ARTICLE
PUBLISHED ONLINE: 17 NOVEMBER 2013 | DOI: 10.1038/NCHEMBIO.1387

nature chemical biology

Revealing the hidden functional diversity of an enzyme family

Karine Bastard[1–3,6], Adam Alexander Thil Smith[1–3,5,6], Carine Vergne-Vaxelaire[1–3], Alain Perret[1–3], Anne Zaparucha[1–3], Raquel De Melo-Minardi[1–4], Aline Mariage[1–3], Magali Boutard[1–3], Adrien Debard[1–3], Christophe Lechaplais[1–3], Christine Pelle[1–3], Virginie Pellouin[1–3], Nadia Perchat[1–3], Jean-Louis Petit[1–3], Annett Kreimeyer[1–3], Claudine Medigue[1–3], Jean Weissenbach[1–3], François Artiguenave[1–3], Véronique De Berardinis[1–3], David Vallenet[1–3] & Marcel Salanoubat[1–3]*

Millions of protein database entries are not assigned reliable functions, preventing the full understanding of chemical diversity in living organisms. Here, we describe an integrated strategy for the discovery of various enzymatic activities catalyzed within protein families of unknown or little known function. This approach relies on the definition of a generic reaction conserved within the family, high-throughput enzymatic screening on representatives, structural and modeling investigations and analysis of genomic and metabolic context. As a proof of principle, we investigated the DUF849 Pfam family and unearthed 14 potential new enzymatic activities, leading to the designation of these proteins as β-keto acid cleavage enzymes. We propose an *in vivo* role for four enzymatic activities and suggest key residues for guiding further functional annotation. Our results show that the functional diversity within a family may be largely underestimated. The extension of this strategy to other families will improve our knowledge of the enzymatic landscape.

The rate of protein functional elucidation lags far behind the rate of gene and protein sequence discovery, leading to an accumulation of proteins with no known function[1]. To address this problem, a call for community action was launched in 2004 for the annotation of genes of unknown function in microbial genomes[2]. The same year, a complementary call was published to associate at least one protein with each orphan enzymatic activity[3]. Despite these efforts, the problem has expanded as genome and metagenome sequencing projects have unleashed a deluge of gene and protein sequences with the arrival of next-generation sequencing technologies. Faced with such a seemingly insurmountable task, comparative genomics was proposed as the most effective strategy to predict functions for unknown proteins and genes for orphan activities[4]. Meanwhile, the structural elucidation of representative members of protein families, thanks to the Protein Structure Initiative, has opened up new opportunities for exploring protein function[5]. These two approaches are instrumental parts of the US Enzyme Function Initiative, an effort to address the challenge of assigning reliable functions to enzymes discovered in bacterial genome projects[6,7]. However, this project currently focuses on only 5 superfamilies representing approximately 100 Pfam families[8]. Pfam itself contains over 13,600 families, among which more than 3,000 families are of unknown function. Deciphering the role of these proteins would require a broader community-wide approach, with the development of high-throughput experimental pipelines guided by bioinformatics[9]. An additional difficulty, often underestimated, is that having a function assigned to a representative of a family does not guarantee that all of the other members of the family will have the same function[10,11].

Conversely, a number of biochemically known functions are yet to be associated to a protein family. Indeed, out of the 4,997 Enzyme Commission (EC) numbers presently formalized by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, 1,113 (22.3%) are sequence orphans[12]. We had previously discovered an initial association between a protein member (called Kce) and an enzymatic activity that was a sequence orphan[13]. This activity is involved in the lysine fermentation pathway and catalyzes the condensation of β-keto-5-amino-hexanoate (KAH (**1**)) and acetyl-CoA to produce aminobutyryl-CoA and acetoacetate. Recently, we resolved the three-dimensional structure of the Kce protein in presence of its substrate (i.e., KAH) and proposed a reaction mechanism[14]. Interestingly, Kce belongs to a Pfam family defined by the presence of a conserved domain of unknown function (DUF849), which contains over 900 proteins that are almost all of bacterial origin. Indeed, not all of the host organisms are capable of fermenting lysine, and some organisms contain several homologs of DUF849 proteins. This suggests the existence of a set of diverse biochemical reactions catalyzed by the different members of the family. Thus, DUF849 represented a good case study for the discovery of new activities within a family of unknown function.

Here we bring our contribution to the functional annotation challenge by presenting an integrated strategy that combines bioinformatics methods and experimental procedures (**Supplementary Results**, **Supplementary Fig. 1**). We applied this strategy to explore the functional diversity of the DUF849 family. First, we established a generic chemical reaction that could correspond to the studied family, along with a list of potential substrates. On the basis of clustering into *a priori* isofunctional subfamilies, we selected a set of representative proteins for a high-throughput all-against-all enzymatic screening. We also performed a structural and computational analysis of active sites, including homology modeling and substrate docking, to interpret the diversity of activities found within this

[1]Direction des Sciences du Vivant, Commissariat à l'energie atomique et aux énergies alternatives (CEA), Institut de Génomique, Evry, France. [2]CNRS-UMR8030, Evry, France. [3]Université d'Evry Val d'Essonne, Evry, France. [4]Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil. [5]Present address: Sleep/Wake Research Centre, Massey University, Wellington, New Zealand. [6]These authors contributed equally to this work. *e-mail: salanou@genoscope.cns.fr

family and to identify the structural elements responsible for specificity and promiscuity. Finally, we explored genomic and metabolic contexts, along with classical biochemical characterization, to reveal the *in vivo* role of some discovered activities.

RESULTS

From protein family to activity family

At the time of writing, DUF849 contained 922 proteins. We reduced this set because of the poor alignment quality of several sequences. The remaining 725 proteins are mostly present in bacteria, covering at least ten phyla, mainly Proteobacteria (**Supplementary Fig. 2**). For the vast majority of proteins, the DUF849 domain covers almost the totality of the sequence.

A multiple sequence alignment (MSA) of the 725 protein sequences showed highly conserved amino acids across the family (**Supplementary Data Set 1**). We found that eight amino acids located in the active site pocket, with respect to Kce structure, are conserved in 66% of the DUF849 proteins. We interpreted this good conservation in light of the three-dimensional structure of the Kce protein, mutagenesis experiments and the Kce reaction mechanism[14] (**Fig. 1**). The five most crucial of the active site amino acids, conserved in 86% of the family, are His46, His48 and Glu230, which are responsible for the coordination of a metal ion, and the Asp231–Arg226 charge-relay dyad involved in the abstraction of the pro-*S* proton from the C2 of KAH (**1**), which is mediated by a catalytic water molecule. Mutations in two of these five crucial residues lead to an inactivation of the Kce enzyme[14] and of the two other mutants we tested in this work (**Fig. 1a**). The three other residues are not strictly conserved in the family: Glu143 is hypothesized to be involved in the network of interactions that hold the catalytic water molecule in position[14], and Ser82 and Thr106, two uncharged hydrogen donors, stabilize the β-keto acid moiety of the charged KAH reaction intermediate through hydrogen bonds. Mutants for two of these three residues result in a decrease of the catalytic efficiency of the Kce enzyme[14] (**Fig. 1a**). The observation of these eight highly (but not perfectly) conserved key amino acids around the β-keto acid moiety in the active site strongly suggests that the reaction mechanism could be conserved in all the enzymes of the family involving substrates similar to KAH, i.e., with a β-keto acid moiety. We thus hypothesized a generic reaction type for all of the proteins of the family, involving the condensation of a β-keto acid with acetyl-CoA to produce a CoA ester and acetoacetate (**Fig. 1b**). We then named the family 'BKACE' for β-keto acid cleavage enzyme.

Seventeen candidate substrates that would fit this generic reaction (i.e., β-keto acids or β-amino or β-hydroxy acids that could act as possible precursors) were then selected from metabolic databases[15,16] and commercial resources (**Supplementary Note 1**). The selected substrates cover most of the diversity of the available metabolic β-keto acids (**Fig. 2**).

High-throughput enzymatic activity screening

We guided the selection of proteins representative of the functional diversity of the BKACE family by a partition of the family into putative isofunctional subfamilies. We implemented a bioinformatics strategy integrating various data sources, such as protein sequence similarity and phylogenetic analyses, as well as two new methods, one based on genomic context[17] and the other based on structural classification of active sites by the active sites modeling and clustering (ASMC) method[18]. Aggregated information from each of these methods was compiled by a cluster ensemble approach[19], as illustrated in **Supplementary Figure 3**. It yielded 32 'consensus' subfamilies with sizes varying from 3 to 130 proteins (**Supplementary Data Set 2**). **Supplementary Figure 4a–c** illustrates the coverage of the 725 BKACE proteins by the ASMC and genomic context (GC) clusters as well as by the cluster ensemble subfamilies.

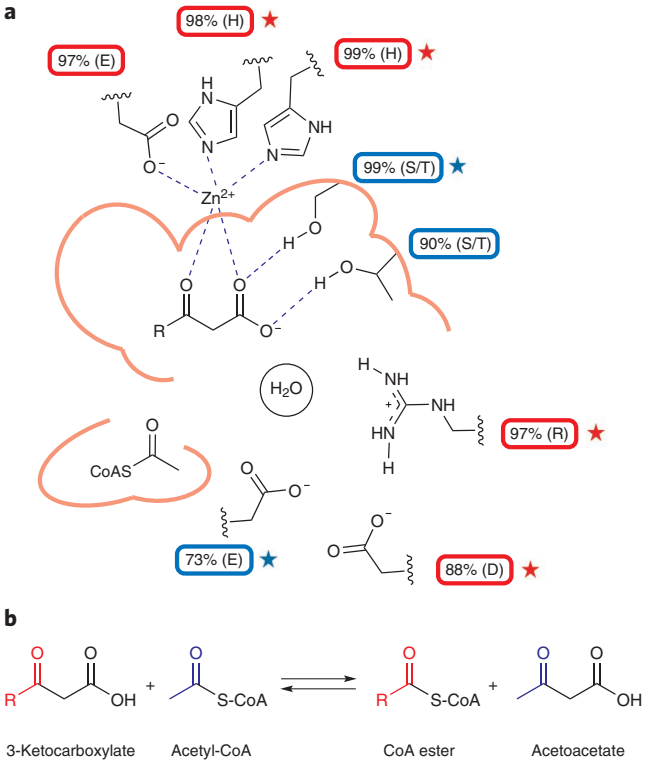

Figure 1 | The BKACE active site pocket and reaction. (**a**) Outline of the active site of the putative common core of the DUF 849 family based on the Kce structure. The colored rectangles indicate the highly conserved residues in the whole DUF 849 family, and the percentage of amino acid conservation is given. Red and blue rectangles indicate the five crucial residues and the three important residues for the reaction mechanism, respectively. Stars indicate the residues analyzed by mutagenesis on Kce protein. Red stars indicate that the mutation resulted in an inactive enzyme, whereas blue stars report a decrease in the catalytic efficiency. (**b**) A postulated generic reaction that could be catalyzed by the whole protein family. Color coding indicates the way in which the atoms are exchanged in the reaction.

Among the 322 candidate proteins (**Supplementary Data Set 3**), we successfully cloned 163 in an expression vector using our platform facilities. Out of these, 124 can be reproducibly overexpressed using SDS-PAGE. **Supplementary Table 1** provides explanations of cloning and expression failures for candidate proteins. These proteins still cover most of the diversity of the family and lead to a final coverage of at least one protein expressed for 24 of the 32 subfamilies (75%) (**Supplementary Fig. 4d,e**). When possible, we tested both forward and reverse reactions for the 17 candidate compounds (**Fig. 2**). For simplicity, we will refer by default to the β-keto acid transformation in the following text. We performed all of the experiments in duplicate. Because the amount of overexpressed proteins in cell lysates varied from protein to protein, activity values were not directly comparable between proteins. Using an activity measure distribution–based threshold, we qualitatively determined which proteins were likely to be active on a given substrate. The preprocess data are in **Supplementary Data Set 4**. We then tested for putative reactions in an all-enzyme versus an all-substrate screen. Enzymatic activities were observed for 15 of the 17 different proposed reactions (some in both forward and reverse directions), and 80 of the 124 proteins tested (65%) present an activity for at least one substrate (**Fig. 3**). Two substrates (4-hydroxybenzoylacetate (**14**) and 2-formamidobenzoylacetate (**17**)) gave enzymatic screening results that were not very reliable, and we did not analyze them further.

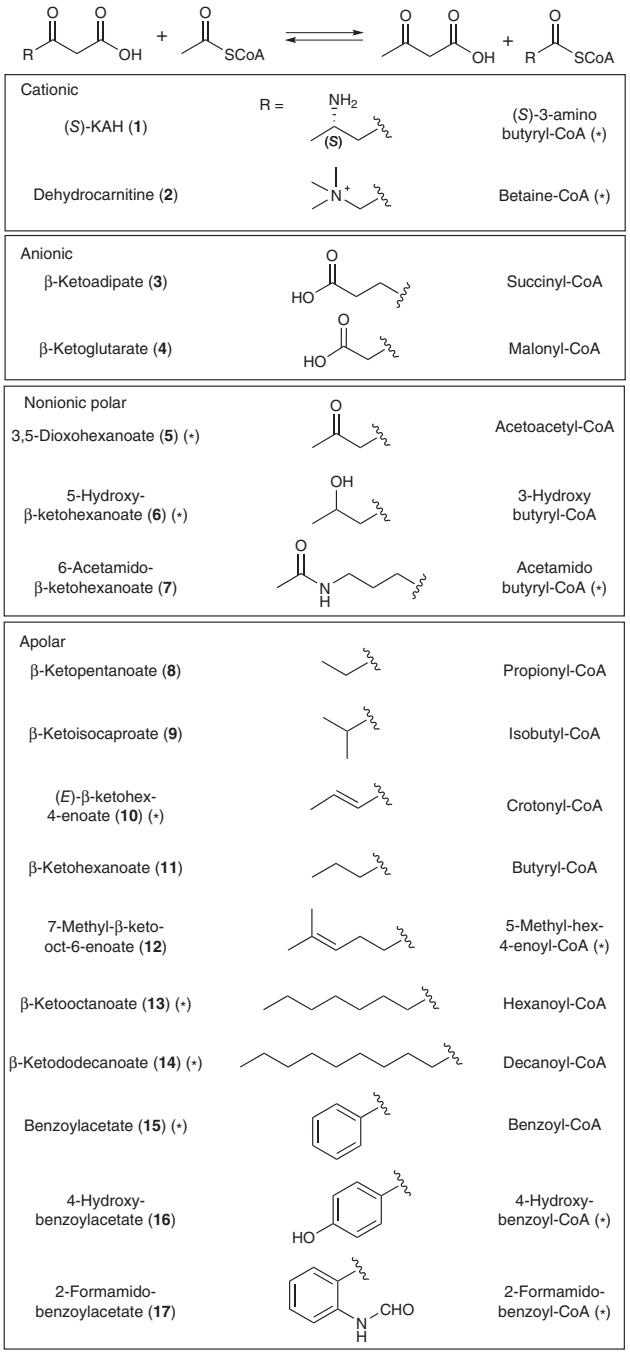

Figure 2 | Substrates used in enzymatic assays. Reactions were attempted in both forward and reverse directions, except for those where an asterisk is indicated on the compound not tested.

Results were quite similar between the two enzymatic tests for forward and reverse reactions, showing that the experiment is highly reproducible. We observed a strong association between consensus subfamilies and measured activities (corrected Fisher tests, most P values <0.05; **Supplementary Table 2**). This clearly indicates that our integrated bioinformatics strategy successfully separated the family into distinct isofunctional subfamilies.

Prediction of key residues responsible for specificities

As templates for homology modeling, we used the protein Kce[14] plus six other protein structures belonging to this family. These additional structures were neither crystallized with a ligand nor functionally annotated in any way and were deposited in the Protein Data Bank (PDB) as part of the Protein Structure Initiative. We divided the 'active site hierarchical tree' into seven main groups (referred to as G1 to G7), each containing between 50 and 156 sequences (**Fig. 4** and **Supplementary Data Set 5**). Amino acid sequence identities were moderately conserved within a group and were more distant between groups (**Supplementary Table 3**). Moreover, this structural partition is in agreement with the phylogenetic tree, except for some proteins of groups G2, G3 and G7, and suggests structural and evolutionary relationships that led to the functional diversification of the family (**Fig. 3** and **Supplementary Fig. 4a**). The observed discrepancies are linked to the differences between the two methodologies: one relies on a functional base by gathering common active sites, and the other relies on global evolutionary relationships among sequences (**Supplementary Note 2**).

We projected linearly the three-dimensional superposition of active sites belonging to the same group to form conservation patterns (represented by logo sequences on **Fig. 4** and **Supplementary Fig. 5**). For four of these groups (G1, G2, G4 and G5), the eight important residues are conserved (except for a subset of G5, which shows a glycine in position 10 in the active site pattern). The specificity of each group can be explained by one or several specific residues (specificity-determining positions (SDPs)) that are not conserved in the other groups at the same position in the active site pattern. Indeed, a strong correlation was found between the nature of the transformed compounds and our classification (**Fig. 3**). Almost all of the proteins from G1 transform hydrophobic and non-charged polar substrates, whereas G4 and G5 proteins catalyze only negatively and positively charged substrates, respectively. In some cases, proteins of a group show high substrate specificity, such as those from G2 that catalyze almost exclusively the transformation of KAH or those from G6 that act exclusively on β-ketoglutarate (**4**). In contrast, we detected various enzymatic activities in G3. Finally, the proteins from G7 do not transform any of the tested substrates. Notably, proteins from G1, G3, G4 and G5 transformed the 6-acetamido-β-ketohexanoate (**7**). The polarizable character of this substrate makes it a potential substrate for different types of active sites (electronegative or electropositive).

To explain how the presence of SDPs could be linked to the activity specificities in groups, we ran *in silico* docking simulations to position the substrate in the pocket (**Fig. 4**). In parallel, we analyzed the potential plasticity of the active sites to describe how some groups are strictly specific or, inversely, how some can catalyze a wide range of reactions. Enzymes from G4 are capable of transforming β-ketoadipate (**3**) and β-ketoglutarate. In our model of association, β-ketoadipate is stabilized by the SDP arginine found in position 8 (for about two-thirds of the G4 sequences) or in position 9 (for the remaining third) in the active site pattern. The active site can easily accept smaller substrates, such as β-ketoglutarate, owing to the flexibility of the active site loop[14], which carries the SDP arginine residue and surrounds the active site. In contrast to G4, G6 proteins catalyze the acetoacetate formation in the forward reaction only with β-ketoglutarate. Moreover, additional biochemical experiments have shown that acetyl-CoA is not required for this reaction. This suggests that G6 enzymes catalyze the decarboxylation of β-ketoglutarate and are not, in fact, BKACEs. The absence of the catalytic residues (i.e., aspartate and glutamate in position 21 and 14, respectively) supports the fact that G6 is not a BKACE. Docking simulations have also shown that the position of the β-ketoglutarate in G6 enzymes is quite different from that in the other β-keto acids docked into the other BKACEs (**Fig. 4**). Experiments have shown that the other group (i.e., G4) that catalyzes a BKACE reaction on β-ketoglutarate and contains the eight essential amino acids does not perform the simple decarboxylation into acetoacetate.

G5 enzymes catalyze the transformation of dehydrocarnitine (**2**) and, infrequently, KAH (**1**). In most of the G5 enzymes, our model of association shows that the SDP aspartate, found in position 9 in

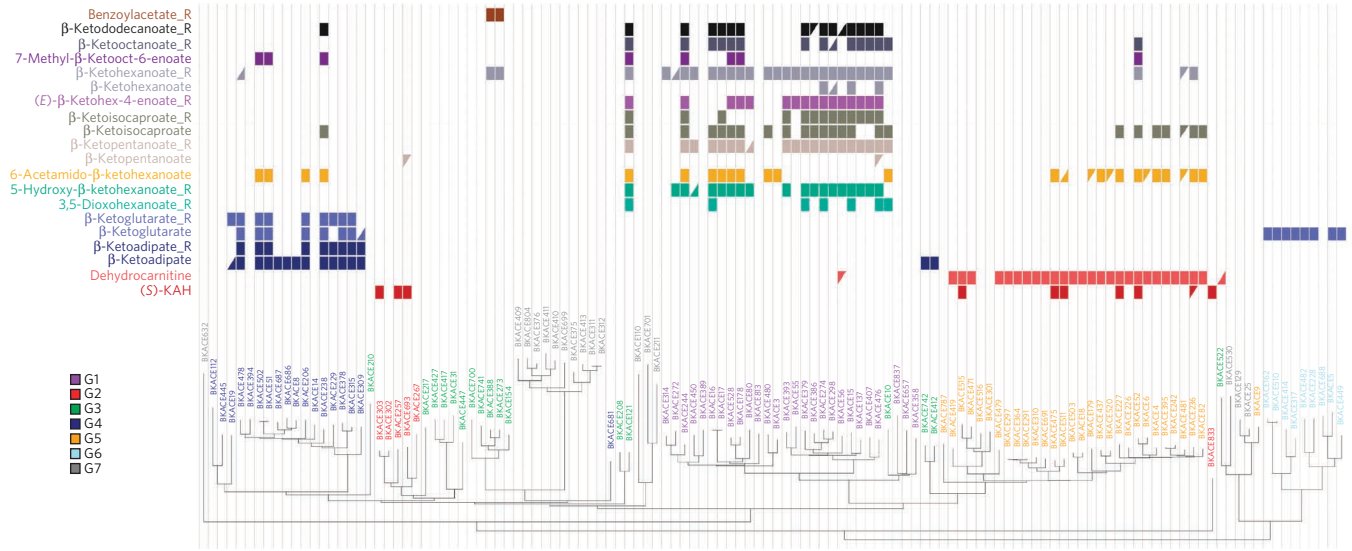

Figure 3 | Enzymatic screening of the BKACE family. Confirmed activities for each enzyme-substrate pair for each replicate test set are represented by colored triangles in the matrix. The phylogenetic tree of BKACE proteins (referred with identification numbers between 1 and 922) was calculated with QuickTree[32]. 'R' indicates that the substrate was tested in the reverse reaction. Hydrophobic substrates are colored from dark to light gray according to their degree of steric hindrance. Among hydrophobic substrates, those colored purple contain a double bond in their aliphatic chain. The main character of the substrates in yellow is polarizability, whereas that of substrates in green is polarity. Negatively charged substrates are in blue, whereas positively charged substrates are in red. BKACE groups G1 to G7 are shown by color on the protein phylogenetic tree.

the active site pattern, could interact with the dehydrocarnitine. In contrast to G5, G2 has a SDP glutamate in position 2 that forms a very specific interaction with KAH[14]. Another specificity of G5 is the presence of a strictly conserved SDP cysteine in position 16, which is part of the acetyl-CoA binding pocket; this cysteine does not seem to be involved in substrate selection (**Supplementary Fig. 6**). In the G1 group, a cap domain is present on the top of the active site and largely increases its size, which makes interactions easier with hydrophobic substrates of any size, including β-ketopentanoate (**8**) and β-ketododecanoate (**9**). Furthermore, the SDP of G1, which is an isoleucine in position 2, is in a key position at the edge of the active site for selecting hydrophobic substrates. G1 enzymes also show affinities for nonionic polar substrates such as 3,5-dioxohexanoate and 5-hydroxy-β-ketohexanoate. The polar residues found on the flexible loop (positions 6 to 9 in the active site pattern) probably attract these compounds. G3 has the particularity of containing a mixture of ASMC subgroups, from which a common active site pattern could not be precisely determined. Some members of G3 did not transform any tested substrate, which was probably due to the absence of a key residue in position 10. Other G3 proteins, which share structural similarities with G4 or G1, transform either β-ketoadipate or hydrophobic substrates (**Supplementary Note 3**). Proteins from G7 do not display activities toward any of the tested substrates; all of the sequences from this group are missing at least one of the eight important residues for catalyzing the BKACE reaction (an explanation of the heterogeneity of G3 and G7 is in **Supplementary Note 4**). Consequently, G7 proteins are probably not BKACE enzymes. This hypothesis is supported by the recent discovery of a *cis*-epoxysuccinate hydrolase[20,21], an enzyme of the G7 group.

In light of these observations, it seems clear that proteins containing the eight conserved amino acids important for the reaction mechanism of the original Kce have the expected BKACE activity. We thus propose an active site consensus pattern (i.e., a BKACE pattern) for active enzymes, '--HH[ST]----[STG]---E----RED', which can be used as a filter to discriminate BKACE from non-BKACE proteins in the family. In our biochemical study, the majority (66 out of 84) of the enzymes that fit this pattern show BKACE activity (**Supplementary Table 4**). The 'inactive' enzymes that do have the BKACE pattern (18 out of 84) are distributed equally in the five BKACE groups. For ~50% of the cases, we found an experimental or statistical explanation for failures. For the other 50%, it is impossible to rule out that the appropriate substrate was not tested (**Supplementary Table 5**). However, these 18 proteins are generally phylogenetically closed to groups of enzymes that show a common profile of activity. Thus, it is likely that these 18 cases would catalyze the same reaction as the group to which they belong. We emphasize that one should interpret the screening results as a whole and not on a one-to-one basis. In some particular cases, we detected an activity on BKACE pattern-free members (**Supplementary Table 4**).

From our initial set of 725 sequences, 470 (i.e., 65%) belong to five groups (G1, G2, G3, G4 and G5) that have the consensus pattern and thus are expected to perform BKACE activity. We found non-BKACE proteins (35%) exclusively in G6 and G7 and in a subset of G3. The G2 group that has the original Kce activity represents only 7% of the DUF849 family, indicating that the actual functions attributed to this family were largely underestimated.

These results are consistent with the observed activities within the family (**Supplementary Fig. 4f**). Furthermore, our structural analysis revealed important positions in the active site that are probably responsible for function specificity. These key residues (SDPs) are located in position 2 (with isoleucine, glutamate and glycine for G1, G2 and G5, respectively), position 8 or 9 (with arginine for G4 and aspartate for G5) or position 16 (with cysteine for G5). They may be used as markers to help the specific functional assignment of new BKACE family members. This structural classification into seven groups was used as a starting point for a deeper investigation of the *in vivo* roles of the BKACEs.

Discovery of new metabolic functions

We conducted an exhaustive genomic context analysis on the entire family to detect which *in vitro* functions could be associated to a cellular role. For each of the Genomic Context clusters (GCclusters) (**Supplementary Data Set 2**), we analyzed the BKACE neighbor gene functions to find metabolic pathways where BKACE activity may occur. In parallel, enzyme kinetics characteristics were determined

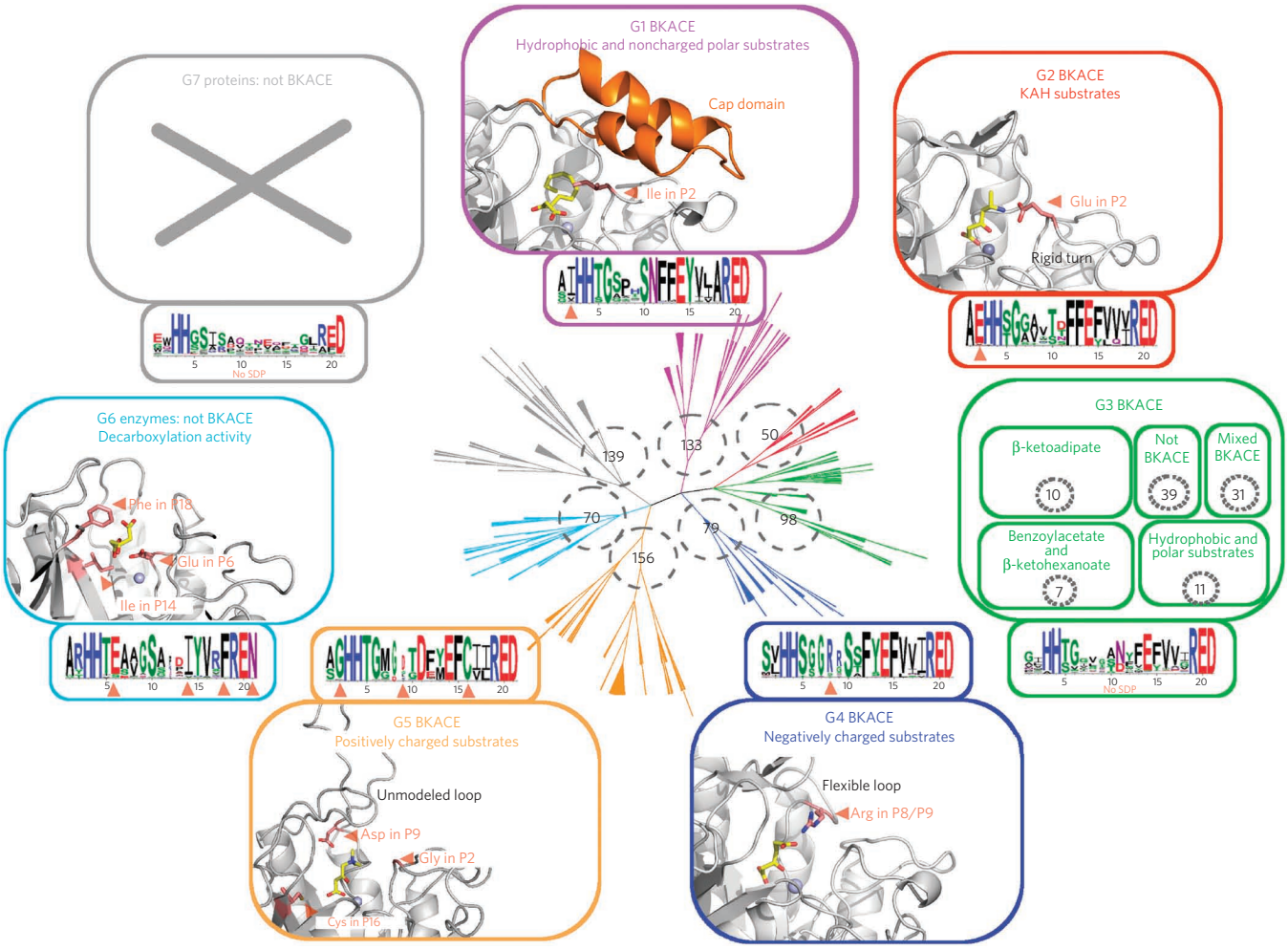

Figure 4 | Dividing the BKACE family into groups of similar active sites. The active site tree of the 725 sequences was generated by the ASMC method. The number of sequences belonging to each group is written at the bottom of each subtree. The nature of the active substrates is indicated below the BKACE group name. A sequence logo represents the conservation of the active site residues. Salmon-colored arrows indicate SDPs both in the active site pattern and in the three-dimensional models. For G3 and G7, no SDP was found as these groups do not show conservation. No models for G3 and G7 are presented as the activity within G3 is various, and G7 proteins do not show proper BKACE activity. Models of association between a representative BKACE for each group with one of the active substrate are drawn as follows: G1 with β-ketododecanoate; G2, crystal structure of Kce with KAH (**1**)[14]; G4 with β-ketoadipate (**3**); G5 with dehydrocarnitine (**2**) and G6 with β-ketoglutarate (**4**). In most of the G5 enzymes, one loop was difficult to model, explaining the low conservation of residues around position 9 in the logo. For G6, the position of the β-ketoglutarate (**4**) is quite different from that in the other β-keto acids docked into the other BKACEs. Glu in P6 pushes away the substrate, which then occupies the acetyl-CoA pocket, preventing the interaction with the Zn. Consequently, Zn cannot stabilize the keto-acid moiety and prevent the formation of a cyclic six-atom transition state, which is the first step of decarboxylation.

for selected enzymes, representatives of BKACE groups in various genomic contexts (**Supplementary Table 6**).

G2 contains the Kce protein (BKACE_267) of Candidatus *Cloacamonas aminovorans*[22], which was previously demonstrated to catalyze the cleavage of KAH in the context of lysine fermentation[13]. This protein has a catalytic efficiency over 10^4 M^{-1} s^{-1} for KAH, whereas no transformation has been detected for other substrates. Among all BKACEs of G2, a large proportion (48 out of 50) is found in organisms where most, if not all, of the genes necessary to ferment lysine are also present. This evidence suggests that G2 BKACEs are Kce proteins involved in lysine fermentation.

Almost all of the tested proteins from group G5 respond positively to the enzymatic screening with dehydrocarnitine as substrate. A large majority of these proteins (~70%) share a common genomic context with a conserved pattern composed of two additional colocalized genes encoding a carnitine dehydrogenase and a thiolase. In some organisms, these two genes are fused[23]. In *Pseudomonas aeruginosa*, which contains this basic pattern, a genomic region containing the carnitine dehydrogenase gene along with several other genes is identified[24].

However, the carnitine degradation pathway is not fully understood. Here, we show that one of these genes encodes a BKACE, which catalyzes the condensation of dehydrocarnitine and acetyl-CoA into acetoacetate and betainyl-CoA. This latter compound is then cleaved into betaine and coenzyme A via the previously mentioned thiolase. Biochemical studies on three BKACEs of the group G5 have shown that two of them metabolize dehydrocarnitine with k_{cat}/K_M ~10^4 M^{-1} s^{-1} (**Supplementary Table 6a**), which is a consistent value for most of the enzymes in the presence of their physiological substrates[25]. Their catalytic efficiencies are lower when KAH or β-ketohexanoate (**11**), which is nonpolar, are used as substrate (**Supplementary Table 6a**). In a few organisms, this carnitine degradation route could be part of two larger pathways for the γ-butyrobetaine degradation (**Fig. 5a**). First, an additional gene encoding an α-ketoglutarate–dependent dioxygenase may be present as in *Burkholderia ambifaria* and would lead to the direct

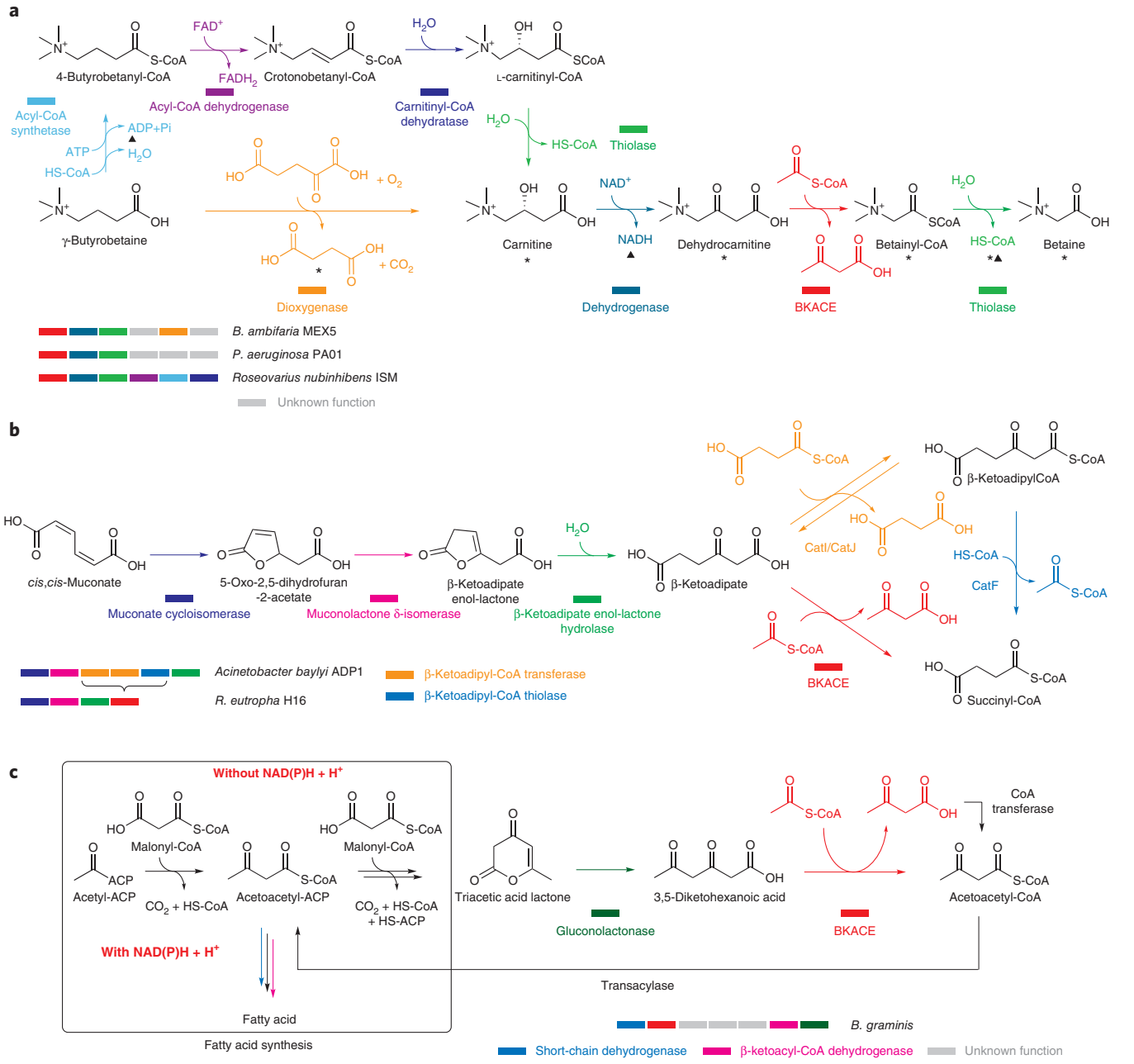

Figure 5 | Proposed *in vivo* roles of the newly discovered BKACEs. A schematic representation of metabolic pathways where BKACE activity occurs is given with their corresponding genomic contexts. (**a**) BKACE-catalyzed conversion of dehydrocarnitine to betainyl-CoA within γ-butyrobetaine and carnitine degradation. (**b**) BKACE-catalyzed conversion of β-ketoadipate to succinyl-CoA within catechol degradation. (**c**) BKACE-catalyzed conversion of diketohexanoic acid to acetoacetyl-CoA as a step in fatty acid biosynthesis rescue via triacetic acid lactone. Asterisk denotes LC/MS assay; ▲ denotes spectrophotometric assay.

conversion of γ-butyrobetaine into carnitine[23] (**Fig. 5a**). We validated this pathway *in vitro*. Indeed, by means of a single LC/MS–based assay, we identified various compounds produced by the reconstituted metabolic pathway using the accurate mass and retention time of commercial standards. In the positive ionization mode, carnitine was detected at 162.11183 Da at 9.15 min (162.11186 Da and 9.24 min for the standard); betaine was detected at 118.08581 Da at 7.39 min (118.08578 Da and 7.33 min for the standard); and finally coenzyme A was detected at 768.11935 Da at 9.01 min (768.11967 Da and 8.96 min for the standard). In negative ionization mode, succinate was identified at 117.0199 Da at 9.11 min (117.01977 Da and 8.76 min for the standard). These compounds were absent in the negative control. In conclusion, we confirmed the identity of these metabolites, validating in turn the involvement of each gene in the degradation of γ-butyrobetaine (**Fig. 5a**). Second, an alternative route may occur using a set of three genes[26] (**Fig 5a**). All of these data are in agreement with the role of G5 BKACEs in dehydrocarnitine cleavage. These pathways are found in about 110 organisms in which a G5 BKACE is present.

The enzymatic screening of G4 BKACEs reveals that they metabolize preferentially negatively charged compounds (i.e., β-ketoadipate and β-ketoglutarate). By studying their genomic context, we found that a BKACE from *Ralstonia eutropha* H16 (BKACE_378) colocalized with genes involved in catechol catabolism to β-ketoadipate. This latter compound is known to be incorporated in the central metabolism via a two-step conversion, which leads to succinate and

acetyl-CoA and involves *catIJF*[27] (**Fig. 5b**). These corresponding genes are absent in the *R. eutropha* H16 operon and are replaced by a BKACE. Biochemical studies confirm that BKACE_378 preferentially metabolizes β-ketoadipate with the formation of acetoacetate and succinyl-CoA (**Supplementary Table 6a**). Thus, it is likely that the *in vivo* role of this BKACE is the degradation of β-ketoadipate through an alternative pathway (**Fig. 5b**). Likewise, this function can be extended to other G4 BKACEs (i.e., GCcluster 4, which contains exclusively 23 Alphaproteobacteria BKACEs). In their genomes, we found conserved syntenies with genes encoding the two subunits of a protocatechuate 3,4-dioxygenase, a 3-carboxy-*cis*,*cis*-muconate cycloisomerase and a β-ketoadipate enol-lactonase. These enzymes are involved in a second pathway of aromatic compound degradation via protocatechuate to β-ketoadipate.

G1 BKACE screening shows large substrate promiscuity for hydrophobic compounds. This enzymatic activity diversity is also reflected by the various genomic contexts that we observed. One of them (BKACE_274), found in *Burkholderia graminis* C4D1M, presents a synteny conservation of four genes (**Fig. 5c** and **Supplementary Data Set 2**). During the synthesis of fatty acids, acetoacetyl-ACP is synthesized from acetyl-CoA and malonyl-CoA. When NAD(P)H is available, fatty acid synthesis is carried out by the canonical pathway. However, if the NAD(P)H concentration is limited, the acetoacetyl-ACP is transformed to triacetic acid ACP by the fatty acid synthase[28,29]. This product is then transformed spontaneously into triacetic acid lactone. At this stage, fatty acid biosynthesis is stopped. The observed synteny with a putative gluconolactonase suggests that the lactone could be cleaved into 3,5-diketohexanoate and then converted to acetoacetyl-CoA by a BKACE. Thus, acetoacetyl-CoA may be transacetylated into acetoacetyl-ACP (**Fig. 5c**). In this hypothetic pathway, BKACE activity would rescue the fatty acid biosynthesis when the reducing power is low. Biochemical studies carried out with acetoacetate and acetoacetyl-CoA as substrates show a notable catalytic efficiency of BKACE_274 on acetoacetyl-CoA (k_{cat}/K_M is 5.7 10^3 M^{-1} s^{-1}), which is in agreement with this proposed metabolic role. Beside, some G1 BKACEs show enzymatic activities with β-ketoisocaproate (**9**). This result could indicate a role in a pathway for the conversion from leucine to valine through β-ketoisocaproate[30]. Unfortunately, we did not find corresponding genomic contexts.

Among the non-BKACE proteins (G7, G6 and a subset of G3), we only observed a β-ketoglutarate decarboxylation activity for G6 proteins. Their genomic context analysis shows synteny conservation with genes encoding aminotransferases and could indicate a pathway for the degradation of β-amino acids via β-keto acids.

DISCUSSION

As far as we know, no integrated study dedicated to the characterization of an entire enzymatic family has been published before our work. We extrapolated knowledge based on a single member's function to the whole family by using new inference methods. Our family partitioning approach, based on the cluster ensemble algorithm, captures diverging and converging information from different sources and is fully automatic. In contrast, the literature only includes functional investigations carried out on simple selections of representatives based on phylogenetic trees.

The general modus operandi behind the strategy presented here could be adapted and improved to explore the functional diversity of other families. To be applied, our approach needs at least one known enzymatic activity in a family. The reaction generalization is an important point as only compounds derived from this generalized reaction will be tested for an activity. The hypothesis is that the majority of the proteins of a family would catalyze reactions that fit with the defined generic transformation. The more favorable situation would be to have structural evidence that gives clues about the reaction mechanism. Thus, the conservation of catalytic residues can be verified within the family to further confirm that the reaction may also be conserved. In less favorable cases, a reaction generalization could be attempted by determining conserved amino acids in the family using the ASMC method. Then, the diversity of the family can be estimated by the characterization of the SDP residues. In parallel, the genomic context method can be used to compute the number of different genomic organizations, which may give clues about the diversity of physiological functions within the family.

It is thus our hope that this approach will be of use in extending knowledge of enzyme families both old and new. We estimate that a couple of hundred families of unknown function could be investigated by our strategy using protein or domain families of unknown function from Pfam (A and B sections) and UniProt[31].

Received 12 July 2013; accepted 2 October 2013;
published online 17 November 2013

METHODS

Methods and any associated references are available in the online version of the paper.

References

1. Galperin, M.Y. & Koonin, E.V. From complete genome sequence to 'complete' understanding? *Trends Biotechnol.* **28**, 398–406 (2010).
2. Roberts, R.J. Identifying protein function–a call for community action. *PLoS Biol.* **2**, E42 (2004).
3. Karp, P.D. Call for an enzyme genomics initiative. *Genome Biol.* **5**, 401 (2004).
4. Hanson, A.D., Pribat, A., Waller, J.C. & de Crecy-Lagard, V. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list–and how to find it. *Biochem. J.* **425**, 1–11 (2010).
5. Gifford, L.K., Carter, L.G., Gabanyi, M.J., Berman, H.M. & Adams, P.D. The Protein Structure Initiative Structural Biology Knowledgebase Technology Portal: a structural biology web resource. *J. Struct. Funct. Genomics* **13**, 57–62 (2012).
6. Gerlt, J.A. *et al.* The Enzyme Function Initiative. *Biochemistry* **50**, 9950–9962 (2011).
7. Lukk, T. *et al.* Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proc. Natl. Acad. Sci. USA* **109**, 4122–4127 (2012).
8. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
9. Furnham, N., Garavelli, J.S., Apweiler, R. & Thornton, J.M. Missing in action: enzyme functional annotations in biological databases. *Nat. Chem. Biol.* **5**, 521–525 (2009).
10. Huang, H. *et al.* Divergence of structure and function in the haloacid dehalogenase enzyme superfamily: *Bacteroides thetaiotaomicron* BT2127 is an inorganic pyrophosphatase. *Biochemistry* **50**, 8937–8949 (2011).
11. Schnoes, A.M., Brown, S.D., Dodevski, I. & Babbitt, P.C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).
12. Lespinet, O. & Labedan, B. Orphan enzymes? *Science* **307**, 42 (2005).
13. Kreimeyer, A. *et al.* Identification of the last unknown genes in the fermentation pathway of lysine. *J. Biol. Chem.* **282**, 7191–7197 (2007).
14. Bellinzoni, M. *et al.* 3-Keto-5-aminohexanoate cleavage enzyme: a common fold for an uncommon Claisen-type condensation. *J. Biol. Chem.* **286**, 27399–27405 (2011).
15. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
16. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **40**, D742–D753 (2012).
17. Deniélou, Y.P., Sagot, M.F., Boyer, F. & Viari, A. Bacterial syntenies: an exact approach with gene quorum. *BMC Bioinformatics* **12**, 193 (2011).
18. de Melo-Minardi, R.C., Bastard, K. & Artiguenave, F. Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinformatics* **26**, 3075–3082 (2010).
19. Strehl, A. & Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining partitionings. *J. Mach. Learn. Res.* **3**, 583–617 (2002).
20. Pan, H., Bao, W., Xie, Z., Zhang, J. & Li, Y. Molecular cloning and characterization of a *cis*-epoxysuccinate hydrolase from *Bordetella* sp. BK-52. *J. Microbiol. Biotechnol.* **20**, 659–665 (2010).
21. Bao, W. *et al.* Analysis of essential amino acid residues for catalytic activity of *cis*-epoxysuccinate hydrolase from *Bordetella* sp. BK-52. *Appl. Microbiol. Biotechnol.* http://dx.doi.org/10.1007/s00253-013-5019-2 (2013).

22. Pelletier, E. *et al.* "Candidatus *Cloacamonas acidaminovorans*": genome sequence reconstruction provides a first glimpse of a new bacterial division. *J. Bacteriol.* **190**, 2572–2579 (2008).
23. Uanschou, C., Frieht, R. & Pittner, F. What to learn from a comparative genomic sequence analysis of L-carnitine dehydrogenase. *Monatsh. Chem.* **136**, 1365–1381 (2005).
24. Wargo, M.J. & Hogan, D.A. Identification of genes required for *Pseudomonas aeruginosa* carnitine catabolism. *Microbiology* **155**, 2411–2419 (2009).
25. Bar-Even, A. *et al.* The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).
26. Kleber, H.P. Bacterial carnitine metabolism. *FEMS Microbiol. Lett.* **147**, 1–9 (1997).
27. Collier, L.S., Gaines, G.L. III & Neidle, E.L. Regulation of benzoate degradation in Acinetobacter sp. strain ADP1 by BenM, a LysR-type transcriptional activator. *J. Bacteriol.* **180**, 2493–2501 (1998).
28. Yalpani, M., Willecke, K. & Lynen, F. Triacetic acid lactone, a derailment product of fatty acid biosynthesis. *Eur. J. Biochem.* **8**, 495–502 (1969).
29. Xie, D. *et al.* Microbial synthesis of triacetic acid lactone. *Biotechnol. Bioeng.* **93**, 727–736 (2006).
30. Monticello, D.J. & Costilow, R.N. Interconversion of valine and leucine by *Clostridium sporogenes*. *J. Bacteriol.* **152**, 946–949 (1982).
31. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009 (2011).
32. Howe, K., Bateman, A. & Durbin, R. QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics* **18**, 1546–1547 (2002).

Acknowledgments

We are grateful to M. Besnard-Gonnet, D. Baud and A. Fossey for excellent technical assistance and to S. Tricot and L. Stuani for MS analyses. We also thank P.L. Saaidi, G. Cohen and M. Stam for helpful discussions and D. Roche, P. Bowe and A. Tolonen for useful comments on the manuscript. This work was supported by grants from Commissariat à l'énergie atomique et aux énergies alternatives (CEA), the CNRS and the University of Evry. A.A.T. Smith has been supported by the MICROME project, European Union Framework Program 7 Collaborative Project (222886-2).

Author contributions

D.V. and M.S. designed and supervised this study. K.B. and A.A.T.S. conducted all of the bioinformatics analyses. F.A. and R.D.M.-M. initiated structural bioinformatics. C.V.-V. and A.Z. contributed to the chemical sections. K.B., A.A.T.S., D.V. and M.S. wrote the manuscript. C.M. and J.W. contributed to the design of this study. A.K. gave support to metabolic analysis. A.P. and V.D.B. supervised experiments, and J.-L.P., V.P., A.D., A.M., N.P., M.B., C.L. and C.P. performed the experiments.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to M.S.

ONLINE METHODS

Sequence selection. Though DUF849 is part of the curated component of Pfam, meaning that the initial seed alignment was revised by an expert curator, the aggregated set of proteins was not submitted to expert analysis. We examined a MSA of proteins from the family (built using MAFFT[33], with 1,000 iterations and the genafpair option) and removed several proteins from the subsequent analysis, deemed as incorrect additions because of poor alignment quality. Finally, only 725 proteins remained.

BKACE subfamilies. Several bioinformatics data sources were used and led to the creation of a 'primary' clustering of the selected DUF849 proteins.

'SIM' clusters. All protein-versus-protein alignments were calculated with gapped BLASTp and the BLOSUM62 scoring matrix. Results were parsed into a 725 × 725 similarity matrix, using the −log10 of the hit *e* values (note that *e* values were averaged in the case of multiple HSPs; <2.5% of protein-protein scores were affected by this very rough approximation). The similarity matrix was then filtered, removing low values (threshold 50) before being rendered symmetric. A distance matrix was then derived from it and passed to the basic complete linkage algorithm available in R. The resulting tree was cut to build the SIM clustering.

'*PHYLO*' clusters. A MSA for the 725 DUF849 proteins with QuickTree[32] established a phylogenetic tree (1,000 bootstraps, kimura option for the translation of pairwise distances). This tree was then inspected and cut, taking into account bootstrap values.

'SCIPHY' clusters. These clusters built and automatically cut another phylogenetic tree. Sci-Phy[34] software processes the same MSA, using the options '-dist tre -subfam ecost' to build and cut the tree into Sci-Phy clusters.

'ASMC' clusters. These used the ASMC method[18] to group proteins with similar active sites.

'GC' clusters. These used a new genomic-context clustering method focused on grouping proteins according to the conserved syntenies[17]. The protocol used to derive DUF849 protein clusters based on genomic context sharing combines the calculation of local syntenies with a similarity graph-clustering method. Syntenies containing DUF849 proteins were isolated using the Syntonizer software on the protein sequences encoded by genes included within a 10,000-bp window around the DUF849-coding genes (which typically captured an average of ten genes), with protein similarities being calculated with BLASTp, as previously presented. Similarities were restricted to those presenting more than 30% identity over at least 80% of the length of the smallest of two compared sequences, and only syntenies including at least three proteins, and never more than three successive gaps, were kept. Between each pair of DUF849-encoding genes, we chose as a similarity measure the number of genes included in any syntenies between them (with an imposed maximum of 20 to limit bias introduced by closely related organisms), counts that we rendered symmetrical by averaging to counter the inherent dissymmetry of both the BLAST and Syntonizer algorithms used. These counts were parsed into a 725-node similarity graph where each node represents a DUF849 gene and its neighborhood. This graph was iteratively filtered, removing below-weight-threshold edges and below-degree-threshold nodes until stabilization (thresholds were set empirically to the following values: minimal edge weight = 4, minimal node degree = 3). This allowed us to render the graph more robust in respect to the nontransitiveness of synteny gene counts between neighborhoods. Furthermore, the filtered graph was then submitted to a spectral clustering algorithm, with the largest clusters being iteratively subclustered, to obtain final genomic context clusters.

However, each clustering method led to different results, and simply taking the Cartesian product would result in highly granular and uninformative clustering. To properly integrate the previous clustering into single consensus subfamily-defining clustering, we used methods from the cluster ensembles statistical framework[19]. We used the R package 'clue' to integrate our clustering results. This package allows the assignment of different weights to the primary clustering results to give them more or less influence on the final consensus. We manually assigned the following weights: Genomic Context, 2.00 × fraction of non-NAs; Sci-Phy, 0.33 × fraction of non-NAs; SIM, 0.66 × fraction of non-NAs; Phylo, 0.33 × fraction of non-NAs; ASMC, 0.67 × fraction of non-NAs.

Non-clustered DUF849 proteins (NAs) for each classification were each assigned to their own newly formed singleton cluster to fit processing requirements. Clue offers several different methods for combining numerous clustering results, varying in how they measure distances between clusterings, in whether the result is a partition or a 'soft' (probabilistic) clustering and in algorithms. In this work, we used a 'hard' (to obtain deterministic cluster assignments) Manhattan distance–based method (to exaggerate disagreements between primary clusterings, as these can be seen as particularly informative), aiming for approximately 40 clusters.

Statistical analysis of enzymatic activities. *Activity preprocessing.* As the quantity of DUF849 protein in each set of protein wells could not be controlled precisely, activity measures cannot be readily compared between proteins of different origin. However, it is possible to extract from the data the binary presence or absence of a given activity for each protein using a mixed distribution approach. We hypothesize that each protein can be either active or inactive for a given activity, independent of its other activities or inactivities. Owing to the noncomparability of measures, when a protein is active on a given substrate, the measure can be modeled as coming from a wide uniform distribution across many possible values. When a protein is inactive on a given substrate, the measure is taken from a normal distribution centered on 0 or a value close to 0. Fitting such a mixture of distributions to the data for each activity, it is then possible to establish a cutoff for the measures, thus separating actives from inactives. The mixture of two distributions was fitted in R using the fixmix function from the package 'mixdist'. Unfortunately, mixdist does not support mixing different types of distributions (i.e., normal and uniform), so both were taken as normal (a normal distribution with a large variance can be assimilated to a uniform distribution for our purposes). The distribution with the lowest mean was taken as the 'inactive' distribution, with the one with the highest termed 'active'. We established a lower cutoff value for distinguishing samples from these two distributions by taking the activity level where the 'active' distribution is 50× higher than the inactive distribution (**Supplementary Fig. 7**).

Fisher test. To verify *a posteriori* that our initial partition of the DUF849 family is accurate in terms of isofunctionality, we carried out Fisher's test for each activity, crossing the active and inactive partition with the subfamily clusters. Fisher's test is designed to test for preferential (or avoided) combinations of modalities for two categorical variables. In this case, the test's base hypothesis is 'there are no preferential nor avoided combinations between active and inactive groups and the clusters'. Wherever this hypothesis is rejected, it means that globally, there are clusters that are preferentially (or not) active (or inactive) for the given activity. This would indicate that our working hypothesis of isofunctional clusters was correct. The Fisher tests were carried out in R using the base function 'fisher.test'. **Supplementary Table 2** gives crude and corrected *P* values (using the Bonferonni Hochberg method for multiple tests).

Structural and modeling analyses of active sites. The set of BKACE proteins was separated into 7 groups (G1 to G7) on the basis of the spatial positions of the amino acids that make up their active sites. The ASMC method[18] was run over the initial collection of 725 sequences using seven experimentally determined structures as templates for homology modeling: the holo form of Kce (presented in format PDB code:chain; 2Y7F:A), and six uncharacterized proteins, determined by the Protein Structure Initiative project, that all belong to the DUF849 Pfam family (3FA5:A, 3CHV:A, 3E49:A, 3E02:A, 3LOT:A, 3C6C:A). Amino acids of candidate models were aligned with the residues of the Kce active site and projected linearly to constitute a structure-based sequence alignment. These 'active site' sequences were subsequently classified with the COBWEB algorithm[35], an incremental system for hierarchical conceptual clustering that uses each position of the structure-based sequence alignment as supports. The resulting tree is called the 'active site hierarchical tree'. BKACE groups were obtained by cutting manually the tree as a function of the root, except for one branch, which was divided into two groups (G2 and G3) depending on the important position 2 in the active site sequence. To feed the cluster ensemble approach presented above with a higher diversity of active site profiles, ASMC clusters were generated automatically using a cutoff C = 0.47062. This cutoff was chosen because the more populated clusters generated, the better the representation of the BKACE groups created above. Eighty-four ASMC clusters were generated. Key amino acids responsible for the segregation between the BKACE groups (called SDPs) were detected by comparison of the conservation patterns obtained from the groups[18].

Models of association with substrates. Substrates (β-ketododecanoate (**14**), β-ketoadipate (**3**), dehydrocarnitine (**2**) and β-ketoglutarate (**4**)) were retrieved from the PubChem Compound database (http://pubchem.ncbi.nlm.nih.gov). For groups G1, G4, G5 and G6, a representative structure was

doi:10.1038/nchembio.1387

chosen; for G1, BKACE _386 (which corresponds to PDB code:chain 3E49:A); for G4, BKACE_14 (i.e., 3FA5:A); for G5, BKACE_291 (model having 56% of sequence identity with the closest structure), for G6, BKACE_317 (29% of sequence identity).

Constraint-driven docking was used for modeling substrates into the active site of a representative of groups G4 and G5. Ballon[36] generated an ensemble of possible conformations. The conformation that looked most like the KAH (**1**) configuration in the KAH (**1**)-Kce crystal structure was used for the docking. Using the Pair fitting wizard of PyMOL1.3 (Schrödinger, LLC), atoms of the keto acid moiety were superimposed onto the atoms of the keto acid moiety of KAH (**1**) in the KAH (**1**)-Kce crystal structure. The structure of the representative was then aligned with that of Kce. The system (representative and substrate) was then submitted to Steepest Descent energy minimization using the MMTK routine from Chimera software[37]. AMBER parameters were used for standard residue. Antechamber was used for assigning partial charges. We fixed the maximum number of steps to 5,000 (with a step size of 0.02 Å and an update interval of 10); all of the atoms of the enzyme and substrates were allowed to move.

Classical docking, using Autodock Vina software[38], was performed for representatives of groups G1 and G6. Gasteiger charges and nonpolar hydrogen atoms were added using AutoDockTools. Side chain protonation states assumed a negative charge for aspartate and glutamate, a positive charge for lysine and a neutral charge for histidine. For the G1 protein, the docking grid encompassed the pocket formed by the cap domain, whereas for the G6 protein, the docking grid included the entire pocket. The 2^+ charge of the Zn ion was added to Autodock Vina's atomic parameters. Protein side chains were allowed to move, and all degrees of freedom were possible for the ligand. The obtained docking poses were analyzed on the basis of scoring values of Autodock's scoring function. For group G1, the pose with the keto acid moiety bound to Zn was selected.

Cloning, expression and enzymatic screening. *Construction of the expression vectors.* Primers were designed using the Primer[39] program and chosen to be as close as possible to a length of 20 bases and to a melting temperature of 55 °C and to maintain homogeneous PCR conditions between all amplifications. The forward primers introduced a His_6 sequence in the proteins after the initial methionine for purification purposes. Genomic DNA used as the PCR templates were prepared by Multiple Displacement Amplification (GenomiPhi HY, GE Healthcare) from prokaryote strains mainly purchased from DSM and ATCC collections (**Supplementary Data Set 2**). PCR conditions for the genes presenting a GC content >65% were used in a second round[40]. The amplified sequences were inserted into pET22b(+) vector (Novagen) modified for ligation-independent cloning[41], and the sequences of the resulting plasmids were verified. **Supplementary Data Set 3** presents all of the primers for BKACE and the primers for PA5385 and PA5386 from *P. aeruginosa* and for Bamb_4455 from *B. ambifaria* (strain ATCC BAA-244 / AMMD).

Expression of the recombinant proteins for enzymatic screening. For each construction, the modified pET22b(+) vector was transformed into *E. coli* BL21 DE3 pLysE (Invitrogen). Transformed cells were grown in 96-well plates (HT96 BL21(DE3). Competent cells (Novagen) in 1.6 ml of Terrific Broth medium containing 0.5 M sorbitol, 5 mM betaine and 100 µg/ml carbenicillin at 37 °C until reaching an $A_{600\,nm}$ of 0.8. Isopropyl β-D-thiogalactopyranoside was added at a concentration of 500 µM to induce protein production, and the cells were further grown at 20 °C overnight. The cells were washed and suspended in 0.3 ml of 50 mM Tris-HCl, pH 7.5, containing 10% (v/v) glycerol, 1 mM Pefabloc SC (Roche Applied Science) and 0.2 µl of Lysonase TM bioprocessing reagent (Novagen), and then were sonicated using a Branson 2510 sonication water bath. After centrifugation, the clarified lysate was analyzed by SDS-PAGE to check for recombinant protein production. Protein concentration was determined by the Bradford method with BSA as the standard (Bio-Rad).

Enzymatic activity screening. Enzymatic activities were screened by testing the potential transformation of each substrate by each candidate protein. Depending on compound availability, the enzymatic test was performed using as substrates either the β-keto acid and acetyl-CoA (forward reaction) or the acetoacetate and CoA ester (reverse reaction). Enzymatic activities were estimated by measuring the initial reaction speed for either acetoacetate production (forward reactions) or for acetyl-CoA formation (reverse reactions).

Activity toward (*S*)-KAH (**1**) (β-keto-5-aminohexanoate), β-ketoadipate (**3**), β-ketoglutarate (**4**), 6-acetamido-β-ketohexanoate (**7**), β-ketopentanoate (**8**), β-ketoisocaproate (**9**), β-ketohexanoate (**11**), 7-methyl-β-ketooct-6-enoate (**12**), 4-hydroxybenzoylacetate (**16**) and 2-formamidobenzoylacetate (**17**) was assayed monitoring acetoacetate formation as previously described[14] in 50 mM Tris-HCl pH 7.5 containing 300 µM NADH, 250 µM acetyl-CoA, 0.035 U β-hydroxybutyrate dehydrogenase, 0.5 µg of cell lysate proteins and 1 mM of substrate. For β-ketopentanoate (**8**), β-ketoisocaproate (**9**), β-ketohexanoate (**11**), 7-methyl-β-ketooct-6-enoate (**12**), 6-acetamido-β-ketohexanoate (**7**), 4-hydroxybenzoylacetate (**16**) and 2-formamidobenzoylacetate (**17**), the substrates were previously prepared from their ester form and transformed *in situ* to their acid form by action of lipase (porcine liver esterase from Sigma-Aldrich, 0.4 –7 U per µmol of ester, depending on the ester), and then incubated for 30 min at room temperature. The assay reaction conditions were the same as those described above except for HBDH (0.035 U). Activity toward malonyl-CoA, acetoacetyl-CoA, butyryl-CoA, crotonyl-CoA, isobutyryl-CoA, decanoyl-CoA, D/L-3-hydroxy-butyrylCoA, *n*-propionyl-CoA, benzoyl-CoA, hexanoyl-CoA and succinyl-CoA was assayed by continuously monitoring the formation of acetyl-CoA according to Moriyama and Srere[42], in 50 mM Tris-HCl, pH 7.5, containing 400 µM 5,5′-dithiobis-(2-nitrobenzoic acid) (DTNB), 200 µM oxaloacetate, 1 mM acetoacetate, 0.1 U citrate synthase from porcine heart, 0.5 µg of cell lysate proteins and 250 µM of substrate.

Cleavage of dehydrocarnitine[43] was assayed monitoring CoA-SH formation using DTNB in a continuous coupled enzymatic assay. Briefly, dehydrocarnitine was produced from L-carnitine using the purified L-carnitine dehydrogenase from *P. aeruginosa* (PA5386). After dehydrocarnitine cleavage, the resulting betainyl-CoA was cleaved into betaine and CoA-SH using the betainyl-CoA thiolase from *P. aeruginosa* (PA5385). The reactions were conducted in 50 mM Tris-HCl, pH 9.0, containing 400 µM DTNB, 250 µM acetyl-CoA, 0.26 µg PA5385, 0.26 µg PA5386, 2 mM NAD^+, 0.5 µg of cell lysate proteins and 50 mM L-carnitine.

All of these reactions were conducted in duplicate experiments, at 25 °C, in 384-well plates and in a final volume of 70 µl using a SpectraMax Plus384 absorbance microplate reader (Molecular Devices). To determine the specific activity of the β-KACE for each substrate to be tested, two kinetics were recorded. The first kinetic was monitored in the presence of the substrate, and the second kinetic rate was recorded in its absence. The second kinetic was subtracted from the first to correct for background. The rectified rate of reaction was then used to calculate the specific activity.

Purification of the recombinant proteins. Cell culture, cell extracts, and protein purification were conducted as previously reported[13].

Enzyme assays for the purified proteins. α-Ketoglutarate–dependent dioxygenase activity from *B. ambifaria* MEX-5 was assayed monitoring the consumption of α-ketoglutarate with the use of glutamate dehydrogenase. First, reactions were performed in 200 µl 50 mM HEPES-NaOH, pH 7.6, containing 5 mM γ-butyrobetaine, 1 mM α–ketoglutarate, 5 mM sodium ascorbate, 1 mM $(NH_4)_2Fe(SO_4)_2$ and 23 µg of dioxygenase for 18 h at 25 °C. For the quantification of the remaining α–ketoglutarate, 90 µl of the mixture was combined with 130 µl 50 mM HEPES-NaOH, pH 7.6, containing 85 mM NH4Cl, 1 mM NADH and 4 µg of glutamate dehydrogenase.

For the BKACE reactions involving β-keto-5-aminohexanoate, β-ketoadipate, β-ketoglutarate and β-ketohexanoate as substrates, experiments were conducted in 200 µl 100 mM Tris-HCl, pH 7.5, containing 300 µM NADH and 0.25 U of β-hydroxybutyrate dehydrogenase. Initial rates were calculated using a molar extinction coefficient of 6,220 M^{-1} cm^{-1} for NADH at 340 nm.

Experiments involving dehydrocarnitine as substrate were conducted as follows: dehydrocarnitine was prepared enzymatically using L-carnitine and purified L-carnitine dehydrogenase (PA5386). L-carnitine (30 mM) was incubated in 2 ml 50 mM glycine-NaOH buffer adjusted to pH 12.0 in the presence of 6 mM NAD^+ and 50 µg PA5386 for 1 h at 25 °C. The reaction was stopped with 1% (v/v) trifluoroacetic acid and neutralized with 5 M K_2CO_3. The concentration of the synthesized dehydrocarnitine was estimated by the NADH concentration. Dehydrocarnitine was thus used as a substrate for β-KACE in 200 µl 100 mM Tris-HCl, pH 9.0, containing 450 µM DTNB and varying concentrations of acetyl-CoA, in a coupled assay in the presence of 1.3 µg betainyl-CoA thiolase (PA5385). Under these conditions, the rate of formation of CoA-SH was strictly proportional to the amount of β-KACE. Initial rates were calculated using a molar extinction coefficient of 13,600 M^{-1} cm^{-1} at 412 nm.

For the reactions involving acetoacetyl-CoA as substrate, experiments were conducted in 200 µl 100 mM Tris-HCl, pH 7.5, in the presence of 450 µM

DTNB, 500 µM oxaloacetate, 0.1 U citrate synthase from porcine heart and varying concentrations of acetoacetate. All of the kinetic parameters were determined by varying one substrate concentration while keeping the other at a fixed concentration. Kinetic constants were obtained from duplicate experiments by nonlinear analysis of initial rates using SigmaPlot 9.0 (Systat Software, Inc.). All of the reactions were performed at 25 °C in a Safas UV mc^2 double beam spectrophotometer.

Functional assay of the reconstituted γ–butyrobetaine degradation pathway. A multienzymatic assay, combining γ-butyrobetaine dioxygenase, carnitine dehydrogenase, BKACE (ASMSG5) and betainyl-CoA thiolase, was set up to generate betaine from γ-butyrobetaine. The coupled assay was conducted using 1 mM α-ketoglutarate, 5 mM γ-butyrobetaine, 1 mM $(NH4)2Fe(SO4)_2$, 5 µg carnitine dehydrogenase (PA5386 from *P. aeruginosa*), 2 mM NAD^+, 800 µm acetyl-CoA, 2 µg BKACE (BKACE 82, ASMCG5) and 1.4 µg betainyl-CoA thiolase (PA5385 from *P. aeruginosa*) in 200 µl 50 mM Tris-HCl pH 7.5. Reactions were initiated by the addition of α-ketoglutarate. Two different experiments were set up: in the first, 23 µg γ-butyrotetaine dioxygenase (from *B. ambifaria* MEX-5) was added, whereas in the second one, this enzyme was omitted (negative control). The reactions were stopped after 2 h by centrifugation on a 3-kDa molecular weight cutoff membrane. The global reaction was monitored by LC/MS in the positive and negative ionization modes.

LC/MS analyses were carried out using a LTQ/Orbitrap mass spectrometer coupled to an Accela LC system (Thermo-Fisher). Chromatographic separation was conducted using a ZIC-pHILIC column (150 × 4.6 mm × 5 µm; Merck Chemicals) kept at 40 °C. A mobile phase gradient was used with a flow rate of 0.5 ml/min, in which mobile phase A consisted of 10 mM ammonium carbonate and mobile phase B consisted of acetonitrile. The gradient started at 20% A for 2 min followed by a linear gradient at 60% A for 14 min and finally 8 min at 60% A. The entire eluant was sprayed into the mass spectrometer using a heated electrospray ionization source (250 °C) at ± 4 kV with sheath, auxiliary and sweep gases set at 60, 50 and 0 arbitrary units, respectively. Desolvation of the droplets was further aided by setting the heated capillary temperature at 275 °C. The metabolites were detected in both the positive and negative ionization mode by full-scan mass analysis from *m/z* 50–1,000 at a resolving power of 30,000 at *m/z* = 400.

Chemicals. Chemicals and enzymes were purchased from Sigma-Aldrich. Oligonucleotides were purchased from Sigma-Genosys. The other chemicals that were not commercially available were synthesized in house (**Supplementary Note 1**).

33. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
34. Brown, D.P., Krishnamurthy, N. & Sjolander, K. Automated protein subfamily identification and classification. *PLoS Comput. Biol.* **3**, e160 (2007).
35. Fisher, D. Knowledge acquisition via incremental conceptual clustering. *Mach. Learn.* **2**, 139–172 (1987).
36. Puranen, J.S., Vainio, M.J. & Johnson, M.S. Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. *J. Comput. Chem.* **31**, 1722–1732 (2010).
37. Pettersen, E.F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
38. Trott, O. & Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
39. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).
40. Ralser, M. *et al.* An efficient and economic enhancer mix for PCR. *Biochem. Biophys. Res. Commun.* **347**, 747–751 (2006).
41. Aslanidis, C. & de Jong, P.J. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* **18**, 6069–6074 (1990).
42. Moriyama, T. & Srere, P.A. Purification of rat heart and rat liver citrate synthases. Physical, kinetic, and immunological studies. *J. Biol. Chem.* **246**, 3217–3223 (1971).
43. Swart, M., Snijders, J.G. & van Duijnenb, Th.P. Polarizabilities of amino acid residues. *J. Comp. Meth. Sci. Eng.* **4**, 419–425 (2004).

doi:10.1038/nchembio.1387