the same reaction using the same substrate—they often catalyze mechanistically related reactions using substrates with conserved reactive groups, e.g., abstraction of the α -proton of a carboxylate group in the enolase superfamily^{7,8}. The conserved mechanistic features can provide important clues about the reactions catalyzed by homologues. However, identification of the substrate for and reaction catalyzed by the uncharacterized homologues requires additional information.

We are using a “genomic enzymology” strategy^{9,10} to discover functions of microbial proteins/enzymes that constitute the majority of the UniProt database (67,155,056 bacterial, 2,310,947 archaeal, and 7,485,472 fungal proteins in Release 2017_11). We use the ligand specificities of the solute binding proteins (SBPs) of microbial transport systems to guide discovery of catabolic pathways for the ligands—the ligand identifies the substrate for the pathway, as the genes that encode the transport systems and pathways often are colocated. In this strategy we use sequence similarity networks (SSNs^{10–12}) of protein families to identify isofunctional (orthologous) families/clusters of SBPs and pathway enzymes and genome neighborhood networks (GNNs^{10,13}) of the genes encoding these clusters to identify the pathways. We provide two community-accessible “genomic enzymology” web tools to enable broad use of this strategy: EFI-EST (<http://efi.igb.illinois.edu/efi-est/>) for generating SSNs^{10,12} and EFI-GNT (<http://efi.igb.illinois.edu/efi-gnt/>) for generating GNNs¹⁰.

We discovered three SBPs for ABC transport systems (from Pfam family PF13407) that bind D-apiose. D-Apiose is a branched pentose (Fig. 1a) found in rhamnogalacturonan-II (RG-II) in the cell walls of higher plants and apigalacturonan in the cell walls of aquatic monocots, e.g., *Zostera* and *Lemna*¹⁴ (Fig. 1b). D-Apiose increases cell wall stability by participating in interstrand RG-II crosslinking *via* boronate esters. The biosynthetic pathway for D-apiose is known^{15–18}; UDP-D-apiose/UDP-D-xylose synthase that generates the branched structure has been mechanistically characterized. Polysaccharide utilization loci (PULs) that degrade RG-II to monosaccharides have been identified in species of *Bacteroides* in the human gut

microbiome^{19,20} (Supplementary Fig. 1); *Bacteroides vulgatus* and *Bacteroides dorei* utilize D-apiose establishing the existence of uncharacterized D-apiose catabolic pathways.

The D-apiose-binding SBPs enabled discovery of a catabolic pathway involving a transketolase that converts D-apiose to intermediates in central carbon metabolism; this pathway is present in *B. vulgatus* and *B. dorei*. We also discovered three pathways that diverge from 3-oxo-isoapionate, a β -ketoacid, generated from D-apionate by a novel D-apionate oxidoisomerase: two pathways involve decarboxylases from different enzyme families, the third involves a transcarboxylase/hydrolase that shares mechanistic features with RuBisCO. This study provides a compelling example of how SSNs and GNNs can be used to discover novel enzymes in novel metabolic pathways.

Results

Strategy for functional annotation.

Our functional assignment strategy is based on 1) identifying the ligand for a SBP of an ABC, TRAP, or TCT transport system (Fig. 1c), 2) discovering the ligand's catabolic pathway by using sequence similarity networks (SSNs) (Fig. 1d) and genome neighborhood networks (GNNs) (Fig. 1e) to locate the genome neighborhoods that encode the pathway enzymes (Fig. 1f).

The number and types of pathways that can be discovered is determined by the sizes/diversities of the libraries of 1) genomic DNAs (gDNA) that encode SBPs and 2) small molecules for screening their ligand specificities. Our gDNA library (408 species) was assembled from the ATCC and DSMZ culture collections. Our ligand library (405 compounds)²¹ includes all D- and L-tetroses, pentoses, and hexoses, their aldonic and aldaric acids, all D- and L-hexuronic acids, and other commercially available aldoses, including D-apiose, ketoses and oligosaccharides. We previously used these libraries to discover catabolic pathways for tetrityls²², hexitols²¹, and tetronic acids²³.

We identified D-apiose as the ligand for three ABC SBPs from Pfam family PF13407 (specific for carbohydrates; 53,665 sequences in UniProt release 2017_07). Using SSNs and GNNs, the SBPs leveraged the discovery of four catabolic pathways for D-apiose that are subsequently outlined in Figures 2–5 in >1350 microbial species.

Identification of SBPs for D-apiose.

We chose 414 targets to sample sequence-function space in PF13407; the SSN filtered with an alignment score of 115 (~63% pairwise sequence identity) is shown in Supplementary Figure 2. We could clone the genes encoding 73% (303) of the targets; the SBPs encoded by 39% of the genes (117; highlighted in blue, magenta, or red in the SSN) could be purified and were subjected to ligand screening. The identities of the screened SBPs are provided in Supplementary Data Set 1.

Ligand hits (>5 °C ligand-induced stabilization) were identified for 89 SBPs (highlighted in magenta or red), including three that bound D-apiose, D-ribose, and D-ribulose (highlighted in red; located in a single cluster): UniProt ID A6VKQ8 from *Actinobacillus succinogenes*

ATCC 55618, B1G898 from *Burkholderia graminis* C4D1M, and Q2JZQ5 from *Rhizobium etli* CFN42) (Supplementary Table 1).

We determined that D-apiose is the physiological ligand for the SBPs. The SBP from *Rhizobium etli* CFN42 (UniProt ID Q2JZQ5) was co-crystallized with D-ribose and D-apiose (Supplementary Table 2); however, hydrogen-bonding patterns were sufficiently similar that neither could be inferred as the physiological ligand (Supplementary Fig. 3). We then disrupted the gene in *Agrobacterium radiobacter* K84 that encodes an SBP (UniProt ID B9JK76) sharing 92% sequence identity with Q2JZQ5; we also disrupted four proximal genes encoding putative pathway enzymes. For all mutants, growth with D-ribose and D-ribulose as carbon source was unaffected; however, growth with D-apiose was reduced or abolished (Supplementary Fig. 4). The wild type phenotype was restored by complementation. Because all three SBPs are located in the same SSN cluster (>63% pairwise sequence identity), we assumed that they and the other sequences in the cluster (total of 356) are functional orthologues.

A flowchart describing how SSNs and GNNs were used to leverage the use of these SBPs to discover the catabolic pathways for D-apiose is provided in Supplementary Fig. 5. The pathways are shown in Figures 2a, 3a, 4a and 5a; the pathway enzymes that were characterized are encoded by the genome neighborhoods shown in Figures 2b, 3b, 4b, and 5b.

Two types of catabolic pathways.

The single SSN cluster containing the D-apiose-binding SBPs (alignment score 115, >63% pairwise sequence identity; Supplementary Fig. 6) was used to generate a GNN (Supplementary Fig. 7a) that identified several Pfam-curated enzyme families (spoke-nodes): two families of kinases, two families of dehydrogenases, the fucose isomerase family (aldose/ketose isomerase), two families of transketolase domains, and the RuBisCO superfamily. This complexity suggested multiple pathways, i.e., the SBPs in the SSN cluster participate in more than one catabolic pathway.

Using an increased alignment score (130; ~72% sequence identity), two major SSN clusters were obtained (Supplementary Fig. 6c and 6d). The GNN (Supplementary Fig. 7b) identified two genome neighborhoods associated with two types of catabolic pathways:

1. Transketolase pathway. The blue SSN/GNN cluster identified three enzyme families: a) *aldose/ketose isomerase* (PF02952 or PF01261), b) bidomain *kinase* (PF00370-PF02782), and c) heterodimeric *transketolase* (PF00456 and PF02779-PF02780). These participate in the pathway in Figure 2a.
2. Oxidative pathways. The red SSN/GNN cluster identified three types of enzyme families: a) two *dehydrogenase* families (one with a PF01408 domain and one with a PF16896 domain), b) three *kinase* families (PF07005-PF17042, PF025733-PF02734, and PF00370-PF02782), and c) *RuBisCO-like proteins* (*RLPs*) from the RuBisCO superfamily (PF00016-PF02788). These participate in pathways initiated by oxidation of D-apiose (Fig. 4a) or oxidation of D-apionate (Figs. 3a and 5a).

We subjected the proteins encoded by the genome neighborhoods in Figures 2b, 3b, 4b, and 5b to *in vitro* enzymatic assays and *in vivo* genetic analyses (knockouts) so that the pathways could be established.

Transketolase pathway characterization.

Proteins from *Pectobacterium atrosepticum* SCRI 1043 and *Actinobacillus succinogenes* ATCC 55618 (genome neighborhoods in Fig. 2b) were assayed (a complete set of pathway enzymes could not be successfully purified from either organism). The aldose/ketose isomerase (Q6D5T7 from *P. atrosepticum* SCRI 1043; PF01261) catalyzes isomerization of D-apiose to D-apulose (Supplementary Fig. 8). The kinase (Q6D5T8 from *P. atrosepticum* SCRI 1043) catalyzes phosphorylation of D-apulose to form D-apulose 4-phosphate (Supplementary Fig. 8). And, the heterodimeric transketolase (A6VKQ3 and A6VKQ4 from *A. succinogenes* ATCC 55618) catalyzes transfer of the glycolaldehyde group from D-apulose 4-phosphate to D-glyceraldehyde 3-phosphate, generating dihydroxyacetone phosphate and D-xylulose 5-phosphate (Fig. 2c and Supplementary Fig. 9; Supplementary Table 3).

Pectobacterium carotovorum WPP14 that also utilizes D-apiose and contains the same genome neighborhood found in *P. atrosepticum* SCRI 1043 was used for *in vivo* studies because it has a faster doubling time. Mutant strains were generated in which the genes encoding the pathway enzymes were disrupted; the mutants were unable to utilize D-apiose. When complemented, they recovered wild type growth (Fig. 2d and Supplementary Fig. 10). Taken together, the *in vitro* enzymatic and *in vivo* phenotypic characterizations establish the pathway in Figure 2a.

Transketolase pathway in gut microbiome species.

Not all of the transketolase polypeptides (from PF00456 (9,813 sequences) and PF02779-PF02780 (31,525 sequences); Supplementary Figs. 11 and 12) identified by the D-apiose-binding SBPs are encoded by genome neighborhoods that include ABC transport systems. These include transketolases encoded by species of Bacteroides and Clostridia from the human gut microbiome; however, their genome neighborhoods encode orthologues of both D-apiose isomerase and D-apulose kinase, e.g., *B. vulgatus* ATCC 8482 and *Blautia hydrogenotrophica* DSM 10507 (genome neighborhoods in Fig. 2b). The study characterizing the PULs for RG-II degradation reported that *B. vulgatus* ATCC 8482 is able to utilize D-apiose. Although we were not able to purify the isomerase and kinase from *B. vulgatus*, we determined that its transketolase catalyzes the conversion of D-apulose 4-phosphate and glyceraldehyde 3-phosphate to dihydroxyacetone phosphate and D-xylulose 5-phosphate (Fig. 2a; Supplementary Fig. 9). We also observed D-apiose-dependent growth and D-apiose dependent up-regulation of transcripts for the isomerase, kinase, and transketolase (Supplementary Fig. 13).

The conservation of the isomerases and the sequence similarities of the kinases encoded by *B. vulgatus*, *P. carotovorum*, *P. atrosepticum*, and *A. succinogenes* together with our experimental data for *B. vulgatus* provide persuasive evidence that *B. vulgatus* uses the pathway in Figure 2a. The study characterizing the PULs for RG-II degradation²⁰ reported

that 1) *B. dorei* DSM 17855 also is able to utilize D-apiose but 2) fourteen other species of Bacteroides are not. Of these sixteen species, only *B. vulgatus* and *B. dorei* encode isomerase, kinase, and transketolase orthologues, thereby providing additional evidence for the pathway in Figure 2a and its presence in human gut microbiota.

We could not purify a Clostridial transketolase; however, genes predicted to encode orthologues are colocated with those encoding kinases as well as isomerases from the L-arabinose isomerase family (PF02610), e.g., the genome neighborhood of *B. hydrogenotrophica* DSM 10507 in Figure 2b. We propose that species of Clostridia in the human gut microbiome also use the pathway in Figure 2a to catabolize D-apiose released from RG-II by the community.

From the number of proteins in the SSN clusters for the transketolase polypeptides, the pathway in Figure 2a is encoded by 860 bacterial species (UniProt Release 2017_07; July 2017; a spreadsheet listing these species is included in Supplementary Data Set 2), including species from the human gut microbiome and soil microbiomes. Both communities are expected to have access to plant cell walls.

Dehydrogenases in oxidative pathways.

We next characterized the oxidative pathways that are encoded by the genome neighborhoods in Figures 3b, 4b, and 5b. The catabolism of D-apiose requires both dehydrogenases identified in the GNN for the SBPs (Fig. 4a), one with a PF01408 domain and one with a PF16896 domain; the catabolism of D-apionate requires only the dehydrogenase with a PF16896 domain (Figs. 3a and 5a).

Both dehydrogenases from *A. radiobacter* K84 were assayed (genome neighborhood in Fig. 4B). The dehydrogenase with a PF01408 domain (B9JK80) catalyzed oxidation of D-apiose to D-apionolactone using NAD⁺ (¹H NMR spectrum in Supplementary Fig. 14; kinetic constants in Supplementary Table 4).

Catabolism of D-apionolactone should require its hydrolysis to D-apionate (Fig. 4a). The genome neighborhoods that encode D-apiose dehydrogenase include a “hypothetical” protein (e.g., Fig. 3b). The “hypothetical” protein from *A. radiobacter* K84 could not be purified, so we purified an orthologue from the same genome neighborhood in *Ochrobactrum anthropi* ATCC 49188 (A6X3G3); it hydrolyzed D-apionolactone to D-apionate (¹H NMR spectrum in Supplementary Fig. 14). The D-apionolactonases are members of a family not yet curated by Pfam (Supplementary Fig. 15).

When D-apionate was incubated with NAD⁺ and the second dehydrogenase with a domain from PF16896 (UniProt ID B9JK75), NADH was generated. D-Apionate was converted to “3-oxo-isoapionate” (Fig. 4a; a β -ketoacid) by oxidation of the 2-OH group and migration of a hydroxymethyl group (¹H NMR spectrum in Supplementary Fig. 16; kinetic constants in Supplementary Table 4); we designate this protein “D-apionate oxidoisomerase”. The reaction is reminiscent of reductoisomerase reactions in branched chain amino acid synthesis^{24–26}. The hydroxymethyl group migration may be either concerted or stepwise *via* formation of formaldehyde and an enolate anion intermediate (Fig. 6a).

D-Apionate oxidoisomerases are present in the catabolic pathways in Figures 3a, 4a, and 5a (genome neighborhoods in Figs. 3b, 4b, and 5b). The following sections describe identification of the downstream enzymes in the pathways.

Downstream enzymes in oxidative pathways.

Because the GNNs for the D-apiose-binding SBPs do not identify the complete oxidative pathways (Fig. 3a, 4a, and 5a; Supplementary Fig. 7), we used the SSN and GNN for D-apionate oxidoisomerase, (PF16896; 500 sequences in UniProt release 2017_07) to identify the remaining pathway enzymes. With an alignment score of 85 (~50% pairwise sequence identity), the SSN contains two major clusters (Supplementary Fig. 17), the larger including the characterized D-apionate oxidoisomerase from *A. radiobacter* K84 (Fig. 4b). Small clusters also are present, including one containing proteins from Clostridia in the human gut microbiome. Although the D-apiose-binding SBPs are genome neighbors of only the largest oxidoisomerase cluster (red nodes in Supplementary Fig. 17c), we hypothesized that the family is isofunctional.

Pathway with a decarboxylase from the xylose isomerase family.

The (smaller, not proximal to D-apiose-binding SBPs) blue SSN/GNN cluster (SSN in Supplementary Fig. 17d; GNN in Supplementary Fig. 18) identifies four enzyme families: a) a *decarboxylase* (PF01261); b) a bidomain *kinase* (PF02733-PF02734); and c) two *isomerases* (PF00121 (triose phosphate isomerase family) and PF02502 (LacAB_rpiB family)). (Members of PF13561 (*adh_short_C2* family) are identified, but these are distal from the oxidoisomerases and functionally irrelevant.) We previously established that orthologues of the isomerases catalyze conversion of L-erythrulose 1-phosphate to D-erythrulose 4-phosphate and of D-erythrulose 4-phosphate to D-erythrose 4-phosphate, respectively²².

A decarboxylation is required to convert 3-oxo-isoapionate to D-erythrose 4-phosphate, the expected product of the pathway. Therefore, 1) L-erythrulose (or L-erythrulose 1-phosphate) is generated from 3-oxo-isoapionate (or 3-oxo-isoapionate 4-phosphate) and 2) the member of PF01261 is a β -ketoacid decarboxylase (SwissProt-curated enzymes catalyze reactions with enolate anion intermediates). The pathway with L-erythrulose is shown in Fig. 4a.

P. atrosepticum SCRI 1043 utilizes D-apionate (Supplementary Fig. 19). The decarboxylase (Q6D8V4), kinase (Q6D8V6), and both isomerases [Q6D8V5 (TIM family) and Q6D8V9 (LacAB_rpiB) family] were assayed (genome neighborhood in Fig. 4B). Q6D8V4 catalyzed decarboxylation of 3-oxo-isoapionate to L-erythrulose (Fig. 3b and Supplementary Fig. 20), Q6D8V6 catalyzed phosphorylation of L-erythrulose to L-erythrulose 1-phosphate, and Q6D8V5 and Q6D8V9 catalyzed the expected isomerizations of L-erythrulose 1-phosphate to D-erythrulose 4-phosphate and of D-erythrulose 4-phosphate to D-erythrose 4-phosphate, respectively (Supplementary Fig. 21, kinetic constants in Supplementary Tables 4 and 5). We designate the decarboxylase “3-oxo-isoapionate decarboxylase”.

P. carotovorum WPP14 that contains the same genome neighborhood was used for *in vivo* studies. Mutant strains were generated that separately disrupted the gene for each enzyme. They were unable to use D-apionate; complemented strains recovered wild type growth (Fig.

3d and Supplementary Fig. 19). Taken together, the *in vitro* enzymatic and *in vivo* phenotypic characterizations confirm the pathway in Figure 4a.

P. carotovorum also utilizes D-apiose via the transketolase pathway (genome neighborhood in Fig. 2a); however, it does not encode D-apiose dehydrogenase or D-apionolactone hydrolase, which would be required to convert D-apiose to D-apionate. Therefore, D-apiose and D-apionate are independent carbon sources.

A SSN was generated for PF01261; the 3-oxo-isoapionate decarboxylase clusters (genome-proximal to PF16896) were identified (Supplementary Fig. 22). From the number of proteins in these clusters, the pathway in Figure 4a is encoded by 176 bacterial species (UniProt Release 2017_07; a spreadsheet listing these species is included in Supplementary Data Set 2), including species of Actinobacteria and Proteobacteria in the soil microbiome.

Two pathways involving RuBisCO-like proteins (RLPs). The GNN for the (larger, SBP proximal) red PF16896 SSN cluster (SSN in Supplementary Fig. 17; GNN in Supplementary Fig. 18) identified a *kinase* (PF07005-PF17042) and *RuBisCO-like proteins* (RLPs) from the RuBisCO superfamily (PF00016-PF02788).

Two genome contexts are identified (Fig. 4a and 5b), each with a kinase but different RLPs (the former containing D-apiose dehydrogenase and the D-apionolactonase). Because reactions catalyzed by RuBisCO²⁷ and the two previously characterized RLPs^{28–31} generate enolate anions from ketose 1-phosphate substrates, we predicted that the kinases phosphorylate 3-oxo-isoapionate to 3-oxo-isoapionate 4-phosphate that is the substrate for the RLPs (pathways in Fig. 4a and 4b).

Remaining enzymes in the pathways were identified using SSNs and GNNs for the RuBisCO superfamily (PF00016-PF02788). Filtered with an alignment score of 95 (Supplementary Fig. 23a), the SSN contains clusters for Form I (red), II (green), and III (blue) RuBisCOs as well as RLPs (Form IV RuBisCOs)^{32,33}. Members of two RLP clusters have been characterized, 2,3-diketo-5-methylthio-D-ribulose 1-phosphate tautomerase^{28,29} (magenta) and 5-methylthio-D-ribulose 1-phosphate isomerase^{30,31} (cyan). The genome neighbors of the oxidoisomerase are located in a third RLP cluster (yellow nodes); when this cluster is filtered with an alignment score of 140, several clusters are generated. The majority of the oxidoisomerase neighbors are located in two clusters (SSN in Supplementary Fig. 23b; GNN in Supplementary Fig. 24).

Pathway with an RLP decarboxylase.

The red RLP SSN/GNN cluster (Supplementary Figs. 23 and 24) identified the same isomerase families in the pathway that includes 3-oxo-isoapionate decarboxylase (Fig. 4a); however, the GNN identified a different kinase (PF07005-PF17042). We predicted a pathway (Fig. 3a) in which 3-oxo-isoapionate is phosphorylated to 3-oxo-isoapionate 4-phosphate prior to decarboxylation; erythrose 1-phosphate, the product of the RLP-catalyzed reaction, is converted to D-erythrose 4-phosphate by the isomerases.

Proteins from two genome neighborhoods from *A. radiobacter* K84 (Fig. 4b) were assayed. The first neighborhood encodes the oxidoisomerase (B9JK75), kinase (B9JK74), and RLP

(B9JK73); the second encodes L-erythrulose 1-phosphate isomerase (B9JN20) and D-erythrulose 4-phosphate isomerase (B9JN19) in a gene cluster that also participates in erythritol catabolism²². The expected intermediates were identified (¹H NMR spectra in Fig. 4b and Supplementary Figs. 25, 26, and 27); the orthologous kinase from *Burkholderia graminis* C4D1M (B1G889) was used because the *A. radiobacter* kinase (B9JK74) was insoluble. We designated the RLP “3-oxo-isoapionate 4-phosphate decarboxylase”.

Strains of *A. radiobacter* K84 were generated in which pathway genes were disrupted; the strains were unable to utilize D-apiose. The complemented strains recovered wild type growth (Fig. 4e and Supplementary Fig. 28). Taken together, the *in vitro* enzymatic and *in vivo* phenotypic characterizations confirm the pathway in Figure 4a.

From the number of proteins in the SSN cluster for 3-oxo-isoapionate 4-phosphate decarboxylase, the pathway in Figure 4a is encoded by 204 bacterial species (UniProt Release 2017_07; a spreadsheet listing these species is included in Supplementary Data Set 2), including species of Proteobacteria in the soil microbiome.

Pathway with an RLP transcarboxylase/hydrolase.

The blue RLP SSN/GNN cluster (Supplementary Fig. 24) identified D-apionate oxidoisomerase and the same kinase family (PF07005-PF17042) identified by the red RLP cluster (Members of PF13561 (adh_short_C2) are identified that are genome distal and functionally irrelevant.)

A gene cluster from *Ralstonia eutropha* N-1 (Fig. 5b) encodes the RLP, D-apionate oxidoisomerase, and the kinase. Both the kinase and RLP were insoluble when expressed, so the kinase (B1G889) in the RLP decarboxylase pathway (Fig. 4a) was used to generate 3-oxo-isopropionate 4-phosphate, and the orthologous RLP from *Xanthobacter autotrophicus* ATCC BAA-1158 (A7IJG7; Fig. 5b) was used. The RLP converted 3-oxo-isoapionate 4-phosphate to 3-phosphoglycerate (3-PGA) and glycolate (Fig. 5a; ¹H NMR spectrum in Fig. 5c and Supplementary Fig. 29). The reaction is initiated by decarboxylation to generate a stabilized enediolate intermediate, with the sequestered CO₂ carboxylating the adjacent enediolate carbon atom. The resulting 3-ketose 1-phosphate intermediate is hydrolyzed, as in the authentic RuBisCO-catalyzed reaction, to generate 3-PGA and glycolate. In support of this mechanism (Fig. 6b), [1-¹³C]-3-PGA was generated using 3-oxo-isoapionate 4-phosphate prepared from [1-¹³C]-(D/L)-apiose. Also, when the reaction was performed with unlabeled 3-oxo-isoapionate 4-phosphate but in the presence of [¹³C]-bicarbonate and carbonic anhydrase, no ¹³C could be detected in the 3-PGA by ¹H NMR spectroscopy, consistent with intramolecular carboxylate group transfer (Supplementary Fig. 30). We designate this RLP “3-oxo-isoapionate 4-phosphate transcarboxylase/hydrolase”. Disruption of the genes encoding the pathway in *R. eutropha* (Fig. 5b) resulted in strains that could not utilize D-apionate (Fig. 5d and Supplementary Fig. 31).

The genome of *R. eutropha* also encodes (in a distal gene cluster) both glycolate oxidase that catalyzes conversion of glycolate to glyoxylate and glyoxylate carboligase that catalyzes conversion of two molecules of glyoxylate to glycerate and CO₂. Disruption of the gene encoding the FAD-binding subunit of glycolate oxidase (G0EVT6) reduced the growth rate

and yield of cells with D-apionate as carbon source (Supplementary Fig. 31), consistent with retention of the ability to assimilate the 3-PGA product but not the glycolate product of the transcarboxylase/hydrolase. Taken together, the *in vitro* enzymatic and *in vivo* phenotypic characterizations confirm the pathway in Figure 5a.

From the number of proteins in the SSN cluster for the 3-oxo-isopropionate 4-phosphate transcarboxylase/hydrolase, the pathway in Figure 5a is encoded by 108 bacterial species (UniProt Release 2017_07; a spreadsheet listing these species is included in Supplementary Data Set 2), including species in the soil microbiome.

Pathway involving 3-oxo-isoapionate and a transketolase.

The oxidoisomerase SSN contains a small cluster with proteins from Clostridia from the human gut microbiome (orange cluster in Supplementary Fig. 18c and 18d). The GNN (Supplementary Fig. 6) identifies 1) a *heterodimeric transketolase* (PF00456 and PF02779-PF02780) and 2) a *dehydrogenase* (PF00389-PF02826). The transketolase polypeptides are in the same Pfam families as the transketolases in the pathway in Figure 2a; however, they are not in the same clusters (circled in orange in Supplementary Fig. 12). Thus, this transketolase, dehydrogenase, and D-apionate oxidoisomerase are predicted to participate in another catabolic pathway for D-apionate (Supplementary Fig. 32a).

The genome neighborhood selected for experimental characterization is shown in Supplementary Figure 32b. We assayed the oxidoisomerase, dehydrogenase, and hydroxypyruvate reductase from *B. hydrogenotrophica* DSM 10507; they catalyze the reactions shown in Supplementary Figure 32a. For the transketolase, D-glycerate is generated from 3-oxo-isoapioinate in the presence of D-glyceraldehyde 3-phosphate; however, we were unable to determine the identity of the second product. The kinetic constants are summarized in Supplementary Table 4; the ¹H NMR spectra showing the formation of D-glycerate by the transketolase are shown in Supplementary Figure 33. Growth of *B. hydrogenotrophica* with D-apionate and the genetics of the pathway were not investigated. Although not fully developed, the data provide evidence for this pathway in another species of the human gut microbiome.

From the number of proteins in SSN cluster for the orthologues of D-apionate oxidoisomerase family (PF16896) that participate in this pathway, we estimate that the pathway in Supplementary Figure 32a is encoded by 13 Clostridial species (UniProt Release 2017_07; a spreadsheet listing these species is included in Supplementary Data Set 2).

Discussion

Without targeting any specific environmental niche, our large-scale strategy enabled identification of pathways by which D-apiose (and/or D-apionate) are degraded by members of the human gut (species of Bacteroides and Clostridia) as well as soil (*A. radiobacter*, *P. carotovorum*/ *P. atrosepticum*, and *R. eutropha*) microbiomes. The specificities of three SBPs were sufficient to discover four pathways in >1350 microbial species and identify eleven new enzymatic functions. The identities of the pathways were not biased by a focus on specific organisms or environmental niches—the phylogenetic distribution of orthologues,

e.g., D-apionate oxidoisomerases, enabled discovery of multiple metabolic strategies, e.g., two pathways involving previously uncharacterized families of RLPs (Fig. 4a and Fig. 5a). The EFI-EST and EFI-GNT web tools then allowed identification of orthologous targets for functional characterization by providing facile access to conserved genome neighborhoods in different organisms.

The D-apionate oxidoisomerase catalyzes the previously unreported migration of a hydroxymethyl group (Fig. 2a and Fig. 6a). The members of this family (500 sequences; Supplementary Fig. 17) contain a conserved C-terminal substrate-binding domain ((PF16896; 6-phosphogluconate dehydrogenase (decarboxylating)) and an N-terminal NAD(P)-binding domain from one of several Pfam families, including PF07991 (acetohydroxy acid isomeroeductase, NADPH-binding) and PF03446 (6-phosphogluconate dehydrogenase, NADP-binding). The family appears to be isofunctional; however, only a fraction of the members of this family are encoded by the same genome neighborhoods that encode members of the D-apiiose-binding SBP clusters (compare SSNs in Supplementary Figs. 6 and 17).

We also discovered two novel functions for RLPs (Fig. 4a and 5a; Fig. 6b). The two previously characterized functions (2,3-diketo-5-methylthio-D-ribulose 1-phosphate tautomerase and 5-methylthio-D-ribulose 1-phosphate isomerase) are involved in pathways for methionine salvage; both catalyze enolization reactions. We identified 3-oxo-isoapionate 4-phosphate decarboxylase and 3-oxo-isoapionate 4-phosphate transcarboxylase/hydrolase in pathways for D-apionate catabolism. Both reactions catalyze decarboxylation of a β -ketoacid to generate a stabilized enediolate intermediate, the “reverse” of the carboxylation reaction catalyzed by RuBisCO. The active site of the decarboxylase protonates the intermediate to generate L-erythrulose 1-phosphate; the active site of the transcarboxylase/hydrolase uses the sequestered CO_2 to carboxylate the enediolate and hydrolyze the resulting β -ketoacid to glycolate and 3-PGA, partial reactions that are shared with RuBisCO. We expect that the “decarboxylase clade” of RLPs (Supplementary Fig. 23B) includes uncharacterized enzyme families that catalyze other decarboxylation/carboxylation reactions involving stabilized enediolate intermediates; the functions of these clades may provide useful insights into the evolution of the carboxylation function and the ability of members of this superfamily to distinguish between O_2 and CO_2 ^{34–36}.

From a biological perspective, our elucidation of these pathways provides evidence for the evolution of several diverse catabolic pathways for D-apiiose with multiple strategies for debranching D-apiiose. Three pathways exploit the oxidoisomerization of D-apionate to generate β -ketoacids that undergo decarboxylation reactions (Fig 3a, 4a, and 5a); two pathways, one characterized (Fig. 2a) and the second proposed (Supplementary Fig. 32a), use homologous but non-orthologous transketolases to accomplish “debranching”. The independent evolution of multiple pathways is intriguing and worthy of study; Supplementary Data Set 2 provides lists of the organisms in which these pathways were discovered to facilitate these efforts.

No single strategy can be expected to be successful for prediction and subsequent experimental assignment of functions to uncharacterized proteins discovered in genome

projects, e.g., not all proteins are enzymes in metabolic pathways nor are the genes that encode the components of metabolic pathways always genome proximal, so functional linkages that facilitate hypothesis generation and testing may not be readily apparent. However, we suggest that given appropriate diversity and size of ligand and gDNA libraries, this strategy could be broadly useful for additional SBPs as well as transcriptional regulators.

To enable the general use of this approach, we provide the EFI-EST and EFI-GNT web tools to members of the experimental community who are not experts in bioinformatics.

Uniprot Accession IDs

Functions have been assigned to proteins with following UniProt IDs (Supplementary Table 7): A6VKQ3/A6VKQ4 (D-apulose 4-phosphate transketolase), A6VKQ8 (D-apiose binding SBP), A6X3G3 (D-apionate lactonase), A7IJG7 (3-oxo-isoapionate 4-phosphate transcarboxylase), B1G889 (3-oxo-isoapionate kinase), B1G894 (D-apiose dehydrogenase), B1G898 (D-apiose binding SBP), B9JK73 (3-oxo-isoapionate 4-phosphate decarboxylase), B9JK75 (D-apionate oxidoisomerase), B9JK80 (D-apiose dehydrogenase), B9JN19 (D-erythrulose 4-phosphate isomerase), B9JN20 (L-erythrulose 1-phosphate isomerase), C0CMQ5/C0CMQ6 (3-oxo-isoapionate transketolase), C0CMQ7 (D-apionate oxidoisomerase), C0CMQ8 (hydroxypyruvate reductase), F8GV06 (D-apionate oxidoisomerase), Q2JZQ0 (3-oxo-isoapionate 4-phosphate decarboxylase), Q2JZQ5 (D-apiose binding SBP), Q6D5T7 (D-apiose isomerase), Q6D5T8 (D-apulose kinase), Q6D8V3 (D-apionate oxidoisomerase), Q6D8V4 (3-oxo-isoapionate decarboxylase), Q6D8V5 (L-erythrulose 1-phosphate isomerase), Q6D8V6 (L-erythrulose kinase), Q6D8V9 (D-erythrulose 4-phosphate isomerase), and Q7CK99 (D-apiose isomerase).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Sequence Similarity Networks (SSNs) and Genome Neighborhood Networks (SSNs)

The EFI-EST (<http://efi.igb.illinois.edu/efi-est/>)^{10,12} and EFI-GNT (<http://efi.igb.illinois.edu/efi-gnt/>)¹⁰ web tools were used to generate SSNs and GNNs, respectively. SSNs for Pfam families were generated using Option B of EFI-EST; sequence-function space in the SSNs was analyzed with Cytoscape, a desktop platform for visualizing complex networks (<http://www.cytoscape.org/>), using node attributes from the UniProtKB and other databases to assist segregating the SSN into isofunctional clusters.

EFI-EST uses a local database with sequences and bioinformatic annotations downloaded from the UniProtKB database and other databases; the SSNs in this manuscript were generated using sequences in UniProt 2017_07 (July 5, 2017) and InterPro 64 (July 6, 2017). Sequences homologous to a user-supplied query are collected using Option A of EFI-EST; Option A was used to identify the members of the “D-apionolactone hydrolase” family that

has not been curated by Pfam. SSNs for Pfam families were generated using Option B of EFI-EST. The alignment scores used for filtering the SSNs are given in the Figure legends.

The GNN for the clusters in an input SSN is generated with EFI-GNT. EFI-GNT generates a GNN cluster for each input SSN cluster, with the hub-node of each GNN cluster representing the proteins in the input SSN cluster and the spoke-nodes identifying the proteins encoded by the same genome neighborhoods as the proteins in the SSN cluster (within a $\pm N$ ORF window, where N is specified by the user; 10 by default); the spoke-nodes represent the Pfam families of the neighborhood proteins to assist functional predictions. The co-occurrence frequencies of the SSN cluster proteins and their genome neighbors as well as the distances between their genes in ORFs are provided; functionally related proteins (same metabolic pathway) should have large co-occurrence frequencies and short intergenic distances; unless specified, GNNs were generated with a $> 20\%$ query-neighbor co-occurrence frequency.

EFI-GNT uses a local database with genome sequences downloaded from the European Nucleic Acid (ENA) database; the GNNs in this manuscript were generated using ENA Release 132 (July 6, 2016). The neighborhood windows and co-occurrence frequencies used for collecting and displaying the Pfam neighbors are given in the Figure legends. EFI-GNT also collects proteins that are not members of Pfam families ($\sim 15\%$ of the proteins in the UniProt database are not assigned to a Pfam family); these are designated members of the “none” hub-nodes in the GNN; these have been deleted in the GNNs in the Figures for clarity.

EFI-GNT also provides interactive diagrams of the genome neighborhoods for the bacterial, archaeal, and fungal proteins in each cluster in the input SSN; with these, proteins encoded by genome neighborhoods in experimentally accessible organisms (gDNA availability and genetic tractability) can be selected for *in vitro* assays (proteins from multiple species often are selected to ensure that at least one orthologue can be purified) and *in vivo* characterization (phenotypes associated with genetic knockouts and/or regulation of transcripts).

The SSNs and GNNs were visualized and analyzed using Cytoscape 3.3.0 with a MacPro computer with 128GB RAM.

Sequence-Function Space in PF13407

The SSN for PF13407 (Supplementary Fig. 2) includes 603 clusters with >10 members, so the 117 SBPs that could be purified and screened (nodes highlighted in blue, magenta, or red in Supplementary Fig. 2) provide access to only a small fraction of sequence-function space in PF13407. Its functional complexity cannot be represented adequately with a phylogenetic tree—SSNs provide an easy-to-compute and visualize alternative for surveying sequence-function space in large protein families.

We also constructed a phylogenetic tree (Supplementary Fig. 34) for representative sequences in the D-apiose-binding clusters in Supplementary Figure 6 using the Neighbor-Joining method. The optimal tree with the sum of branch length = 5.4364709 was computed.

The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of number of amino acid substitutions per site. The analysis involved 61 amino acid sequences. All positions containing gaps and missing data were eliminated. The final dataset contained a total of 298 positions. The analysis was performed using MEGA7.

Despite the enhanced accuracy of trees to distinguish orthologues from paralogues we use SSNs: both 1) the greater speed of generating SSNs (pairwise sequence comparisons) instead of trees (multiple sequence alignments) for large families and 2) the ability to construct GNNs for clusters in even very large SSNs and then interactively inspect individual genome neighborhoods for the members of each cluster establish the advantage of the synergistic generation and interpretation of SSNs and GNNs.

Reagents and Analytical Methods

Unless otherwise specified, all solvents and organic chemicals were purchased from Sigma-Aldrich, and used without further purification. D-Apiose and DL-[1-¹³C] apiose were purchased from Omicron Biochemicals, Inc. Isopropyl- β -D-thiogalactoside (IPTG), ampicillin and kanamycin were purchased from Sigma-Aldrich. Primers were ordered from Integrated DNA Technologies. The nucleotide sequence was determined by ACGT. NMR spectra were recorded on Agilent 600MHz NMR. Enzyme assays were performed with a UV-visible spectrophotometer (Varian CARY 300Bio). The consumption or formation of NADH was monitored as the decrease or increase in the absorbance at 340 nm using an extinction coefficient (ϵ) of 6220 M⁻¹cm⁻¹.

Protein Expression and Purification

Expression and Purification of Solute Binding Proteins

The gene encoding the SBP of interest was PCR amplified from genomic DNA (Supplementary Table 6) and inserted into the N-terminal TEV-cleavable 6x-His-tag vector pNIC23-Bsa4, a pET23 based variant of the pNIC28-Bsa4 vector³⁷ by ligation independent cloning (LIC)³⁸. The periplasmic signal sequence as predicted by the webserver SignalP³⁹ was not included in the final construct. All growth media contained 100 μ g mL⁻¹ carbenicillin and 34 μ g mL⁻¹ chloramphenicol. *Escherichia coli* BL21 (DE3) containing the pRIL plasmid (STRATAGENE) was transformed with the cloned target and used to inoculate a 20 mL culture of 2xYT. The overnight growth was used to inoculate 2L of selenomethionine-containing ZYP-5052 (or methionine for native) autoinduction medium⁴⁰, in a LEX 48 airlift fermenter, incubated for 4 hours at 37 °C and then an additional 12–16 hours at 25°C. Purification of the solute binding protein from the pelleted cell mass was as previously described for the TRAP solute binding proteins⁴¹ with purification by metal affinity and size exclusion chromatography. The N-terminal TEV-cleavable 6X-His-tag was not removed prior to DSF experimentation.

Expression and Purification of Pathway Enzymes

Genes encoding enzyme targets were amplified by polymerase chain reaction (PCR) from respective genomic DNA using Phusion® high-fidelity DNA polymerase (New England BioLabs) and the oligonucleotide primers listed in Supplementary Table 6. The amplified genes were digested with appropriate restriction enzymes and ligated into similarly digested pET-28a, pET-15b, or pET23b (Novagen PET vectors) expression vectors for expression with an *N*-terminal or *C*-terminal His₆-tag. For the two-subunit Clostridial transketolase, the genes for both subunits were inserted into a pET-28a vector to be expressed cotranscriptionally. C0CMQ5 was produced with an N-terminal His-tag; C0CMQ6 was produced without a His-tag. Identifying information for each protein and gene is given in Supplementary Table 7.

The proteins were heterologously produced in *E. coli* strain BL21 (DE3). Cells transformed with the expression plasmids described above were grown at 37 °C in 2 L of Luria-Bertani broth (LB; supplemented with 50 µg/mL kanamycin for pET-28a or 100 µg/mL ampicillin for pET-15b or pET 23b) to OD₆₀₀ of 0.5~0.6, and expression of the genes was induced by addition of 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). The cultures were allowed to grow an additional 12 hr at 18 °C before the cells were harvested by centrifugation. The cells were resuspended in 80 mL of binding buffer (5 mM imidazole, 0.3 M NaCl, and 20 mM Tris-HCl, pH 7.9) and lysed by sonication (Fischer Scientific 550 Sonic Dismembrator). The lysate was cleared by centrifugation at 15,000 rpm for 40 minutes at 4 °C. The clarified supernatant containing the His-tagged protein was applied to a column containing 10 mL of Ni-NTA resin (QIAGEN) previously equilibrated with binding buffer. After equilibration of the Ni-NTA resin with the clarified supernatant on a rocking platform for 30 minutes, the flow-through was discarded. Then, the column was washed with 80 mL of wash buffer (25 mM imidazole, 0.3 M NaCl, and 20 mM Tris-HCl, pH 7.9) to elute weakly bound proteins. Resin-bound His-tagged protein was eluted with elution buffer-1 (250 mM imidazole, 0.3 M NaCl, and 20 mM Tris-HCl, pH 7.9) and collected in 4 mL fractions. The absorbance at 280 nm of each fraction was measured by a NanoDrop spectrophotometer; the purity of the fractions with strong absorbance at 280 nm (> 1 mg/ml) was checked by SDS-PAGE gel (mini-protean TGX precast gels from Bio-Rad). The purified fractions were pooled and dialyzed against 4 L of 20 mM Tris-HCl, pH 7.9 buffer three times (membrane tubing from Spectrum Laboratories). The recombinant proteins were concentrated, flash-frozen in liquid nitrogen, and stored at -80 °C prior to use.

For the transketolase (A6VKQ3, A6VKQ4, and C0CMQ5/C0CMQ6), the lysate containing the His-tagged protein was loaded on a 5 mL HisTrap FF column (GE Healthcare) equilibrated with binding buffer. Then the protein was eluted with a linear 100 mL gradient from 0% to 100% of elution buffer-2 (1 M imidazole, 0.3 M NaCl, and 20 mM Tris-HCl, pH 7.9) and collected in 5 mL fractions.

Cloning and Protein Expression for Decarboxylase RLP (B9JK73, Q2JZQ0) and Transcarboxylase RLP (A7IJG7):

DNAs encoding B9JK73, Q2JZQ0 and A7IJG7 were inserted in frame into a T7 promoter based pNYCOMPS23 vector using ligation independent cloning⁴²; the RLP genes were

fused to a leader sequence encoding an TEV cleavable C-terminal His₁₀ tag. The inserted coding region of the corresponding RLPs was sequenced (GENEWIZ) completely to exclude the acquisition of unwanted coding changes during DNA amplification and cloning.

The resultant plasmids were transformed into *E. coli* C2527 (DE3) (New England Biolab), which contained an additional pRIL (STRATAGEN) for optimal protein expression⁴³ and used as pre-cultures grown 37°C overnight in 25 ml selenomethionine containing PSAM media⁴⁴. For large-scale protein preparations, 2-liter bacterial cultures were grown in PSAM-5052 auto-induction media⁴⁴ supplemented with 100 µg/ml carbenicillin, 100 µg/ml chloramphenicol and 1 ml/l antifoam 204 (Sigma) at 37°C in LEX 48 airlift bioreactors (Epiphyte3, Canada). After 6 hours of growth, the temperature of the cultures was increased to 42°C, incubated for 30 min (heat-shock) to activate *E. coli* chaperon systems for optimal protein folding and then reduced to 22°C for overnight incubation.

Cells were harvested by centrifugation at 6,500xg and suspended in buffer containing 20 mM HEPES (pH 7.5), 500 mM NaCl, 20 mM imidazole, 0.1% IGEPAL, 20% sucrose, 1 mM β-mercaptoethanol (βME). Cells were disrupted by sonication and cell-debris was removed by centrifugation at 45,000xg. The supernatants were applied to a chromatography columns packed with 5 ml His60 superflow resin (Clontech) that had been equilibrated with buffer A (20 mM HEPES pH 7.5, 20 mM imidazole, 500 mM NaCl, 1 mM βME). The columns were washed with buffer A and the His₁₀ tagged RLPs were eluted with buffer B (20 mM HEPES pH 7.5, 350 mM NaCl, 250 mM imidazole, 1 mM βME). The C-terminal His₁₀ tags from the proteins was removed by overnight digestion at 4°C with the TEV protease at a 2000:1 ratio of RLP:TEV. The tag-free proteins were then separated from the His₁₀ tag and TEV protease by a Superdex 200 (16/60) size exclusion chromatography column equilibrated with buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl, 5% glycerol and 5 mM DTT. Potential peak fractions containing RLP proteins were assessed by SDS-PAGE, pooled concentrated to 15–20 mg/ml using a 30 kDa Amicon Ultra-15 centrifugal filter device (Millipore).

Crystallization, Data Collection and Structure Determinations of Solute Binding Proteins

Prior to crystallization, recombinant TEV protease⁴⁵ was added at a ratio of 1 to 80 protein/TEV protease and protein was buffer exchanged (into 20 mM Hepes pH 7.5, 5 mM DTT) and concentrated (to 40 mg mL⁻¹) by dilution and centrifugal ultrafiltration. Prior to crystallization with D-apiose, the *Rhizobium etli* CFN42 D-apiose SBP (Q2JZQ5, RheApSBP) was treated to remove copurified D-ribose. RheApSBP was diluted to 5 mg mL⁻¹ and incubated at 37° C for 30 minutes in the presence of 10 mM D-apiose. The protein was then reconcentrated to 40 mg mL⁻¹ by centrifugal ultrafiltration at 4°C and stored on ice prior to crystallization trials.

RheApSBP was crystallized by sitting drop vapor diffusion in 96-well Intelliplates (ART ROBBINS) stored at 18°C. The final crystallization conditions consisted of 0.5 µL of protein (40 mg mL⁻¹, 10 mM D-ribose or D-apiose, 20 mM Hepes pH 7.5, 5 mM DTT) combined with 0.5 µL of reservoir. For the D-ribose complex crystals were obtained in 20% w/v PEG4000, 200 mM CaCl₂, 100 mM TRIS pH 8.5 while for the D-apiose complex crystals were obtained in 20% w/v PEG4000, 200 mM CaCl₂. Crystals were mounted on nylon

loops, streaked through reservoir solution supplemented with glucose to 20% w/v and flash-cooled by plunging directly into liquid nitrogen. Data were collected at beamline 31-ID (LRL-CAT; Advanced Photon Source) using a wavelength of 0.9793 Å at 100 K and a Rayonix 225 HE detector. For the D-ribose complex data were integrated and scaled in HKL3000⁴⁶, initial phases were determined by selenomethionine SAD with HKL3000/SHELX⁴⁷, an initial model was built using HKL3000/ARPWARP⁴⁸ and the structure was refined with REFMAC. For the D-apiose complex data were integrated and scaled in MOSFLM, phases were determined using the D-ribose complex and the structure refined with PHENIX⁴⁹. During the final refinement cycles ligands were built into the observed difference density. Data collection and refinement statistics are given in Supplementary Table 2. There is a monomer per asymmetric unit with residues 26–312 fit to electron density (25–313 cloned). For both the D-ribose and D-apiose complex 99.7% of residues were in allowed regions of the Ramachandran plot.

Coordinates and structure factors have been deposited in the Protein Data Bank, and the accession codes are presented in Supplementary Table 2.

Differential Scanning Fluorimetry (DSF)

DSF utilized 384-microwell plates on an Applied Biosystems 7900HT real time PCR system with excitation at 490 nm and emission at 530 nm. The screening library consisted of 405 metabolites²¹, some as mixtures of up to six compounds, with each condition in duplicate and with eight control wells (no ligand). The final assay mix (10 µL) contained 10 µM protein, 1 mM ligand and 5X SYPRO ORANGE (Thermo-Fisher) in 100 mM HEPES pH 7.5, 150 mM NaCl, and 5 mM DTT. The temperature was increased at 3°C per minute from 22°C to 99°C with the melting temperature of the protein (T_m) calculated by fitting the melting curve to the Boltzmann equation⁵⁰. Among others, the D-apiose SBPs from *R. etli* CFN42 (RheApSBP, Q2JZQ5), *Actinobacillus succinogenes* 130Z (AsucApSBP, A6VKQ8), and *Burkholderia graminis* C4D1M (BgramApSBP, B1G898) were screened via this method (Supplementary Table 1).

Kinetic Assays and Production Identification by ¹H NMR

¹H NMR Assay for D-Apiose Isomerase and D-Apulose Kinase

The coupled reactions used for the ¹H NMR assays for the isomerase and kinase are shown in Supplementary Figure 8. The reaction mixture contained 50 mM sodium phosphate buffer, pH 8.0, 20 mM D-apiose, 2 mM MgCl₂, 5 mM KCl, 0.5 mM ATP, 19.6 mM PEP, 3 µM Q6D5T7, 3 µM Q6D5T8, and 2 U pyruvate kinase (PK) from rabbit muscle (Lyophilized powder, Sigma-Aldrich) in a total volume of 600 µL of D₂O. The reaction was incubated at 25 °C for 3 hr, and the ¹H NMR spectrum (water suppression) was recorded (Supplementary Figure 8).

¹H NMR Assay for Transketolase Activity with D-Apulose 4-Phosphate

The coupled reactions used for the ¹H NMR assay of the transketolase are shown in Supplementary Figure 9. The reaction mixture contained 50 mM Tris-Cl buffer (d₁₁), pH 7.5, 20 mM D-apulose 4-phosphate, 2.5 mM MgCl₂, 0.2 mM thiamine pyrophosphate

(ThDP), 0.2 mM dihydroxyacetone phosphate (DHAP), 2 U triosephosphate isomerase (TPI) from rabbit muscle (Lyophilized powder, Sigma-Aldrich), 2 μ M A6VKQ3, and 2 μ M A6VKQ4 in a total volume of 600 μ L of H₂O. The reaction was incubated at 25 °C for 18 hr. Then the reaction mixture was lyophilized, and the residue was dissolved in 600 μ L of D₂O before the ¹H NMR spectrum was recorded (Supplementary Figure 9).

Kinetic Assay for Transketolase Activity with D-Apulose 4-phosphate.

Transketolase activities were assayed by measuring the consumption of NADH. The formation of DHAP product was coupled with glycerol-3-phosphate dehydrogenase (G3PDH) from rabbit muscle (Sigma-Aldrich). For the D-glyceraldehyde 3-phosphate, D-erythrose 4-phosphate, and D-ribose 5-phosphate kinetic assays, the reaction mixtures (25 °C) contained 50 mM Tris-HCl buffer (pH 8.0), 2.5 mM MgCl₂, 0.1 mM thiamine pyrophosphate (ThDP), 0.16 mM NADH, variable concentrations of D-glyceraldehyde 3-phosphate / D-erythrose 4-phosphate / D-ribose 5-phosphate, 2.5 mM D-apulose 4-phosphate, 3 U G3PDH, and transketolase enzymes (pre-mixed A6VKQ3 and A6VKQ4) in a final volume of 200 μ L. For the D-glyceraldehyde 3-phosphate assays, the rate of the background reaction (absence of the D-apulose 4-phosphate) was subtracted. Data were fit to the Michaelis-Menten equation (kinetic constants are shown in Supplementary Table 3).

For the D-apulose 4-phosphate kinetic assay, the reaction mixture (25 °C) contained 50 mM Tris-HCl buffer (pH 8.0), 2.5 mM MgCl₂, 0.1 mM thiamine pyrophosphate (ThDP), 0.16 mM NADH, variable concentrations of D-apulose 4-phosphate, 5 mM D-erythrose 4-phosphate, 3 U G3PDH, and 0.5 μ M transketolase enzymes (pre-mixed A6VKQ3 and A6VKQ4) in a final volume of 200 μ L. Data were fit to the Michaelis-Menten equation (kinetic constants are shown in Supplementary Table 3).

Kinetic Assay for Oxidative Activity of D-Apiose Dehydrogenase with D-Apiose

Oxidation activities of D-apiose dehydrogenase proteins were assayed by measuring the formation of NADH. The reaction mixture (25 °C) contained variable concentrations of furanose D-apiose, 50 mM Tris-HCl buffer (pH 8.5), 1.5 mM NAD⁺, and enzyme in a final volume of 200 μ L. Data were fit to the Michaelis-Menten equation (kinetic constants are shown in Supplementary Table 4).

Kinetic Assay for Oxidative Activity of D-Apionate Oxidoisomerase with D-Apionate

Oxidative activities of D-apionate oxidoisomerase proteins were assayed by measuring the formation of NADH. The reaction mixture (25 °C) contained variable concentrations of D-apionate, 50 mM Tris-HCl buffer (pH 9.0), 1.5 mM NAD⁺, 1 mM MnCl₂, and enzyme in a final volume of 200 μ L. Data were fit to the Michaelis-Menten equation (kinetic constants are shown in Supplementary Table 4).

Kinetic Assay for Reductive Activity of Hydroxypyruvate Reductase

Reduction activity of hydroxypyruvate reductase protein was assayed by measuring the consumption of NADH. The reaction mixture (25 °C) contained variable concentrations of hydroxypyruvate, 50 mM Tris-HCl buffer (pH 8.0), 2 mM MgCl₂, 0.16 mM NADH, and

3.65×10^{-8} M C0CMQ8 in a final volume of 200 μ L. Data were fit to the Michaelis-Menten equation (kinetic constants are shown in Supplementary Table 4).

^1H NMR Assay for Lactonase Activity of a Hypothetical Protein with D-Apionolactone.

The reactions used for the ^1H NMR assay of the lactonase activity of the hypothetical protein are shown in Supplementary Figure 14. Two parallel reactions were conducted. One reaction mixture contained 50 mM phosphate-Na buffer, pD 8.0, 5 mM furanose D-apiose, 1 mM MgCl_2 , 0.25 mM NAD^+ , 4.8 mM Pyruvate, 1 μM B9JK80, and 2 U L-lactate dehydrogenase (LDH) from rabbit muscle (Lyophilized, Sigma-Aldrich) in a total volume of 600 μL of D_2O . The other reaction mixture contained 50 mM phosphate-Na buffer, pD 8.0, 5 mM furanose D-apiose, 1 mM MgCl_2 , 0.25 mM NAD^+ , 4.8 mM Pyruvate, 1 μM B9JK80, 1 μM A6X3G3, and 2 U LDH in a total volume of 600 μL of D_2O . The reactions were incubated at 25 $^\circ\text{C}$ for 1 hr before the ^1H NMR spectra were recorded (Supplementary Figure 14).

^1H NMR Assay for Oxidoisomerase Activity of D-Apionate Oxidoisomerase

The reactions used for the ^1H NMR assay of the oxidoisomerase are shown in Supplementary Figure 16. The reaction mixture contained 50 mM Tris-Cl (d_{11}) buffer, pD 8.5, 5 mM D-apionate, 1 mM MgCl_2 , 0.25 mM NAD^+ , 4.8 mM Pyruvate, 3 μM B9JK75 or Q6D8V3, and 2 U LDH in a total volume of 600 μL of D_2O . The reactions were incubated at 25 $^\circ\text{C}$ for 2 hr before the ^1H NMR spectrum was recorded (Supplementary Figure 16).

^1H NMR Assay for Transketolase Activity with 3-Oxo-isoapionate

The coupled reactions used for the ^1H NMR assay of the transketolase (C0CMQ5/C0CMQ6) are shown in Supplementary Figure 33. The reaction mixture contained 50 mM Tris-Cl buffer (d_{11}), pH 8.5, 5 mM D-apionate, 5 mM transketolase acceptor (D-glyceraldehyde 3-phosphate or D-erythrose 4-phosphate or D-ribose 5-phosphate in three reactions), 2.5 mM MgCl_2 , 0.25 mM NAD^+ , 0.2 mM thiamine pyrophosphate (ThDP), 3 μM C0CMQ8, 3 μM B9JK75, and 2 μM C0CMQ5/C0CMQ6 in a total volume of 600 μL of D_2O . The reaction was incubated at 25 $^\circ\text{C}$ for 24 hr before the ^1H NMR spectrum was recorded (Supplementary Figure 34).

^1H NMR Assay for the Decarboxylase Activity of Decarboxylase from the Xylose Isomerase Superfamily

The reactions used for the ^1H NMR assay of the decarboxylase are shown in Supplementary Figure 20. The reaction mixture contained 50 mM Tris-Cl (d_{11}) buffer, pD 8.5, 5 mM D-apionate, 1 mM MgCl_2 , 0.25 mM NAD^+ , 4.8 mM Pyruvate, 3 μM Q6D8V3, 1 μM Q6D8V4, and 2 U LDH in a total volume of 600 μL of D_2O . The reactions were incubated at 30 $^\circ\text{C}$ for 1 hr before the ^1H NMR spectrum was recorded (Supplementary Figure 20).

^1H NMR Assay for Kinase Activity and Isomerase Activities Following the Decarboxylase from the Xylose Isomerase Superfamily

The reactions used for the ^1H NMR assay of the kinase and isomerase are shown in Supplementary Figure 21. The reaction mixture for preparation of L-erythrulose 1-phosphate

or D-erythrose 4-phosphate contained 25 mM phosphate-Na buffer, pH 8.0, 10 mM L-erythrose or D-erythrose, 2.5 mM MgCl_2 , 50 mM KCl, 0.5 mM ATP, 9.5 mM PEP, 5 U/mL pyruvate kinase and 1 μM Q6D8V6 in a total volume of 600 μL of D_2O . The reactions were incubated at 30 °C for 2 hr, and ^1H NMR spectra show that the phosphorylation reactions were complete. Then the corresponding isomerases were added to a final concentration of 1 μM , and the ^1H NMR spectra were recorded after incubation at 30 °C for 30 min (Q6D8V5) or 1 hr (Q6D8V9) (Supplementary Figure 21).

Kinetic Assay for Kinase Activity of Q6D8V6

ATP-Dependent kinase activity was determined spectrophotometrically by following consumption of NADH. The formation of ADP was coupled to the oxidation of NADH via PK and LDH in the presence of ATP, PEP, and NADH. The reaction solution (25 °C) contained variable concentrations of substrate, 100 mM Tris-HCl buffer (pH 8.0), 0.16 mM NADH, 1.0 mM MgCl_2 , 1.0 mM KCl, 2.5 mM ATP, 2.5 mM PEP, 5 U/mL PK/LDH from rabbit muscle (Sigma), and enzyme (Q6D8V6) in a final volume of 200 μL . Data were fit to the Michaelis-Menten equation (kinetic constants are shown in Supplementary Table 5).

Kinetic Assay for Decarboxylation Activity of Q6D8V4

A stock solution of 3-oxo-isoapionate was enzymatically prepared with the following procedure: The reaction solution (600 μL) containing 50 mM Tris-DCI (D11) buffer (pH 8.5), 5.0 mM D-apionate, 1.0 mM MgCl_2 , 0.25 mM NAD^+ , 4.8 mM Pyruvate, 10U/ml LDH, 5.0 μM Q6D8V3 was incubated at 30 °C for 1 hr. Enzymes were removed using an Amicon® Ultra 0.5 mL centrifugal filter (10,000 NMWL, Merck Millipore Ltd.). The product concentration was quantitated based on the integration of the ^1H NMR. The decarboxylation activity of Q6D8V4 (0.05 μM) was acquired for variable concentrations of 3-oxo-isoapionate in the presence of erythrose kinase (Q6D8V6, 5 μM) assay kit describes above. The value of $k_{\text{cat}}/K_{\text{M}}$ was approximated from the initial slope of the linear range of the Michaelis-Menten curve (kinetic constants are shown in Supplementary Table 5).

^1H NMR Assay for ATP-Dependent Kinase Activity of 3-Oxo-Isoapionate Kinase

The coupled reactions used for the ^1H NMR assay for the kinase are shown in Supplementary Figure 25. The reaction mixture contained 50 mM Tris-Cl (d_{11}) buffer, pH 8.5, 5 mM D-apionate, 2 mM MgCl_2 , 5 mM KCl, 0.25 mM ATP, 4.8 mM PEP, 0.25 mM NAD^+ , 2.5 mM Pyruvate, 3 μM B9JK75, 3 μM B1G889, and 2 U PK/LDH in a total volume of 600 μL of D_2O . The reactions were incubated at 25 °C for 4 hr before the ^1H NMR spectrum (water suppression) was recorded (see Supplementary Figure 25).

^1H NMR Assay for Decarboxylase Activity of the RLP

The coupled reactions used for the ^1H NMR assay are shown in shown in Supplementary Figure 26. The reaction mixture contained 50 mM Tris-Cl (D11) buffer, pH 8.5, 5 mM D-apionate, 2 mM MgCl_2 , 5 mM KCl, 0.25 mM ATP, 4.8 mM PEP, 0.25 mM NAD^+ , 2.5 mM Pyruvate, 3 μM B9JK75, 3 μM B1G889, 2 μM B9JK73 or 2 μM Q2JZQ0, and 2 U LDH/PK from rabbit muscle (lyophilized powder, Sigma-Aldrich) in a total volume of 600 μL of

D₂O. The reaction was incubated at 25 °C for 4 hr, and ¹H NMR spectrum (water suppression) was recorded (see Supplementary Figure 26).

¹H NMR Assay for the Isomerase following the RLP Decarboxylase

The coupled reactions used for the ¹H NMR assay for the isomerases are shown in is shown in Supplementary Figure 27. For B9JN20, the reaction mixture contained 50 mM Tris-Cl (D11) buffer, pD 8.5, 5 mM D-apionate, 2 mM MgCl₂, 5 mM KCl, 0.25 mM ATP, 4.8 mM PEP, 0.25 mM NAD⁺, 2.5 mM pyruvate, 3 μM B9JK75, 3 μM B1G889, 2 μM B9JK73, 2 μM B9JN20, and 2 U LDH/PK from rabbit muscle (Lyophilized powder, Sigma-Aldrich) in a total volume of 600 μL of D₂O. For B9JN19, the reaction mixture contained 2 μM B9JN19 with the reagents in the B9JN20 reaction mixture. The reactions were incubated at 25 °C for 4 hr before the ¹H NMR spectra (water suppression) were recorded (see Supplementary Figure 27).

¹H NMR Assay for Transcarboxylase Activity of RLP

The coupled reactions used for the ¹H NMR assay for the transcarboxylase are shown in Supplementary Figure 29. The reaction mixture contained 50 mM Tris-Cl (d₁₁) buffer, pD 8.5, 5 mM D-apionate, 2 mM MgCl₂, 5 mM KCl, 0.25 mM ATP, 4.8 mM PEP, 0.25 mM NAD⁺, 2.5 mM Pyruvate, 3 μM B9JK75, 3 μM B1G889, 2 μM A7IJG7, and 2 U LDH/PK from rabbit muscle (lyophilized powder, Sigma-Aldrich) in a total volume of 600 μL of D₂O. The reaction was incubated at 25 °C for 4 hr before the ¹H NMR spectrum (water suppression) was recorded (see Supplementary Figure 29).

Mechanistic Study of Transcarboxylase Activity of RLP by ¹H NMR.

To prove intramolecular transfer of a sequestered CO₂ moiety, the transcarboxylation reaction was performed in H₂O solvent (600 μL total) using DL-[1-¹³C]-apionate as substrate. The reaction contained 50 mM Tris-Cl (d₁₁) buffer, pH 8.5, 10 mM DL-[1-¹³C]-apionate, 2 mM MgCl₂, 5 mM KCl, 0.25 mM ATP, 4.8 mM PEP, 0.25 mM NAD⁺, 2.5 mM Pyruvate, 5 μM B9JK75, 3 μM B1G889, 3 μM A7IJG7, and 2 U LDH/PK from rabbit muscle (lyophilized powder, Sigma-Aldrich) in a total volume of 600 μL of H₂O. The reaction was incubated at 25 °C for 18 hr. Then the reaction mixture was lyophilized; the residue was dissolved in 600 μL of D₂O before the ¹H NMR spectrum was recorded (see Supplementary Figure 30).

Biological and Genetic Methods

Strains and Growth Conditions

Escherichia coli strains were grown aerobically in LB at 37 °C with 100 μg/mL ampicillin or 50 μg/mL kanamycin when necessary. *P. carotovorum* WPP14⁵¹ and its derived strains were grown aerobically at 30 °C in LB (rich medium) or M9 salts buffer (Sigma) (defined media) with 10 mM of a carbon source and 50 μg/mL ampicillin or 25 μg/mL kanamycin when necessary. *A. radiobacter* K84 (ATCC BAA_868) and its derived strains were grown aerobically at 26 °C in TSB (rich medium) or defined medium²³ with 10 mM of a carbon source and 200 μg/mL kanamycin when necessary. *R. eutropha* N-1 (DSM 13513) and its derived strains were grown aerobically at 30 °C in LB and defined medium²³ with 10 mM of

a carbon source and 300 µg/mL kanamycin when necessary. *B. vulgatus* (ATCC 8482) was grown anaerobically in a Vinyl Anaerobic Chamber (Coy) (5% H₂, 10% CO₂, 85% N₂) in brain heart infusion media (rich media) or in defined media⁵² with indicated concentrations of carbon sources.

Isolation of Mutant Strains

Mutants of each species were isolated as described previously via double homologous recombination allelic exchange between the chromosome and a constructed suicide plasmid²³. Briefly, a suicide plasmid (pYMD1 or pK18mobsacB)⁵³ was constructed by inserting 750 bp – 1000 bp of the upstream and downstream regions of the coding region to be deleted (see Supplementary Table 6 for primers). The plasmid was transferred into a wild type target strain via electroporation or conjugation with S17–1 that contained the suicide plasmid. Transformed or recipient cells with the plasmid recombined into the genome were isolated via ampicillin (for pYMD1-based plasmids) or kanamycin (for pK18mobsacB-based plasmids) selection. Cells were then grown in nonselective conditions in rich medium to permit a second recombination event. Cells that experienced a second recombination event were isolated via sucrose selection. Resulting wild type revertants were distinguished from mutants by PCR amplification of the truncated genomic region. Mutant genotypes were confirmed by sequencing the region of the genome available for recombination with the plasmid. Mutants were complemented with the corresponding genes cloned into pSRKKm⁵⁴ or pBBR1MCS2⁵⁵. Complementation studies performed with pSRKKM included 0.1 µM IPTG.

Isolation of RNA

RNA was isolated from 4 mL cultures to which 4 mL of RNA later (Qiagen) was added. Cells were harvested by centrifugation for 5 min at 5,000 x g. Cells were resuspended in 1 mL TRIzol (Thermo-Fisher). The remaining protocol was performed via the PureLink RNA Mini Kit manufacturer's instructions.

qRT PCR quantification of transcripts

RNA was isolated from *B. vulgatus* cells grown to an OD₆₀₀ ~0.6. For each culture, reverse transcription reactions were performed from 350 ng of RNA using random primers as part of the ProtoScript First Strand cDNA Synthesis Kit (NEB) per manufacturer's instructions. Using a LightCycler 480 (Roche), threshold cycle values were observed from PCR reaction that contained Power Sybr Green Mix (Thermo-fisher), 2 µL of cDNA, and gene appropriate primers. Amplification parameters were 95°C for 10 min followed by 45 cycles of 95°C for 15 s, 50°C for 15 s, and 60°C for 1 min with fluorescence measurement. Product purity was established observing a melt curve. RT-PCR primers were designed with Beacon Designer 8.2 demo version, using default parameters. Primer concentrations were optimized for the lowest threshold cycle values, and amplification efficiency was determined and used for data analysis. Data were analyzed using the comparative cycle threshold method and normalized against three reference genes: *gyrA*, *ftsZ* and *rpoB*^{56,57}. Primers used are given in Supplementary Table 6.

Data availability

Atomic Coordinates and Structure Factors

The atomic coordinates and structure factors for the D-apiose- and D-ribose-liganded forms of the D-apiose-binding solute binding protein from *Rhizobium etli* CFN42 D-apiose SBP (UniProt ID Q2JZQ5) have been deposited in the Protein Data Bank (PDB 5IBQ and 4RY0, respectively).

Uniprot Accession IDs

Functions have been assigned to proteins with following UniProt IDs (Supplementary Table 7): A6VKQ3/A6VKQ4 (D-apulose 4-phosphate transketolase), A6VKQ8 (D-apiose binding SBP), A6X3G3 (D-apionate lactonase), A7IJG7 (3-oxo-isoapionate 4-phosphate transcarboxylase), B1G889 (3-oxo-isoapionate kinase), B1G894 (D-apiose dehydrogenase), B1G898 (D-apiose binding SBP), B9JK73 (3-oxo-isoapionate 4-phosphate decarboxylase), B9JK75 (D-apionate oxidoisomerase), B9JK80 (D-apiose dehydrogenase), B9JN19 (D-erythrulose 4-phosphate isomerase), B9JN20 (L-erythrulose 1-phosphate isomerase), C0CMQ5/C0CMQ6 (3-oxo-isoapionate transketolase), C0CMQ7 (D-apionate oxidoisomerase), C0CMQ8 (hydroxypyruvate reductase), F8GV06 (D-apionate oxidoisomerase), Q2JZQ0 (3-oxo-isoapionate 4-phosphate decarboxylase), Q2JZQ5 (D-apiose binding SBP), Q6D5T7 (D-apiose isomerase), Q6D5T8 (D-apulose kinase), Q6D8V3 (D-apionate oxidoisomerase), Q6D8V4 (3-oxo-isoapionate decarboxylase), Q6D8V5 (L-erythrulose 1-phosphate isomerase), Q6D8V6 (L-erythrulose kinase), Q6D8V9 (D-erythrulose 4-phosphate isomerase), and Q7CK99 (D-apiose isomerase).

Other data are available from the authors upon reasonable request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants U54GM093342 (to S.C.A. and J.A.G.) and P01GM118303 (to S.C.A. and J.A.G.) from the National Institutes of Health.

References

1. Zhao S et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* 502, 698–702, doi:10.1038/nature12576 (2013). [PubMed: 24056934]
2. Bastard K et al. Revealing the hidden functional diversity of an enzyme family. *Nature chemical biology* 10, 42–49, doi:10.1038/nchembio.1387 (2014). [PubMed: 24240508]
3. Sevin DC, Fuhrer T, Zamboni N & Sauer U Nontargeted in vitro metabolomics for high-throughput identification of novel enzymes in *Escherichia coli*. *Nature methods* 14, 187–194, doi:10.1038/nmeth.4103 (2017). [PubMed: 27941785]
4. Calhoun S et al. Prediction of enzymatic pathways by integrative pathway mapping. *eLife* 7, doi:10.7554/eLife.31097 (2018).
5. Zallot R, Harrison KJ, Kolaczowski B & de Crecy-Lagard V Functional Annotations of Paralogs: A Blessing and a Curse. *Life* 6, doi:10.3390/life6030039 (2016).

6. Schnoes AM, Brown SD, Dodevski I & Babbitt PC Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology* 5, e1000605, doi:10.1371/journal.pcbi.1000605 (2009). [PubMed: 20011109]
7. Babbitt PC & Gerlt JA Understanding enzyme superfamilies. *Chemistry As the fundamental determinant in the evolution of new catalytic activities. The Journal of biological chemistry* 272, 30591–30594 (1997). [PubMed: 9388188]
8. Gerlt JA & Babbitt PC Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual review of biochemistry* 70, 209–246, doi:10.1146/annurev.biochem.70.1.209 (2001).
9. Gerlt JA & Babbitt PC Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Current opinion in chemical biology* 2, 607–612 (1998). [PubMed: 9818186]
10. Gerlt JA Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence-Function Space and Genome Context to Discover Novel Functions. *Biochemistry* 56, 4293–4308, doi: 10.1021/acs.biochem.7b00614 (2017). [PubMed: 28826221]
11. Atkinson HJ, Morris JH, Ferrin TE & Babbitt PC Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PloS one* 4, e4345, doi:10.1371/journal.pone.0004345 (2009). [PubMed: 19190775]
12. Gerlt JA et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica et biophysica acta* 1854, 1019–1037, doi:10.1016/j.bbapap.2015.04.015 (2015). [PubMed: 25900361]
13. Zhao S et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife* 3, doi:10.7554/eLife.03275 (2014).
14. Picmanova M & Moller BL Apiose: one of nature's witty games. *Glycobiology* 26, 430–442, doi: 10.1093/glycob/cww012 (2016). [PubMed: 26848180]
15. Choi SH, Ruszczycky MW, Zhang H & Liu HW A fluoro analogue of UDP-alpha-D-glucuronic acid is an inhibitor of UDP-alpha-D-apiose/UDP-alpha-D-xylose synthase. *Chemical communications* 47, 10130–10132, doi:10.1039/c1cc13140k (2011). [PubMed: 21826368]
16. Choi SH, Mansoorabadi SO, Liu YN, Chien TC & Liu HW Analysis of UDP-D-apiose/UDP-D-xylose synthase-catalyzed conversion of UDP-D-apiose phosphonate to UDP-D-xylose phosphonate: implications for a retroaldol-aldol mechanism. *Journal of the American Chemical Society* 134, 13946–13949, doi:10.1021/ja305322x (2012). [PubMed: 22830643]
17. Eixelsberger T, Horvat D, Gutmann A, Weber H & Nidetzky B Isotope Probing of the UDP-Apiose/UDP-Xylose Synthase Reaction: Evidence of a Mechanism via a Coupled Oxidation and Aldol Cleavage. *Angewandte Chemie* 56, 2503–2507, doi:10.1002/anie.201609288 (2017). [PubMed: 28102965]
18. Smith JA & Bar-Peled M Synthesis of UDP-apiose in Bacteria: The marine phototroph *Geminicoccus roseus* and the plant pathogen *Xanthomonas pisi*. *PloS one* 12, e0184953, doi: 10.1371/journal.pone.0184953 (2017). [PubMed: 28931093]
19. Martens EC et al. Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS biology* 9, e1001221, doi:10.1371/journal.pbio.1001221 (2011). [PubMed: 22205877]
20. Ndeh D et al. Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature* 544, 65–70, doi:10.1038/nature21725 (2017). [PubMed: 28329766]
21. Wichelecki DJ et al. ATP-binding Cassette (ABC) Transport System Solute-binding Protein-guided Identification of Novel d-Altritol and Galactitol Catabolic Pathways in *Agrobacterium tumefaciens* C58. *The Journal of biological chemistry* 290, 28963–28976, doi:10.1074/jbc.M115.686857 (2015). [PubMed: 26472925]
22. Huang H et al. A General Strategy for the Discovery of Metabolic Pathways: d-Threitol, l-Threitol, and Erythritol Utilization in *Mycobacterium smegmatis*. *Journal of the American Chemical Society* 137, 14570–14573, doi:10.1021/jacs.5b08968 (2015). [PubMed: 26560079]
23. Zhang X et al. Assignment of function to a domain of unknown function: DUF1537 is a new kinase family in catabolic pathways for acid sugars. *Proceedings of the National Academy of*

Sciences of the United States of America 113, E4161–4169, doi:10.1073/pnas.1605546113 (2016). [PubMed: 27402745]

24. Lv Y et al. Crystal structure of Mycobacterium tuberculosis ketol-acid reductoisomerase at 1.0 Å resolution - a potential target for anti-tuberculosis drug discovery. *The FEBS journal* 283, 1184–1196, doi:10.1111/febs.13672 (2016). [PubMed: 26876563]
25. Tadrowski S et al. Metal Ions Play an Essential Catalytic Role in the Mechanism of Ketol-Acid Reductoisomerase. *Chemistry* 22, 7427–7436, doi:10.1002/chem.201600620 (2016). [PubMed: 27136273]
26. Patel K et al. Crystal structures of Staphylococcus aureus ketol-acid reductoisomerase in complex with two transition state analogs that have biocidal activity. *Chemistry*, doi:10.1002/chem.201704481 (2017).
27. Cleland WW, Andrews TJ, Gutteridge S, Hartman FC & Lorimer GH Mechanism of Rubisco: The Carbamate as General Base. *Chemical Reviews* 98, 549–562 (1998). [PubMed: 11848907]
28. Ashida H et al. A functional link between RuBisCO-like protein of Bacillus and photosynthetic RuBisCO. *Science* 302, 286–290, doi:10.1126/science.1086997 (2003). [PubMed: 14551435]
29. Imker HJ, Fedorov AA, Fedorov EV, Almo SC & Gerlt JA Mechanistic diversity in the RuBisCO superfamily: the “enolase” in the methionine salvage pathway in Geobacillus kaustophilus. *Biochemistry* 46, 4077–4089, doi:10.1021/bi7000483 (2007). [PubMed: 17352497]
30. Imker HJ, Singh J, Warlick BP, Tabita FR & Gerlt JA Mechanistic diversity in the RuBisCO superfamily: a novel isomerization reaction catalyzed by the RuBisCO-like protein from Rhodospirillum rubrum. *Biochemistry* 47, 11171–11173, doi:10.1021/bi801685f (2008). [PubMed: 18826254]
31. Erb TJ et al. A RubisCO-like protein links SAM metabolism with isoprenoid biosynthesis. *Nature chemical biology* 8, 926–932, doi:10.1038/nchembio.1087 (2012). [PubMed: 23042035]
32. Tabita FR, Satagopan S, Hanson TE, Kreeel NE & Scott SS Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *Journal of experimental botany* 59, 1515–1524, doi:10.1093/jxb/erm361 (2008). [PubMed: 18281717]
33. Tabita FR, Hanson TE, Satagopan S, Witte BH & Kreeel NE Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 363, 2629–2640, doi:10.1098/rstb.2008.0023 (2008). [PubMed: 18487131]
34. Erb TJ & Zarzycki J A short history of RubisCO: the rise and fall (?) of Nature’s predominant CO₂ fixing enzyme. *Curr Opin Biotechnol* 49, 100–107, doi:10.1016/j.copbio.2017.07.017 (2018). [PubMed: 28843191]
35. Yokota A Revisiting RuBisCO. *Biosci Biotechnol Biochem* 81, 2039–2049, doi:10.1080/09168451.2017.1379350 (2017). [PubMed: 28950756]
36. Bathellier C, Tcherkez G, Lorimer GH & Farquhar GD Rubisco isn’t really so bad. *Plant Cell Environ*, doi:10.1111/pce.13149 (2018).

Online methods references

37. Savitsky P et al. High-throughput production of human proteins for crystallization: the SGC experience. *Journal of structural biology* 172, 3–13, doi:10.1016/j.jsb.2010.06.008 (2010). [PubMed: 20541610]
38. Aslanidis C & de Jong PJ Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic acids research* 18, 6069–6074 (1990). [PubMed: 2235490]
39. Bendtsen JD, Nielsen H, von Heijne G & Brunak S Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology* 340, 783–795, doi:10.1016/j.jmb.2004.05.028 (2004). [PubMed: 15223320]
40. Studier FW Protein production by auto-induction in high density shaking cultures. *Protein expression and purification* 41, 207–234 (2005). [PubMed: 15915565]

41. Vetting MW et al. Experimental strategies for functional annotation and metabolism discovery: targeted screening of solute binding proteins and unbiased panning of metabolomes. *Biochemistry* 54, 909–931, doi:10.1021/bi501388y (2015). [PubMed: 25540822]
42. Gileadi O et al. High throughput production of recombinant human proteins for crystallography. *Methods Mol Biol* 426, 221–246, doi:10.1007/978-1-60327-058-8_14 (2008). [PubMed: 18542867]
43. Tropea JE, Cherry S, Nallamsetty S, Bignon C & Waugh DS A generic method for the production of recombinant proteins in *Escherichia coli* using a dual hexahistidine-maltose-binding protein affinity tag. *Methods Mol Biol* 363, 1–19, doi:10.1007/978-1-59745-209-0_1 (2007). [PubMed: 17272834]
44. Studier FW Stable expression clones and auto-induction for protein production in *E. coli*. *Methods Mol Biol* 1091, 17–32, doi:10.1007/978-1-62703-691-7_2 (2014). [PubMed: 24203322]
45. Blommel PG & Fox BG A combined approach to improving large-scale production of tobacco etch virus protease. *Protein expression and purification* 55, 53–68, doi:10.1016/j.pep.2007.04.013 (2007). [PubMed: 17543538]
46. Minor W, Cymborowski M, Otwinowski Z & Chruszcz M HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta crystallographica. Section D, Biological crystallography* 62, 859–866, doi:10.1107/S0907444906019949 (2006). [PubMed: 16855301]
47. Sheldrick GM A short history of SHELX. *Acta crystallographica. Section A, Foundations of crystallography* 64, 112–122, doi:10.1107/S0108767307043930 (2008). [PubMed: 18156677]
48. Morris RJ, Perrakis A & Lamzin VS ARP/wARP and automatic interpretation of protein electron density maps. *Methods in enzymology* 374, 229–244, doi:10.1016/S0076-6879(03)74011-7 (2003). [PubMed: 14696376]
49. Adams PD et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* 66, 213–221, doi:10.1107/S0907444909052925 (2010). [PubMed: 20124702]
50. Niesen FH, Berglund H & Vedadi M The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature protocols* 2, 2212–2221, doi:10.1038/nprot.2007.321 (2007). [PubMed: 17853878]
51. Mole B, Habibi S, Dangel JL & Grant SR Gluconate metabolism is required for virulence of the soft-rot pathogen *Pectobacterium carotovorum*. *Molecular plant-microbe interactions : MPMI* 23, 1335–1344, doi:10.1094/MPMI-03-10-0067 (2010). [PubMed: 20636105]
52. Varel VH & Bryant MP Nutritional features of *Bacteroides fragilis* subsp. *fragilis*. *Applied microbiology* 28, 251–257 (1974). [PubMed: 4853401]
53. Yamada K, Kaneko J, Kamio Y & Itoh Y Binding sequences for RdgB, a DNA damage-responsive transcriptional activator, and temperature-dependent expression of bacteriocin and pectin lyase genes in *Pectobacterium carotovorum* subsp. *carotovorum*. *Applied and environmental microbiology* 74, 6017–6025, doi:10.1128/AEM.01297-08 (2008). [PubMed: 18689515]
54. Khan SR, Gaines J, Roop RM, 2nd, & Farrand SK. 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* 74, 5053–5062, doi:10.1128/AEM.01098-08 (2008). [PubMed: 18606801]
55. Kovach ME et al. Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes. *Gene* 166, 175–176 (1995). [PubMed: 8529885]
56. Pfaffl MW A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic acids research* 29, e45 (2001). [PubMed: 11328886]
57. Rocha DJ, Santos CS & Pacheco LG Bacterial reference genes for gene expression studies by RT-qPCR: survey and analysis. *Antonie van Leeuwenhoek* 108, 685–693, doi:10.1007/s10482-015-0524-1 (2015). [PubMed: 26149127]

Figure 1.

(a) Structure of D-apiose. (b) Top: Structure of rhamnogalacturonan-II from higher plants. Reprinted with permission from reference 14. Bottom: Structure of apigalacturonan from aquatic monocots, e.g., *Zostera* and *Lemna*. D-Apiose is represented by the blue pentagons. (c) Ribbon diagram of Q2JZQ5 with D-apiose. (d) SSN for D-apiose-binding SBPs. (e) GNN for the SSN in (d). (f). Partial view of genome neighborhoods for proteins in the red SSN cluster in (d) and the red GNN cluster in (e).

Figure 2. The non-oxidative transketolase pathway.

(a) The metabolites and reactions in the pathway. (b) The genome neighborhoods that were characterized; the proteins that are encoded by the outlined genes were characterized *in vitro*. Gray genes encode D-apiose-binding SBPs; black genes encode components of an ABC transport system. (c) ^1H -NMR spectrum of the products of the transketolase (resonances highlighted by red bars). (d) Growth of *P. carotovorum* WPP14 wild type (black) and deletion strains (colors correspond to deletion of the genes using the colors in

(b) with D-apiose as the sole carbon source. Gene and protein identifying information is provided in Supplementary Table 7.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3. Oxidative pathway with a xylose isomerase family decarboxylase.

(a) The metabolites and reactions in the pathway. (b) The genome neighborhoods that were characterized; the proteins that are encoded by the outlined genes were characterized *in vitro*. White genes encode proteins of unknown function. Black genes encode major facilitator superfamily transporters. (c) ^1H -NMR spectrum of the decarboxylation product of the xylose isomerase family enzyme. (d) Growth of *P. carotovorum* WPP14 wild type (black) and deletion strains (colors correspond to deletion of the genes using the colors in (b)) with D-apionate as the sole carbon source. Gene and protein identifying information is provided in Supplementary Table 7.

Figure 4. Oxidative pathway with an RLP decarboxylase.

(a) The metabolites and reactions in the pathway. (b) The genome neighborhoods that were characterized; the proteins that are encoded by the outlined genes were characterized *in vitro*. The black genes (*A. radiobacter* and *O. anthropi*) encode ABC transport components. The black gene (*B. graminis*) encodes a major facilitator superfamily transporter. Gray genes encode SBPs. Genes for enzymes in the pathway after L-erythrulose-1-phosphate are not shown. (c) ^1H -NMR spectrum of the decarboxylation product of the RLP (resonances highlighted by red bars). (d) Growth of *A. radiobacter* K84 wild type (black) and deletion strains (colors correspond to deletion of the genes colored in (b) with D-apiose as the sole carbon source. Gene and protein identifying information is provided in Supplementary Table 7.

Figure 5. Oxidative pathway with an RLP transcarboxylase/hydrolase.

(a) The metabolites and reactions in the pathway. (b) The genome neighborhoods that were characterized; the proteins that are encoded by the outlined genes were characterized *in vitro*. (c) ^1H -NMR spectrum of the products of the RLP transcarboxylase/hydrolase (resonances highlighted by red and blue bars). (d) Growth of *R. eutropha* N-1 wild type (black) and deletion strains (colors correspond to deletion of the genes colored in (b) with D-apionate as the sole carbon source. The orange symbols illustrate growth of a strain in which the gene for the FAD-binding subunit (GlcE) of the glycolate oxidase was deleted. Gene and protein identifying information is provided in Supplementary Table 7.

Figure 6. Novel reactions and mechanisms in the catabolism of D-apiose.

(a) Reaction catalyzed by the D-apionate oxidoisomerase, with possible mechanisms (concerted or stepwise). (b) Comparison of the mechanisms of the reactions catalyzed by 3-oxo-isoapionate 4-phosphate decarboxylase and 3-oxo-isoapionate 4-phosphate decarboxylase transcarboxylase/hydrolase from the RuBisCO superfamily.