Nontargeted *in vitro* metabolomics for high-throughput identification of novel enzymes in *Escherichia coli*

Daniel C Sévin[1–3], Tobias Fuhrer[1], Nicola Zamboni[1] & Uwe Sauer[1]

Our understanding of metabolism is limited by a lack of knowledge about the functions of many enzymes. Here, we develop a high-throughput mass spectrometry approach to comprehensively profile proteins for *in vitro* enzymatic activity. Overexpressed or purified proteins are incubated in a supplemented metabolome extract containing hundreds of biologically relevant candidate substrates, and accumulating and depleting metabolites are determined by nontargeted mass spectrometry. By combining chemometrics and database approaches, we established an automated pipeline for unbiased annotation of the functions of novel enzymes. In screening all 1,275 functionally uncharacterized *Escherichia coli* proteins, we discovered 241 potential novel enzymes, 12 of which we experimentally validated. Our high-throughput *in vitro* metabolomics method is generally applicable to any purified protein or crude cell lysate of its overexpression host and enables performing up to 1,200 nontargeted enzyme assays per working day.

20 years after the sequencing of the first genome, 30–50% of the gene sequences listed in public databases still await functional annotation (**Supplementary Fig. 1**)[1–5]. Fueled by rapid advances in sequencing technology that by far outpace experimental gene annotation, the number of functionally uncharacterized genes is projected to increase even further[5]. Many functionally uncharacterized genes have only low sequence similarity to genes encoding proteins with known biological function and consequently cannot be annotated by homology-based approaches[6,7]. Nevertheless, ample evidence suggests that uncharacterized genes are likely to encode functionally relevant proteins[3,8]. For instance, over 70% of the encoded uncharacterized proteins can be grouped into families that are phylogenetically conserved[3,9], and this implies the presence of function-imposed evolutionary constraints. Hence, novel approaches are needed to identify the functions of nonannotated genes and make the accumulating genome sequence data accessible to biological interpretation.

A considerable fraction of nonannotated genes can be expected to encode metabolic enzymes; for instance, one-fourth of all known biochemical reactions with assigned Enzyme Commission (EC)[10] numbers have not yet been linked to any specific gene product[11–13]. The annotation of genes encoding such 'orphan' enzymes is expected to improve the effectiveness of metabolic engineering strategies and to enhance the predictive quality of genome-scale metabolic reconstructions[2,14–16]. Consequently, various computational and experimental methods have been developed to discover missing enzymes[17–25]. One approach that provides direct experimental evidence is activity-based metabolic profiling, which consists of incubating purified proteins in a mixture of candidate substrates and monitoring metabolite conversion as a reporter of enzymatic activity[18]. Despite its appealing conceptual simplicity, the scope of previous implementations of this approach was considerably limited by coverage or scalability of the employed analytical techniques. Methods based on chromatography- and electrophoresis-coupled mass spectrometry, which can detect dozens of potential reactants in a single measurement, provided insufficient throughput to screen large protein cohorts for enzymatic activity[22,26]. Conversely, approaches employing biochemical readouts to detect particular reaction types, such as dephosphorylations[27], could screen more proteins but by design were focused on only detecting specific enzyme classes. Furthermore, the use of commercially available substrate mixtures[27] or technical-grade metabolite extracts[22] precluded the identification of reactions involving commercially unavailable, sensitive or yet-undiscovered compounds. Finally, the requirement for purified soluble proteins made membrane-localized or unstable enzymes inherently difficult to discover.

Here we develop an approach based on nontargeted metabolomics that combines high throughput with broad coverage for activity-based enzyme discovery at the proteome scale. We apply our method to screen all functionally uncharacterized *E. coli* proteins.

RESULTS

Method development and optimization

The basic principle of our method is the incubation of one purified protein or protein-overexpressing cell lysate in a metabolite

[1]Institute of Molecular Systems Biology, ETH Zürich, Zürich, Switzerland. [2]PhD Program on Systems Biology, Life Science Zürich, Zürich, Switzerland. [3]Present address: Cellzome, GlaxoSmithKline R&D, Heidelberg, Germany. Correspondence should be addressed to U.S. (sauer@ethz.ch).

RECEIVED 22 APRIL; ACCEPTED 19 OCTOBER; PUBLISHED ONLINE 12 DECEMBER 2016; DOI:10.1038/NMETH.4103

cocktail, followed by mass-spectrometric identification of metabolites with changing abundance (**Fig. 1a**). The metabolite cocktail should contain chemically diverse yet biologically relevant compounds as candidate substrates for enzymatic reactions. Because of the lack of commercial availability and because mixtures such as yeast extract are depleted for unstable but potentially relevant compounds, we prepared a concentrated *E. coli* metabolome extract under conditions that largely preserved natural metabolic diversity. Specifically, we cultivated *E. coli* BW25113 cells in two separate aerobic fermentations with either defined glucose minimal medium or complex amino acid–glycerol medium and extracted soluble metabolites (**Fig. 1a**). Subsequently, the two extracts were pooled to increase metabolite diversity and adjusted to approximately 0.1-fold of the *in vivo E. coli* metabolome concentration (**Supplementary Fig. 2**). Since *in vivo* substrate concentrations often exceed the K_m values of their cognate enzymes more than ten-fold[28], this concentration was sufficient to assure detectable activity of most enzymes. Using flow injection time-of-flight mass spectrometry[29], we detected 4,720 mass-to-charge features (ions) in the pooled extracts, 777 of which were putatively annotated as 962 unique metabolites by matching their accurate masses with sum formulas of compounds in the Kyoto Encyclopedia of Genes and Genomes (KEGG) *E. coli* metabolite database (**Supplementary Tables 1** and **2**)[30]. The number of putatively annotated metabolites represents a potential metabolome coverage of up to 36% (**Supplementary Fig. 3**), with the inherent limitation that in-source fragments are not accounted for[31], and isobaric compounds cannot be distinguished[29].

To obtain purified proteins, we grew individual strains of a genome-wide *E. coli* protein expression library (the ASKA collection[32]) in 96-well plates, lysed the cells and purified overexpressed proteins using 96-well-format His-tag technology (**Fig. 1a** and **Supplementary Table 3**). Since purification of active enzyme forms is not guaranteed in all cases, such as in the cases of heterocomplexes or membrane-bound enzymes, we also assayed cell lysates of the expression strains. Purified proteins or cell lysates were incubated in the metabolite cocktail for 30 min at 22 °C to allow for potential enzyme catalysis (**Fig. 1a**). Accumulated or depleted compounds were detected by nontargeted flow injection time-of-flight mass spectrometry[29], enabling the detection of hundreds of chemically diverse metabolites at a throughput of 1 min per injection.

Method validation and screen of 1,275 uncharacterized proteins

We screened 1,275 uncharacterized *E. coli* proteins for enzymatic activity in purified form and in crude lysates of the *E. coli* overexpression strain in pentuplicates. To validate our method, we included an additional 189 known enzymes that represented the distribution of reaction mechanisms and biological pathways of all *E. coli* enzymes. The resulting data set is an abundance matrix of 4,720 detected metabolite ions in 14,670 assays. After correcting for instrument drift and filtering of spurious systematic and technical errors (**Supplementary Fig. 4**), *Z*-score standardization was applied to quantify the abundance of each ion in a given enzyme assay relative to its abundance and s.d. across all other assays. The data are exemplarily visualized in **Figure 1b** for the ion *m*/*z* 323.029, tentatively annotated as uridine 5′-monophosphate (UMP). As expected, this ion accumulated in the control assay of uridine kinase, an enzyme of known function. Additionally, UMP was considerably depleted in the assay of YgdH, a conserved protein of unknown function; and UMP was accumulated in the assay of YgjP, a predicted metal-dependent hydrolase, suggesting that YgdH and YgjP are novel UMP-consuming and UMP-forming enzymes, respectively (see below).

To systematically separate enzymatic-activity-based ion responses from random artifacts, we determined an empirical *Z*-score cutoff based on the 121 purifiable known enzymes in our data set. Specifically, we first removed all interactions with an absolute *Z*-score < 2 to exclude statistical noise, and we subsequently determined the *Z*-score at which the product of recall (defined as the recovered fraction of known enzyme–reactant pairs) and precision (defined as the ratio of known to unknown recovered enzyme–reactant pairs) was maximal (**Fig. 1c**). Based on this procedure, we defined metabolite ions with *Z*-score < −5 and *Z*-score > 5 as potential reaction substrates and products, respectively. At this cutoff, the false-positive rates for pure enzyme assays and cell lysate assays were 5.8% and 1.5% at true-positive rates of 74.6% and 41.6%, respectively (**Fig. 1c**). As expected, known enzyme–reactant pairs were better recovered by pure enzyme assays than by cell lysate assays, as judged by the areas under the receiver operating characteristic curves of 0.89 for pure enzyme and 0.79 for cell lysate assays (**Fig. 1c**).

For about two-thirds of the known enzymes we could not detect *in vitro* activity; this can be explained by the requirements for additional proteins in heteromultimeric complexes, inactive protein, absence of reactants in our cocktail, or metabolite concentrations too close to thermodynamic equilibrium for the reaction to proceed (**Fig. 1d**). These proteins lacking *in vitro* activity for known reasons are nonetheless included in subsequent analyses to obtain realistic estimates for uncharacterized proteins for which this information is not readily available. For 34 of the 121 purifiable known enzymes, we detected differential metabolite ions passing the stringent *Z*-score cutoff. For 4 enzymes only unknown reactants, for 8 enzymes both known and unknown reactants, and for 22 enzymes only known reactants were detected (**Fig. 1d** and **Supplementary Table 4**). Based on the 34 enzymes with detectable differential metabolite ions, this demonstrates that our discovery method has a precision of 88% and a recall of 25%. Among the 30 correctly recovered enzymes, 6 were detected only by pure enzyme assays, 11 only by cell lysate assays and 13 by both assays (**Fig. 1e**), indicating that the two assay types are complementary with a notable overlap. For 9 enzymes we detected both substrates and products, and for the remaining 21 enzymes we detected either only substrates or only products (**Fig. 1f**). The distributions of EC numbers and metabolic pathway categories assigned to the recovered enzymes matched the distributions of all enzymes in *E. coli* (**Fig. 1g**,**h**), demonstrating that our method can detect enzymes with different catalytic mechanisms and diverse biological roles without strong bias. Isomerization reactions were a notable exception, as they could not be detected because they do not involve mass shifts between substrates and products.

Discovery of novel enzymes and assignment of catalyzed reactions

For 241 of the 1,275 functionally uncharacterized proteins, at least one metabolite ion passed the stringent |*Z*-score| > 5.0 cutoff (**Supplementary Data Sets 1** and **2**), providing evidence for enzymatic activity (**Fig. 2a**,**b** and **Supplementary Table 5**).

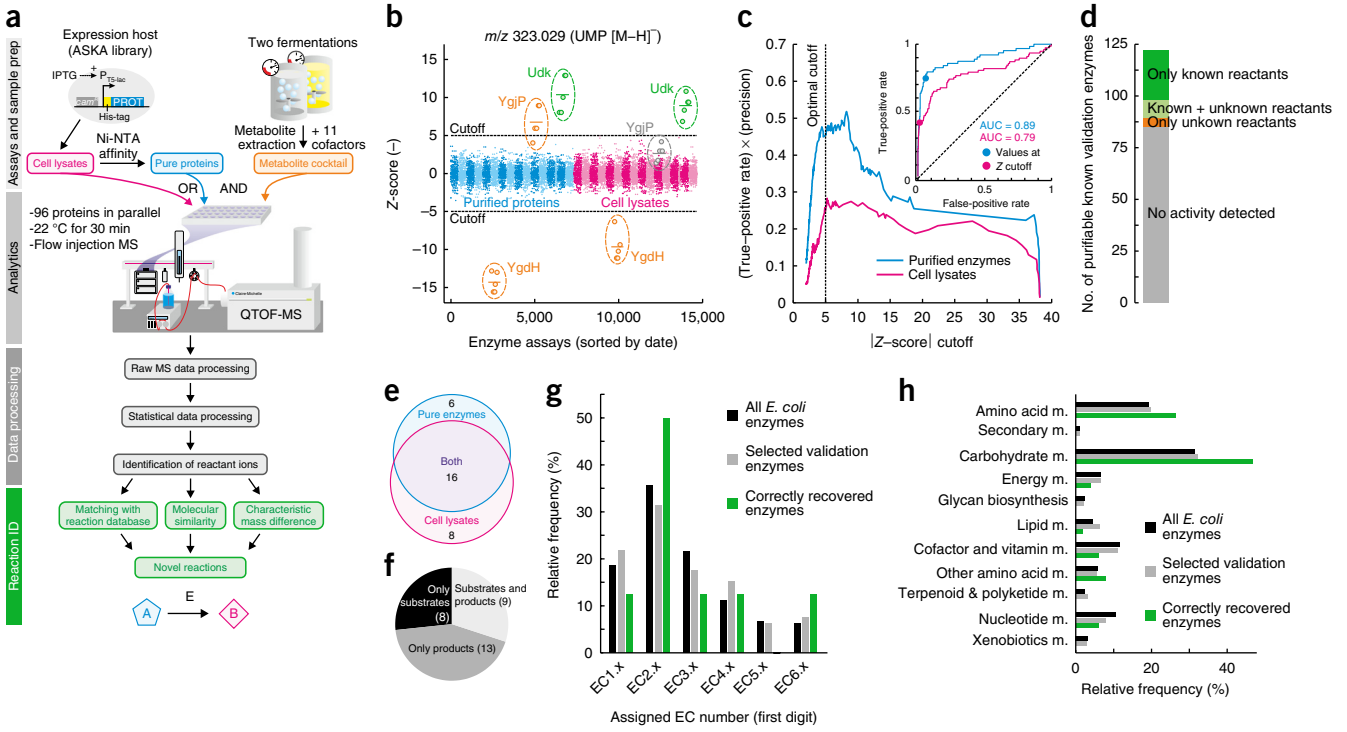

Figure 1 | Enzyme discovery by nontargeted metabolomics. (**a**) Experimental setup. Purified proteins or the lysate of their overexpression strains were incubated in a cocktail consisting of concentrated *E. coli* metabolome extracts and general enzyme cofactors. Differential metabolite ions were subsequently detected by nontargeted flow injection mass spectrometry[29] followed by data processing (**Supplementary Fig. 4**) and reaction identification using different database and chemoinformatics approaches. Ni-NTA, nitrilotriacetic acid; QTOF-MS, quadrupole time-of-flight mass spectrometry. (**b**) Abundance of the metabolite ion *m/z* 323.029, annotated as uridine 5′-monophosphate (UMP). A *Z*-score cutoff of 5.0 was determined to identify changes due to enzymatic reactions. Pentuplicate assays of uridine kinase (Udk), a known enzyme producing UMP, and the two uncharacterized proteins YgdH and YgjP predicted to enzymatically act on UMP, are highlighted. Horizontal lines between pentuplicate values obtained from individual assays represent their means. Data sets of pure enzyme and cell lysate assays are distinguished by coloring. (**c**) Empirical *Z*-score cutoff determination based on 189 known enzymes included in the screen. The inset shows the receiver operating curves of the *Z*-scores in both assays, the areas under the curves (AUC) and the true-positive and false-positive rates at the *Z*-score cutoff. (**d**) Recovery analysis of known enzymes after applying the *Z*-score cutoff, both assays combined. (**e**) Known enzymes recovered correctly (at least one known reactant) using purified proteins or cell lysates. (**f**) Enzymes recovered correctly by change direction of recovered reactants. (**g**) Histogram of Enzyme Commission (EC) numbers associated with correctly recovered enzymes. (**h**) Histogram of metabolic functions (based on KEGG pathway definitions[30]) associated with correctly recovered enzymes. M., metabolism.

We excluded presumably unspecific metabolite ions that were affected by more than five putative enzymes to focus on the most characteristic interactions. Most of the 241 putative enzymes appeared to be specific, as they affected only a few metabolite ions, but 15 enzymes caused abundance changes in more than five ions, suggesting they might catalyze multiple reactions (**Supplementary Fig. 5**). The annotated differential metabolites were chemically diverse and included amino acids, nucleotides, carbohydrates and vitamins (**Fig. 2c**), indicating widespread metabolic functions of putative enzymes.

Annotated differential metabolite ions were promising starting points for assigning metabolic reactions. To identify reactions' substrate–product pairs, we used a relaxed *Z*-score cutoff such that reactant pairs consisting of at least one high-confidence metabolite ion ($|Z\text{-score}| > 5$) and at most one lower confidence metabolite ion ($3 < |Z\text{-score}| < 5$) could be obtained. For each putative enzyme, we then matched all combinations of annotated substrate and product metabolites with the KEGG main reactant pair database[30] that lists pairs of compounds known to participate in common reactions. In total, we found 40 known reactant pairs associated with nine putative enzymes (**Supplementary Table 6**). For example, we predicted (and later confirmed; see below) that YgdH would convert UMP to uracil, as both metabolites form the KEGG reactant pair rp:RP01202 part of the reaction EC 2.4.2.9.

Additionally, to identify reactant pairs not listed in the KEGG database, we computed the molecular similarity between all possible substrate–product pairs of each enzyme using the graph-based SIMCOMP2 algorithm that determines the largest overlapping moieties between two molecules[33,34]. High molecular similarity can suggest possible reactant pairs because most enzymes perform only minor modifications to the chemical backbones of their substrates, and indeed known substrate–product pairs are strongly enriched for highly similar compounds (**Supplementary Fig. 6**). In total, 133 similarity-based potential reactant pairs associated with 23 putative enzymes were identified (**Supplementary Table 7**). For example, we predicted (and later confirmed; see below) that YbiC would convert 2-oxoglutarate to 2-hydroxyglutarate, as both compounds have a molecular similarity of 91%.

As a separate approach, we matched the mass differences among all possible substrate and product ions with a list of mass differences associated with known reactions (**Supplementary Fig. 7**). In total, we thereby discovered 63 mass-difference-based potential reactant pairs associated with 15 putative enzymes (**Supplementary Table 8**). The advantage of this approach is that

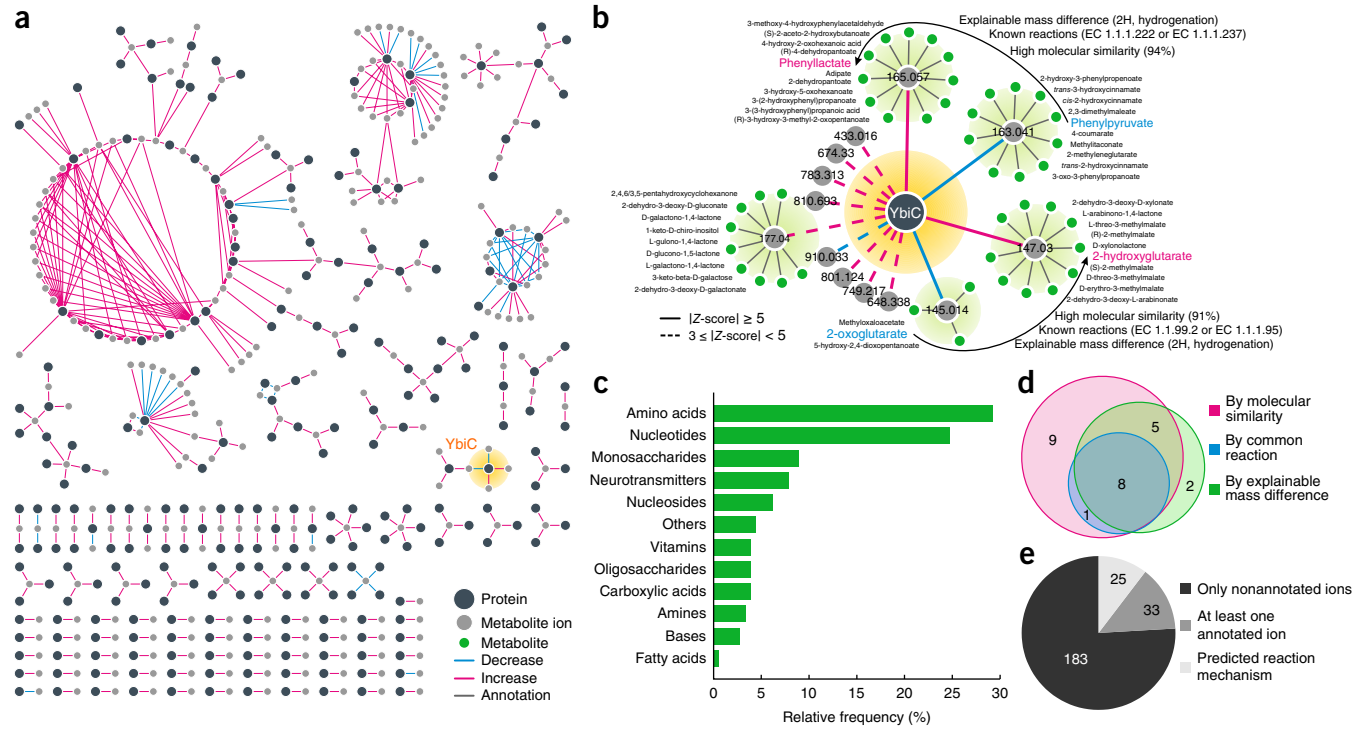

Figure 2 | Nontargeted metabolomics reveals numerous novel enzymes in *E. coli*. (**a**) Network representation of 485 detected associations between 241 functionally uncharacterized proteins and 258 metabolite ions with an absolute *Z*-score > 5 representing predicted enzyme–reactant pairs. Associations detected by pure enzyme assays and cell lysate assays were combined. The underlying data are provided in **Supplementary Table 5**. (**b**) Zoom into the associations detected for the YbiC protein visualizing potential ambiguity of accurate-mass-based metabolite annotation. To systematically identify reactant pairs among hit metabolites, three complementary approaches were used. All three approaches independently identified the reactant pairs [phenylpyruvate → phenyllactate] and [2-oxoglutarate → 2-hydroxyglutarate], thus predicting YbiC to be a multispecific 2-hydroxycarboxylic acid dehydrogenase, which we confirmed in subsequent validation experiments. (**c**) Chemical classifications of annotated hit metabolite ions. (**d**) Overlap of the three reactant pair prediction approaches. Numbers indicate putative novel enzymes with predicted reactant pairs. (**e**) Summary of available evidence for the 241 putative enzymes.

it is also applicable to the numerous nonannotated differential ions that are typically observed in nontargeted mass spectrometry assays[35]. For example, we predicted that YgjP, a novel uridine 5′-triphosphate pyrophosphatase, would additionally dephosphorylate an unknown metabolite with *m/z* 427.020, yielding a likewise unknown metabolite with *m/z* 347.051 because the *m/z* difference is consistent with the loss of one phosphate group. As the number of annotated metabolites increases, our data set can be revisited to identify further orphan enzyme functions from presently nonannotated ions. Reassuringly, our three approaches to reaction prediction provide overlapping results (**Fig. 2d**), increasing confidence for the postulated reactions.

In summary, our screen identified 241 functionally uncharacterized proteins as putative metabolic enzymes. For 33 of these we identified annotated metabolite ions but only one reaction partner, thus providing specific starting points for further enzyme identification. For 25 predicted enzymes we were able to postulate specific substrate–product pairs and hence reaction mechanisms that can be readily validated (**Fig. 2e**).

Biochemical and functional validation of 12 novel enzymes

To verify our reaction predictions, we selected 16 diverse enzymes for which the reaction substrates were commercially available. C-terminally His_6-tagged proteins[32] were purified and then incubated with predicted pure substrates, and reaction progression was monitored by mass spectrometry. For 4 out of the 16 putative enzymes we were unable to confirm predicted activities (listed in **Supplementary Table 9**), likely reflecting the inherent uncertainty in assigning reactions based on ambiguously annotated ions.

We successfully confirmed that the remaining 12 enzymes catalyze 29 unique metabolic reactions (**Fig. 3**, **Table 1** and **Supplementary Figs. 8–19**), representing an accuracy of 75% that is generally consistent with the 88% accuracy we achieved for known enzymes. To further assess the *in vivo* relevance of the predicted and validated novel enzymes, we performed metabolomics analyses of 223 viable single-gene-deletion mutants (obtained from the KEIO collection[36]) grown exponentially in glucose minimal medium supplemented with casein hydrolysate. In 18 mutants, the substrates or products of deleted enzymes were among the 5% most strongly altered metabolite ions, consistent with *in vivo* activity of these enzymes under this condition (**Supplementary Table 10** and **Supplementary Data Set 3**). Moreover, we determined maximum specific growth rate, growth yield and lag phase for all 12 deletion mutants of experimentally validated novel enzymes under environmental conditions in which we anticipated the catalyzed reactions to be relevant for optimal growth physiology (**Supplementary Table 11**). For the Δ*yfbT*, Δ*yncA* and Δ*ygdH* mutants we found that dephosphorylation of hexitols by YfbT contributes to butanol tolerance, acetylation of L-phenylglycine by YncA contributes to

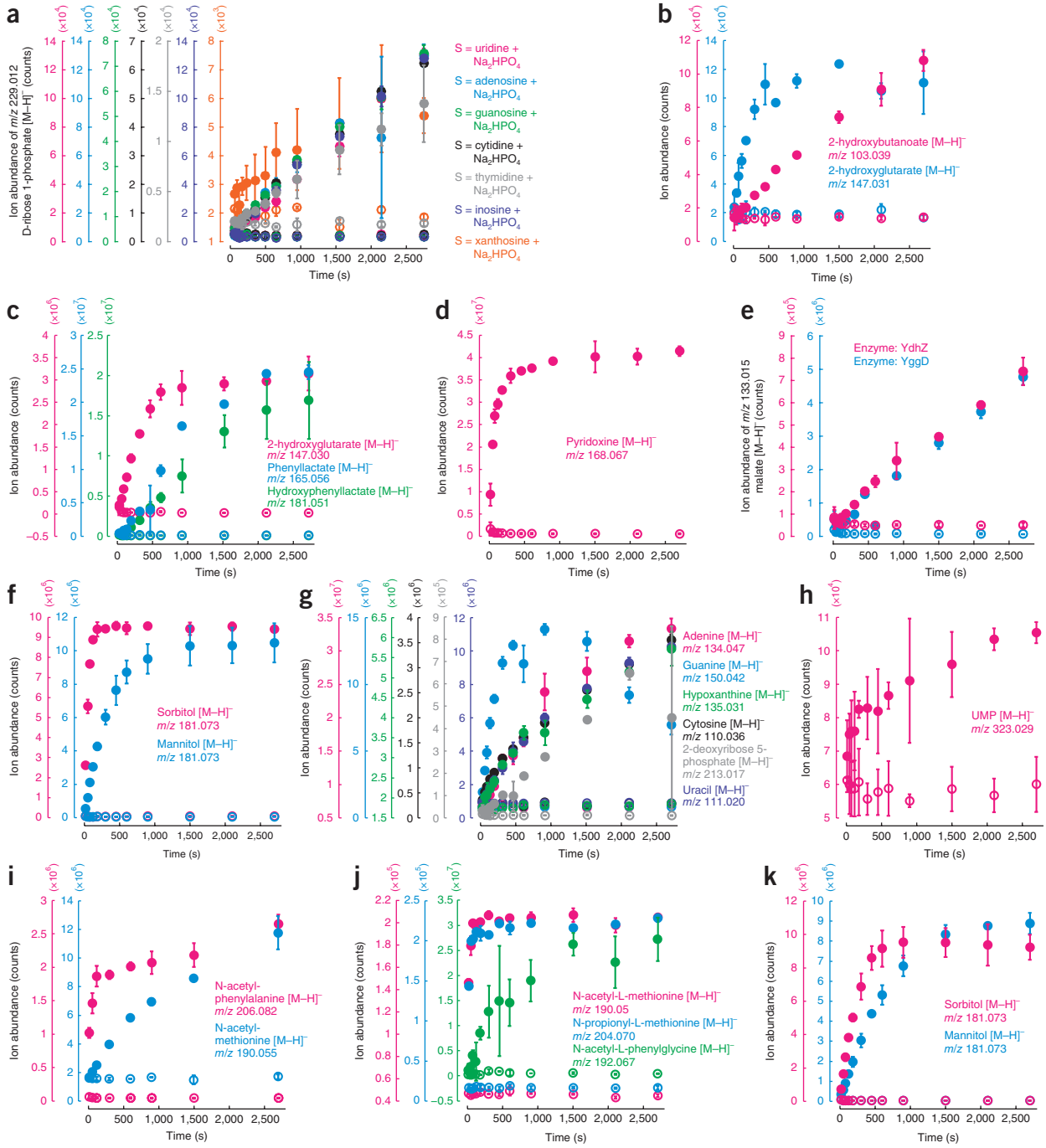

Figure 3 | Functional validation of 12 diverse novel metabolic enzymes. All panels show timecourses of indicated products of selected enzymatic reactions performed with 10 mM of pure substrates. For panels **a**–**h** and **j**–**k**, closed symbols represent data of assays with active enzymes, and open symbols represent data of control assays with heat-inactivated enzymes. Additional data are provided in **Table 1** and **Supplementary Figures 8**–**19**. Data shown as mean and s.d. of three individual assays. (**a**) YaiE, phosphorolysis of diverse nucleosides, yielding D-ribose 1-phosphate and free bases. S, substrate. (**b**) YbdH, NADPH-dependent reduction of 2-oxobutanoate and 2-oxoglutarate to 2-hydroxybutanoate and 2-hydroxygultarate, respectively. (**c**) YbiC, NADPH-dependent reduction of 2-oxoglutarate, phenylpyruvate and 3-(4-hydroxyphenyl)pyruvate to 2-hydroxyglutarate, phenyllactate and 3-(4-hydroxyphenyl)lactate, respectively. (**d**) YdbC, NADPH-dependent reduction of 4-pyridoxal to 4-pyridoxine. (**e**) YdhZ and YggD, fumarate hydratases. (**f**) YfbT, previously reported to have phosphatase activity[27], additionally dephosphorylates sorbitol 6-phosphate and mannitol 1-phosphate to sorbitol and mannitol, respectively. (**g**) YgdH, a conserved protein. Hydrolysis of the N-glycosidic bond of AMP, GMP, IMP, CMP, and UMP, yielding the respective free bases adenine, guanine, hypoxanthine, cytosine and uracil as well as D-ribose 5′-phosphate. For the cleavage of dTMP, 2-desoxyribose 5′-phosphate is shown. (**h**) YgjP, hydrolysis of UTP to UMP and pyrophosphate. (**i**) YhhY, acetyl-coenzyme A-dependent N-acetylation of L-phenylalanine and L-methionine, yielding N-acetyl-L-phenylalanine and N-acetyl-L-methionine, respectively. The YhhY enzyme could not be purified in soluble form; instead the cell lysates of its overexpression strain (closed symbols) and its deletion mutant (open symbols) were assayed. (**j**) YncA, a sulfur-containing amino acid N-acyl-transferase[48]. Acetyl-coenzyme A-dependent N-acetylation of L-methionine and L-phenylglycine yielding N-acetyl-L-methionine and N-acetyl-L-phenylglycine, and propionyl-coenzyme A-dependent N-propionylation of L-methionine yielding N-propionyl-L-methionine. (**k**) YniC, previously reported to have phosphatase activity[27]. Dephosphorylation of sorbitol 6-phosphate and mannitol 1-phosphate to sorbitol and mannitol, respectively.

Table 1 | Reactions catalyzed by 12 functionally validated novel metabolic enzymes

Gene	Current annotation	Catalyzed reaction(s) identified in this study*	EC number(s)	Proposed new annotation
b0391, ***yaiE***	Conserved protein	Uridine + orthophosphate ↔ D-ribose 1-phosphate + uracil	EC 2.4.2.3	*ppnP*, pyrimidine and purine nucleoside phosphorylase
		Adenosine + orthophosphate ↔ D-ribose 1-phosphate + adenine	EC 2.4.2.1	
		Guanosine + orthophosphate ↔ D-ribose 1-phosphate + guanine	EC 2.4.2.15	
		Cytidine + orthophosphate ↔ D-ribose 1-phosphate + cytosine	EC 2.4.2.2	
		Thymidine + orthophosphate ↔ D-ribose 1-phosphate + thymine	EC 2.4.2.4	
		Inosine + orthophosphate ↔ D-ribose 1-phosphate + hypoxanthine	EC 2.4.2.1	
		Xanthosine + orthophosphate ↔ D-ribose 1-phosphate + xanthine	EC 2.4.2.1	
b0599, ***ybdH***	Predicted oxidoreductase	2-oxobutanoate + NADPH + H^+ → 2-hydroxybutanoate + $NADP^+$ [a,#]	EC 1.1.1.27	*hcxA*, 2-hydroxycarboxylic acid dehydrogenase A
		2-oxoglutarate + NADPH + H^+ → 2-hydroxyglutarate + $NADP^+$ [a,#]	EC 1.1.99.2 or EC 1.1.99.39	
b0801, ***ybiC***	Predicted dehydrogenase	2-oxoglutarate + NADH + H^+ → 2-hydroxyglutarate + NAD^+ [b,#]	EC 1.1.99.2 or EC 1.1.99.39	*hcxB*, 2-hydroxycarboxylic acid dehydrogenase B
		2-oxoglutarate + NADPH + H^+ → 2-hydroxyglutarate + $NADP^+$ [b,#]	EC 1.1.99.2 or EC 1.1.99.39	
		Phenylpyruvate + NADH + H^+ → phenyllactate + NAD^+ [#]	EC 1.1.1.222 or EC 1.1.1.237	
		Phenylpyruvate + NADPH + H^+ → phenyllactate + $NADP^+$ [#]	EC 1.1.1.222 or EC 1.1.1.237	
		3-(4-hydroxyphenyl)pyruvate + NADH + H^+ → 3-(4-hydroxyphenyl)lactate + NAD^+ [#]	EC 1.1.1.222 or EC 1.1.1.237	
		3-(4-hydroxyphenyl)pyruvate + NADPH + H^+ → 3-(4-hydroxyphenyl)lactate + $NADP^+$ [#]	EC 1.1.1.222 or EC 1.1.1.237	
b1406, ***ydbC***	Predicted oxidoreductase, NAD(P)-binding	4-pyridoxal + NADH + H^+ → 4-pyridoxine + NAD^+ [b,c]	EC 1.1.1.65	*pdxI*, pyridoxine dehydrogenase
		4-pyridoxal + NADPH + H^+ → 4-pyridoxine + $NADP^+$ [b,c]	EC 1.1.1.65	
b1675, ***ydhZ***	Predicted protein	Fumarate + H_2O → malate	EC 4.2.1.2	*fumD*, fumarase D
b2293, ***yfbT***	Sugar phosphatase	D-sorbitol 6-phosphate + H_2O → D-sorbitol + orthophosphate[d]	EC 3.1.3.50	*hxpA*, hexitol phosphatase A
		D-mannitol 1-phosphate + H_2O → D-mannitol + orthophosphate[d]	EC 3.1.3.22	
b2795, ***ygdH***	Conserved protein, UPF0053 family	Adenosine 5′-monophosphate + H_2O → adenine + D-ribose 5′-phosphate	EC 3.2.2.1	*ppnN*, pyrimidine and purine nucleotide 5′-monophosphate nucleosidase
		Guanosine 5′-monophosphate + H_2O → guanine + D-ribose 5′-phosphate	EC 3.2.2.1	
		Inosine 5′-monophosphate + H_2O → hypoxanthine + D-ribose 5′-phosphate	EC 3.2.2.1	
		Cytidine 5′-monophosphate + H_2O → cytosine + D-ribose 5′-phosphate	EC 3.2.2.10	
		2-deoxythymidine 5′-monophosphate + H_2O → thymine + 2-deoxy-D-ribose 5′-phosphate	EC 3.2.2.10	
		Uridine 5′-monophosphate + H_2O → uracil + D-ribose 5′-phosphate	EC 3.2.2.10	
b2929, ***yggD***	Predicted DNA-binding transcriptional regulator	Fumarate + H_2O → malate	EC 4.2.1.2	*fumE*, fumarase E
b3085, ***ygjP***	Predicted metal-dependent hydrolase	Uridine 5′-triphosphate + H_2O → uridine 5′-monophosphate + pyrophosphate[e]	EC 3.6.1.19	*upp*, UTP pyrophosphatase
b3441, ***yhhY***	Predicted acetyltransferase	L-phenylalanine + acetyl-CoA → N-acetyl-L-phenylalanine + coenzyme A [f,g,#]	EC 2.3.1.53	*aaaT*, L-amino acid N-acetyltransferase
		L-methionine + acetyl-CoA → N-acetyl-L-methionine + coenzyme A [f,#]	EC 2.3.1.1	
b1448, ***yncA***	Predicted acyltransferase with acyl-CoA N-acyltransferase domain	L-methionine + acetyl-CoA → N-acetyl-L-methionine + coenzyme A [h,#]	EC 2.3.1.1	*mnaT*, L-methionine N-acyltransferase[48]
		L-methionine + propionyl-CoA → N-propionyl-L-methionine + coenzyme A [h,#]	EC 2.3.1.1	
		L-phenylglycine + acetyl-CoA → N-acetyl-L-phenylglycine + coenzyme A [f,#]	EC 2.3.1.1	
b1727, ***yniC***	2-deoxyglucose 6-phosphate phosphatase	D-sorbitol 6-phosphate + H_2O → D-sorbitol + orthophosphate[d]	EC 3.1.3.50	*hxpB*, hexitol phosphatase B
		D-mannitol 1-phosphate + H_2O → D-mannitol + orthophosphate[d]	EC 3.1.3.22	

*Data of pure substrate enzyme assays are shown in **Supplementary Figures 8–19**.
[a]NADH not accepted as cofactor. [b]Reaction confirmed to be irreversible. [c]4-pyridoxic acid and 4-pyridoxal phosphate are not accepted as substrates. [d]Previously reported alternative dephosphorylation reactions[27] were not detected. [e]Inosine 5′-triphosphate and adenosine 5′-triphosphate were not accepted as substrates. [f]Propionyl-CoA, succinyl-CoA and (S)-methylmalonyl-CoA were not accepted as acyl donors. [g]L-tyrosine was not accepted as alternative substrate to L-phenylalanine. [h]Activity previously reported in patent application[48].
[#]Reaction represents novel gap in metabolic network model (at least one metabolite is not connected to any other reaction).

L-phenylglycine utilization and YgdH is beneficial for optimal growth on glucose, presumably by contributing to nucleoside pool homeostasis (**Supplementary Fig. 20**).

DISCUSSION

Our methodology allowed us to predict 241 and validate 12 novel enzymes, 3 of which have particularly intriguing functions. The multispecific hydrolase YgdH catalyzes the long missing cleavage of CMP to cytosine and ribose 5-phosphate. Furthermore, the two dehydrogenases YbdH and YbiC catalyze, among other reactions, the irreversible reduction of 2-oxoglutarate to 2-hydroxyglutarate, a metabolite implicated in human neurological disorders[37] and cancer[38]. 2-hydroxyglutarate is thought to be accidentally synthesized in humans[38] and *E. coli*[39], and both species are equipped with specific enzymes to detoxify 2-hydroxyglutarate by oxidation to the nontoxic metabolite 2-oxoglutarate[37,39]. Our unexpected discovery of two enzymes that appear to specifically produce but not degrade 2-hydroxyglutarate indicates that this compound may yet have additional cellular roles.

Our pipeline for unbiased enzyme annotation has a recall of 25%, but its predictions are often correct with a precision in the range of 80%. One limitation of our pipeline is its poor recovery of purified or overexpressed enzymes. For overexpressed enzymes, accumulating products may be immediately consumed by downstream enzymes, preventing their detectable accumulation. Additionally, some reactions may not proceed *in vitro* on account of thermodynamic equilibration or low concentration of substrates and products. This could be addressed by spiking the metabolome cocktail with a diverse mixture of synthetic metabolites or combining extracts from species growing in vastly different environments with different thermodynamic equilibria[40]. Another limitation is that our approach does not allow the determination of absolute reaction rates, but rather it suggests the catalytic potential of enzymes that may serve as starting points for molecular evolution[41] or could be exploited in biotechnology[42]. Finally, the unambiguous prediction of catalyzed reactions is complicated by a large proportion of nonannotated ions and the inherent ambiguity in assigning metabolite names to ions based on accurate mass alone. Generally, we believe these limitations to be acceptable for a discovery method in return for a broad metabolome coverage that few other analytical techniques can currently provide[14,43]; and we were able to reduce ambiguity considerably by implementing reaction-prediction algorithms. Nonetheless, more comprehensive metabolome databases, advances in algorithms for reaction prediction from metabolomics data, structure elucidation approaches[44,45] and computational frameworks to integrate other information sources[14,46] such as protein structures could be exploited in the future to interrogate our data set to identify currently obscured novel enzymes and reactions.

A major advantage of our method is its direct experimental applicability to large sets of candidate proteins or complex biological samples without a priori definition of compounds or reactions of interest. This nontargeted, data-driven approach allows for high-content screens even of poorly characterized samples, for instance, metagenomic expression libraries[47]. Since large numbers of assays can be performed and statistically analyzed, even small reactant concentration differences translate into statistically significant signals, allowing for the robust detection of reactions involving low-abundance metabolites. Our comparative data analysis approach also eliminates the need for highly purified protein preparations, since effects of frequent impurities or reactions catalyzed by general contaminant enzymes are cancelled out, and only specific metabolite changes are reported. For sufficiently large sample cohorts, protein purification can be completely omitted without unacceptably sacrificing precision or recall as we demonstrated with our screen of overexpression cell lysates, thus considerably reducing costs and eliminating purification-inherent drawbacks.

We anticipate that our approach could be important for enzyme discovery in basic research and industrial applications such as high-content functional screening of metagenome libraries or design of industrial bioproduction strains. This method is generally applicable to any purified protein or crude cell lysate of its overexpression host and enables performing up to 1,200 nontargeted enzyme assays per working day, thus enabling the functional screen of entire proteomes within a few weeks. The surprisingly large numbers of identified novel enzymes and potentially undiscovered metabolites even in the well-characterized and relatively simple bacterium *E. coli* underscore that further elucidating network topologies remains an important goal in metabolism research.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Suter and J. Schmitz for technical assistance with validation enzyme assays. Funding was provided by the MetaNetX project of the Swiss Initiative for Systems Biology (SystemsX.ch; http://metanetx.org; evaluated by the Swiss National Science Foundation) and the Swiss Federal Government through the Federal Office of Education and Science.

AUTHOR CONTRIBUTIONS

D.C.S. and T.F. performed the experiments and analyzed the data. D.C.S., T.F. and N.Z. developed data analysis software and algorithms. D.C.S., N.Z. and U.S. designed the research and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Hanson, A.D., Pribat, A., Waller, J.C. & de Crécy-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 (2009).
2. Galperin, M.Y. & Koonin, E.V. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* **32**, 5452–5463 (2004).
3. Jaroszewski, L. *et al.* Exploration of uncharted regions of the protein universe. *PLoS Biol.* **7**, e1000205 (2009).
4. Sorokina, M., Stam, M., Médigue, C., Lespinet, O. & Vallenet, D. Profiling the orphan enzymes. *Biol. Direct* **9**, 10 (2014).
5. Chen, L. & Vitkup, D. Distribution of orphan metabolic activities. *Trends Biotechnol.* **25**, 343–348 (2007).
6. Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333**, 863–882 (2003).
7. Bork, P. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res.* **10**, 398–400 (2000).
8. Blaby-Haas, C.E. & de Crécy-Lagard, V. Mining high-throughput experimental data to link gene and function. *Trends Biotechnol.* **29**, 174–182 (2011).

9. Galperin, M.Y. Conserved 'hypothetical' proteins: new hints and new puzzles. *Comp. Funct. Genomics* **2**, 14–18 (2001).
10. Tipton, K. & Boyce, S. Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999). *Eur. J. Biochem.* **264**, 610–650 (1999).
11. Pouliot, Y. & Karp, P.D. A survey of orphan enzyme activities. *BMC Bioinformatics* **8**, 244 (2007).
12. Shearer, A.G., Altman, T. & Rhee, C.D. Finding sequences for over 270 orphan enzymes. *PLoS One* **9**, e97250 (2014).
13. Lespinet, O. Orphan enzymes? *Science* **307**, 42 (2005).
14. Sévin, D.C., Kuehne, A., Zamboni, N. & Sauer, U. Biological insights through nontargeted metabolomics. *Curr. Opin. Biotechnol.* **34**, 1–8 (2015).
15. Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L. & Palsson, B.Ø. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**, 129–143 (2009).
16. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337**, 1101–1104 (2012).
17. Kuznetsova, E. *et al.* Enzyme genomics: application of general enzymatic screens to discover new enzymes. *FEMS Microbiol. Rev.* **29**, 263–279 (2005).
18. Prosser, G.A., Larrouy-Maumus, G. & de Carvalho, L.P. Metabolomic strategies for the identification of new enzyme functions and metabolic pathways. *EMBO Rep.* **15**, 657–669 (2014).
19. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **8**, 995–1005 (2007).
20. Plata, G., Fuhrer, T., Hsiao, T.-L., Sauer, U. & Vitkup, D. Global probabilistic annotation of metabolic networks enables enzyme discovery. *Nat. Chem. Biol.* **8**, 848–854 (2012).
21. Zhao, S. *et al.* Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* **502**, 698–702 (2013).
22. Saito, N. *et al.* Metabolite profiling reveals YihU as a novel hydroxybutyrate dehydrogenase for alternative succinic semialdehyde metabolism in *Escherichia coli*. *J. Biol. Chem.* **284**, 16442–16451 (2009).
23. Notebaart, R.A. *et al.* Network-level architecture and the evolutionary potential of underground metabolism. *Proc. Natl. Acad. Sci. USA* **111**, 11762–11767 (2014).
24. Guzmán, G.I. *et al.* Model-driven discovery of underground metabolic functions in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **112**, 929–934 (2014).
25. Coelho, P.S. *et al.* A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins *in vivo*. *Nat. Chem. Biol.* **9**, 485–487 (2013).
26. Saito, N. *et al.* Metabolomics approach for enzyme discovery. *J. Proteome Res.* **5**, 1979–1987 (2006).
27. Kuznetsova, E. *et al.* Genome-wide analysis of substrate specificities of the *Escherichia coli* haloacid dehalogenase-like phosphatase family. *J. Biol. Chem.* **281**, 36149–36161 (2006).
28. Bennett, B.D. *et al.* Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.* **5**, 593–599 (2009).
29. Fuhrer, T., Heer, D., Begemann, B. & Zamboni, N. High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. *Anal. Chem.* **83**, 7074–7080 (2011).
30. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
31. Xu, Y.F., Lu, W. & Rabinowitz, J.D. Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. *Anal. Chem.* **87**, 2273–2281 (2015).
32. Kitagawa, M. *et al.* Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): unique resources for biological research. *DNA Res.* **12**, 291–299 (2005).
33. Hattori, M., Okuno, Y., Goto, S. & Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **125**, 11853–11865 (2003).
34. Hattori, M., Tanaka, N., Kanehisa, M. & Goto, S. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.* **38**, W652–W656 (2010).
35. da Silva, R.R., Dorrestein, P.C. & Quinn, R.A. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. USA* **112**, 12549–12550 (2015).
36. Yamamoto, N. *et al.* Update on the Keio collection of *Escherichia coli* single-gene deletion mutants. *Mol. Syst. Biol.* **5**, 335 (2009).
37. Struys, E.A. *et al.* Mutations in the D-2-hydroxyglutarate dehydrogenase gene cause D-2-hydroxyglutaric aciduria. *Am. J. Hum. Genet.* **76**, 358–360 (2005).
38. Dang, L. *et al.* Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739–744 (2009).
39. Linster, C.L., Van Schaftingen, E. & Hanson, A.D. Metabolite damage and its repair or pre-emption. *Nat. Chem. Biol.* **9**, 72–80 (2013).
40. Amend, J.P. & Shock, E.L. Energetics of overall metabolic reactions of thermophilic and hyperthermophilic Archaea and bacteria. *FEMS Microbiol. Rev.* **25**, 175–243 (2001).
41. Khersonsky, O. & Tawfik, D.S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
42. Nobeli, I., Favia, A.D. & Thornton, J.M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167 (2009).
43. Fuhrer, T. & Zamboni, N. High-throughput discovery metabolomics. *Curr. Opin. Biotechnol.* **31**, 73–78 (2015).
44. Dunn, W.B. *et al.* Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **9**, 44–66 (2013).
45. Li, L. *et al.* MyCompoundID: using an evidence-based metabolome library for metabolite identification. *Anal. Chem.* **85**, 3401–3408 (2013).
46. Nichols, R.J. *et al.* Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–156 (2011).
47. Lorenz, P. & Eck, J. Metagenomics and industrial applications. *Nat. Rev. Microbiol.* **3**, 510–516 (2005).
48. Figge, R., Barbier, G. & Bestel-Corre, G. Production of N-acylated sulphur-containing amino acids with microorganisms having enhanced N-acyltransferase enzymatic activity. US patent US20100047880 A1 (2010).

ONLINE METHODS

Media and bacterial cell cultivation. All *E. coli* strains used in this study were obtained from the ASKA collection[32]. These strains were derived from *E. coli* K-12 AG1 and harbored plasmid pCA24N conferring chloramphenicol resistance and carrying a His_6-tagged open reading frame under the control of an isopropyl-β-D-thiogalactoside (IPTG)-inducible T5-lac promoter. Cultures were inoculated from −80 °C glycerol stocks, and cultivation was performed in Luria–Bertani (LB) medium (10 g/L Bacto tryptone, 5 g/L Bacto yeast extract, 5 g/L NaCl, pH 7.4) in presence of 100 µM IPTG and 100 µM chloramphenicol for 16 h at 37 °C shaking at 300 r.p.m. For the screen, 96-well deepwell blocks with 1.5 mL medium/well were used, whereas for the validation experiments 500 mL baffled shake flasks with 50 mL medium were used.

Protein purification (enzyme screen). Cells were pelleted by centrifugation (10 min at 4,000*g*) and resuspended in 400 µL of lysis buffer (20 mM sodium phosphate, 1 mM $MgCl_2$, 20 mM imidazole, 500 mM NaCl, 1 mM phenylmethanesulfonyl fluoride, 2 mg/mL lysozyme (Fluka), 0.2 mg/mL DNAse I (Roche), pH 7.4). After incubation at 30 °C for 30 min, lysis was completed by three repeated cycles of freezing (−80 °C) and thawing (30 °C). 50 µL aliquots of the cell lysates were stored at −80 °C for the cell lysate activity assays. Remaining cell lysates were applied to 96-well format Ni^{2+}-charged nitrilotriacetic acid (Ni-NTA) affinity columns (GE Healthcare) and washed twice with 20 column volumes of washing buffer (20 mM sodium phosphate, 500 mM NaCl, 20 mM imidazole, pH 7.5). Proteins were eluted from the column with elution buffer (20 mM sodium phosphate, 500 mM NaCl, 500 mM imidazole, pH 7.5), which was subsequently replaced by storage buffer (2 mM Tris–HCl, 1 mM $MgCl_2$, pH 7.4) using 96-well ultrafiltration plates with 10 kDa molecular weight cutoff, and proteins were stored at 4 °C for at most 1 d before performing activity assays. For each purified protein, we recorded the growth of its expression strain, the concentration of purified protein and whether the solution had a yellow color (**Supplementary Table 3**). Four 96-well plates per workday can be easily handled in parallel, thus enabling a theoretical throughput of 384 purified proteins using this setup.

Quantification of cell growth and protein concentration. Cell growth was quantified by measuring the absorbance at 595 nm of the expression cultures before harvest. For this, 50 µL of overnight culture were diluted with 100 µL of saline (0.9% NaCl, 1 mM $MgCl_2$) in 96-well clear flat-bottom plates to obtain absorbance values of less than 1. Protein yield was quantified by adding 5 µL of purified protein solution to 150 µL of Bradford reagent (Bio-Rad), incubating for 10 min at room temperature and measuring the absorbances at 590 and 450 nm, respectively. The ratio of A_{590} over A_{450} linearly correlates with the protein concentration over a wide range[49]. Quantification was performed based on a dilution series of bovine serum albumin (BSA). We excluded poorly growing strains with a harvest optical density below 0.5 and purified proteins below 50 µg/mL (for purified protein assays only) from further analyses.

Fed-batch fermentations for metabolite extract preparation. The fermentation vessel used for cultivation of *E. coli* K-12 BW25113 had a nominal volume of 2.4 L (Bioengineering) and was equipped with a three-level radial flow impeller (d = 3 cm) set to 1,000 r.p.m. and aerated with sterile air at 1 L gas/L liquid/min (vvm) at an operating pressure of 1.2 bar. pH was set to 7.0 and controlled by addition of 28% (w/v) NH_4OH or 5 M HCl, and temperature was set to 37 °C. The initial medium for glucose minimal medium fermentation contained 5 g/L glucose, 14.8 g/L KH_2PO_4, 4.44 g/L $(NH_4)_2HPO_4$, 1.9 g/L sodium citrate, 1.33 g/L $MgSO_4$, 5 mg/L thiamine–HCl as well as 9 mL/L trace element solution (0.18 g/L $ZnSO_4 \cdot 7\ H_2O$, 0.12 g/L $CuCl_2 \cdot 2\ H_2O$, 0.12 g/L $MnSO_4 \cdot H_2O$, 0.18 g/L $CoCl_2 \cdot 6\ H_2O$)[50]. The initial medium for glycerol complex medium fermentation contained 3 g/L glycerol, 12 g/L Bactro tryptone, 24 g/L Bacto yeast extract, 2.2 g/L KH_2PO_4, 9.4 g/L K_2HPO_4. Polypropylene glycol was added as required to prevent foam formation. After inoculating 750 mL of initial medium with 50 mL of overnight preculture in LB medium, feed medium of increasing substrate concentration (same composition as initial medium, but with 10–25–100 g/L glucose for the glucose fermentation and 6–15–60 g/L glycerol for the glycerol fermentation) was added at a constant rate of 2 mL/min. The low-concentrated feed was added for a duration of 12 h, the medium-concentrated feed for another 4 h, and finally the high-concentrated feed until the fermentations were aborted at an optical cell density at 600 nm of 20. Fermentation profiles are shown in **Supplementary Figure 21**.

Preparative metabolite extraction. 50 mL aliquots of fermentation broth with an OD_{595} of 20 were transferred to 50 mL tubes directly from the fermenter and pelleted by fast centrifugation (1 min at 10,000*g* at 0 °C). Thus, each tube contained a cell pellet of 380 mg dry cell weight (DCW) using a coefficient of 0.38 mg DCW/OD/mL[51]. The supernatant was discarded, and the pellets were flash frozen in liquid nitrogen. Subsequently, each pellet was extracted with 20 mL of 60:40% (v/v) ethanol:water at 80 °C for 10 min with occasional vortexing. All extracts were pooled, and the samples were dried under vacuum using a SpeedVac (Christ). The extracted metabolites were resuspended in a volume of 600 µL H_2O per tube to attain physiological concentration, as each mg DCW initially corresponded to 1.63 µL of intracellular fluid[51]. Aliquots of the extracts of both fermentations were stored separately at −80 °C before further use.

Nontargeted enzyme activity assays. Each assay mixture contained 39 µL buffer (2 mM Tris–HCl, 2 mM $MgCl_2$, pH 7.5), 5 µL ten-fold concentrated cofactor mix (1 mM of each NADH, NAD^+, NADPH, $NADP^+$, FAD, ATP, GTP, ITP, *S*-adenosyl-L-methionine, acetyl-CoA, CoA), 2.5 µL of *E. coli* metabolite extract cultivated on glucose minimal medium and 2.5 µL of *E. coli* metabolite extract cultivated on glycerol complex medium per well in 96-well plates. Assays were started by adding 1 µL of either purified protein or 1 µL of overexpression strain lysate to each well, allowed to proceed for 1 h at 22 °C and quenched by adding −80 °C methanol to 75%-v/v. Each uncharacterized protein was analyzed in pentuplicates for both assay types.

Flow injection time-of-flight mass spectrometry. The analysis of quenched assays was performed on a platform consisting of a Hitachi L-7100 liquid chromatography pump coupled to a Gerstel MPS2 autosampler and an Agilent 6520 QTOF mass spectrometer (Agilent, Santa Clara, California) operated with published

settings[29]. The isocratic flow rate was 150 µL/min of mobile phase consisting of isopropanol:water (60:40, v/v) buffered with 5 mM ammonium fluoride at pH 9 for negative ionization mode. For online mass axis correction, 2-propanol (in the mobile phase), taurocholic acid and Hexakis(1H, 1H, 3H-tetrafluoropropoxy)–phosphazine were used. Mass spectra were recorded in profile mode from 50 to 1,000 *m/z* with a frequency of 1.4 spectra/s using the highest available resolving power (4 GHz HiRes). Source temperature was set to 325 °C, with 5 L/min drying gas and a nebulizer pressure of 30 psig. Fragmentor, skimmer and octupole voltages were set to 175 V, 65 V and 750 V, respectively.

Spectral data processing and ion annotation. All steps of mass spectrometry data processing and analysis were performed with Matlab (The Mathworks, Natick) using functions embedded in the Bioinformatics and Statistics toolboxes as previously described[29]. Briefly, peak detection was performed once for each sample on the total profile spectrum obtained by summing all single scans recorded over time and using wavelet decomposition as provided by the Bioinformatics toolbox. In this procedure, we applied a cutoff to filter peaks of less than 500 ion counts (in the summed spectrum) to avoid the detection of artifacts. Centroid lists from samples were then merged into a single matrix by binning the accurate centroid masses within the tolerance given by the instrument resolution (about 0.002 amu at *m/z* 300). The resulting matrix lists the intensity of each mass peak in each analyzed sample. An accurate common *m/z* was recalculated with a weighted average of the values obtained from independent centroiding. Because mass axis calibration was applied online during acquisition, no *m/z* correction was applied during processing to correct for potential drifts. After merging, negatively charged ions were tentatively annotated based on accurate mass using 0.001 Da tolerance (**Supplementary Table 1**). Annotation was based on the assumption that $[M - H]^-$ and $[M + F]^-$ are the possible and dominant ionization options for negative mode. The following isotopes were considered: $^{12}C- > ^{13}C$. The following neutral electrospray adducts were considered: exchange of H to K (H– > K), H to Na (H– > Na) and addition of sodium chloride (+NaCl). Ions and adducts were chosen based on occurrence frequency determined previously during method development[29] and are tentative in nature. We then correlated the abundance of nonannotated ions with the abundance of annotated ions across all data sets and, in case of a positive correlation (Pearson's $R > 0.8$), transferred this annotation. Correlation-based annotation is highlighted by [CORR] in **Supplementary Table 2**.

Data processing and standardization. The data set consisting of a matrix of intensities values for 4,720 detected ions in 14,670 assays was processed the following way using Matlab code (available upon request): (i) median filter subtraction to correct for instrument sensitivity drifts over the course of the measurements; (ii) *Z*-score (equation (1)) standardization along the second dimension (data sets) to have comparable numerical abundance values for all ions; (iii) *Z*-score standardization along the first dimension (ions) to correct for systematic sample abundance variations; (iv) median filtering of the five replicates of each enzyme assay to correct technical outliers; and (v) final *Z*-score standardization along the second dimension (data sets) to rescale ion abundance values. The effect of each data processing step is further explained and visualized in **Supplementary Figure 4**. This workflow was established based on our experience with handling such large metabolomics data sets and a good understanding of different sources of technical bias in our particular setup. When applied to data sets generated by different instrumentation, careful validation of each data processing step is advised to ensure desired performance.

$$Z_{\text{ion } i \text{ in assay } a} = \frac{\text{mean}_{i \text{ in 5 replicates of } a} - \text{mean}_{i \text{ in all assays}}}{\text{standard deviation}_{i \text{ across all assays}}} \quad (1)$$

Protein purification (validation experiments). Cells were pelleted by centrifugation (10 min at 4,000*g*), resuspended in 4 mL of lysis buffer (100 mM Tris–HCl, 5 mM $MgCl_2$, 2 mM dithiothreitol, 4 mM phenylmethanesulfonyl fluoride, pH 7.5) and cooled on ice. Lysis was performed by three passages through a FrenchPress system (Simo Aminco) with pressure cells precooled at 4 °C and operated at 1,000 psig. Cell lysates were applied to Co^{2+}-charged TALON affinity columns (0.5 mL column volume, GE Healthcare) providing higher purity than conventional Ni^{2+}-charged resins, and washed twice with 20 column volumes of washing buffer (20 mM sodium phosphate, 500 mM NaCl, 20 mM imidazole, pH 7.5). Subsequently, step gradient elution with elution buffer (20 mM sodium phosphate, 500 mM NaCl, pH 7.5) containing 60, 100 and 500 mM imidazole was performed; and the fractions containing the target protein (as identified by SDS–PAGE) were selected for further purification. Buffer of the selected fractions was exchanged to 10 mM Tris–HCl pH 7.5 by three ultrafiltration steps using spin columns with 10 kDa molecular weight cutoff (Millipore). In a second purification step, the eluates of the affinity purification were applied to an anion exchange column packed with 1.5 mL Q-Sepharose High Performance resin (GE Healthcare) preloaded with Cl^- ions. Proteins were eluted in 4 mL fractions with a step gradient of 100, 200 and 500 mM NaCl in 20 mM phosphate buffer at pH 7.5 supplemented with 1 mM $MgCl_2$ at a flowrate of 1 mL/min. Fractions containing the target proteins (as determined by SDS–PAGE) were pooled and concentrated to a final volume of 1 mL in buffer containing 100 mM Tris–HCl pH 7.5 and 10 mM $MgCl_2$ using ultrafiltration spin columns (Millipore) with a molecular weight cutoff of 10 kDa.

Validation enzyme assays with pure substrates. Purified enzymes (or, in the case of YhhY, lysates of its overexpression and deletion strains[36]) were incubated at 37 °C at a protein concentration of 50 µg/mL in 150 µL buffer containing 10 mM Tris–HCl pH 7.5, 1 mM $MgCl_2$ and 10 mM of each substrate. Compounds were purchased from Sigma-Aldrich at the highest available purity. At indicated timepoints, 10 µL of the reaction solution were transferred to 30 µL methanol cooled by dry ice to quench the reaction by enzyme denaturation. Reactant concentrations were subsequently measured by time-of-flight mass spectrometry[29]. Each experiment was repeated with at least two independent enzyme purification in experimental triplicates. As negative controls, heat-inactivated (60 min at 95 °C) proteins were used.

Gel electrophoresis. Gel electrophoresis (SDS–PAGE) was performed using precast 12-well 1 mm thick NuPAGE 4–12% Bis-Tris Gels (Novex) together with premixed NuPAGE MOPS SDS running buffer (Novex) at a constant voltage of 100 V with a PowerPac 1000 power supply (Bio-Rad). Gels were stained in

doi:10.1038/nmeth.4103

50%:40%:10% (v/v) methanol:water:glacial acetic acid containing 2 g/L Coomassie brilliant blue R-250 at room temperature for 3 h; and they were destained in the same solution without the Coomassie dye until background coloring was negligible. Gels were digitalized using a scanner (Hewlett-Packard).

Metabolomics analysis of gene deletion mutants. The 223 viable single-gene deletion mutants were obtained from the KEIO collection[36] and grown in 96-deepwell plates at 37 °C in 1 mL of minimal medium containing 7.52 g Na_2HPO_4–$2H_2O$, 3 g KH_2PO_4, 0.5 g NaCl, 2.5 g $(NH_4)_2SO_4$, 14.7 mg $CaCl_2 \times 2H_2O$, 246.5 mg $MgSO_4$–$7H_2O$, 16.2 mg $FeCl_3$–$6H_2O$, 180 µg $ZnSO_4$–$7H_2O$, 120 µg $CuCl_2$–$2H_2O$, 120 µg $MnSO_4$–H_2O, 180 µg $CoCl_2$–$6H_2O$ and 1 mg thiamine–HCl per liter of deionized water, supplemented with 4 g/L glucose as main carbon and energy source. Cells were harvested at midexponential growth phase (as monitored via absorbance at 600 nm using a TECAN M200 plate reader) by centrifugation at 4,000*g* and 4 °C for 10 min. Polar metabolites were extracted with 150 µL of water containing 2 µM reserpine and 2 µM taurocholic acid for 10 min at 80 °C with occasional vortexing. After centrifugation (0 °C, 10 min, 4,000*g*), supernatants of these extracts were stored at −80 °C before further analysis by flow injection mass spectrometry as described above. Two independent clones were grown for each mutant, and each metabolite extract was analyzed in technical duplicates. Intensities of 3,169 detected ions were standardized across mutants by calculating *Z*-scores (equation (1)), and tentative annotation as metabolites based on accurate mass was performed as described above. The complete data set is provided as **Supplementary Data Set 3**. The metabolomics data was compared with our enzyme screen data by verifying whether the predicted novel enzyme substrates or products were among the 5% of metabolite ions showing the strongest change in the respective deletion mutant.

Growth phenotyping experiments. All cells were grown at 37 °C in 96-well plates in 200 µL of M9 minimal medium containing 7.52 g Na_2HPO_4–$2H_2O$, 3 g KH_2PO_4, 0.5 g NaCl, 2.5 g $(NH_4)_2SO_4$, 14.7 mg $CaCl_2$–$2H_2O$, 246.5 mg $MgSO_4$–$7H_2O$, 16.2 mg $FeCl_3$–$6H_2O$, 180 µg $ZnSO_4$–$7H_2O$, 120 µg $CuCl_2$–$2H_2O$, 120 µg $MnSO_4$–H_2O, 180 µg $CoCl_2$–$6H_2O$ and 1 mg thiamine–HCl per liter of deionized water, supplemented with 4 g/L glucose unless indicated otherwise. Growth was monitored in a TECAN M200 plate-reading device by measuring the culture absorbance at 600 nm every 10 min. The wild-type strain was *E. coli* BW25113, and the deletion mutants derived from this strain were obtained from the KEIO collection[36]. A summary of tested growth conditions and the experimental outcomes is provided in **Supplementary Table 11**.

Statistics. To identify the appropriate *Z*-score cutoff to predict enzyme reactants from our screen data, we used an empirical approach optimizing the product of precision and recall of known enzyme assays, as described above. If one were to assume normality of the data, the identified *Z*-score cutoff of 5 s.d. above or below mean would translate into a one-tailed *P* value of 3×10^{-7}, pending correction for multiple-hypothesis testing.

Data availability. The mass spectrometry data are deposited in the MetaboLights database and can be accessed using the code MTBLS373. All other data are provided as Excel spreadsheets or comma-separated text files on the journal home page or are available from the authors upon request.

49. Zor, T. & Selinger, Z. Linearization of the Bradford protein assay increases its sensitivity: theoretical and experimental studies. *Anal. Biochem.* **236**, 302–308 (1996).
50. Riesenberg, D. *et al.* High cell density cultivation of *Escherichia coli* at controlled specific growth rate. *J. Biotechnol.* **20**, 17–27 (1991).
51. Kashket, E.R. Effects of aerobiosis and nitrogen source on the proton motive force in growing *Escherichia coli* and *Klebsiella pneumoniae* cells. *J. Bacteriol.* **146**, 377–384 (1981).