Published online 20 October 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 20* ***11495–11514***
doi: 10.1093/nar/gkx937

SURVEY AND SUMMARY

Confronting the catalytic dark matter encoded by sequenced genomes

Kenneth W. Ellens, Nils Christian, Charandeep Singh, Venkata P. Satagopam, Patrick May and Carole L. Linster*

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

Received January 18, 2017; Revised September 28, 2017; Editorial Decision October 02, 2017; Accepted October 03, 2017

ABSTRACT

The post-genomic era has provided researchers with a deluge of protein sequences. However, a significant fraction of the proteins encoded by sequenced genomes remains without an identified function. Here, we aim at determining how many enzymes of uncertain or unknown function are still present in the *Saccharomyces cerevisiae* and human proteomes. Using information available in the Swiss-Prot, BRENDA and KEGG databases in combination with a Hidden Markov Model-based method, we estimate that >600 yeast and 2000 human proteins (>30% of their proteins of unknown function) are enzymes whose precise function(s) remain(s) to be determined. This illustrates the impressive scale of the 'unknown enzyme problem'. We extensively review classical biochemical as well as more recent systematic experimental and computational approaches that can be used to support enzyme function discovery research. Finally, we discuss the possible roles of the elusive catalysts in light of recent developments in the fields of enzymology and metabolism as well as the significance of the unknown enzyme problem in the context of metabolic modeling, metabolic engineering and rare disease research.

INTRODUCTION

Over fifty years ago, Gordon Moore predicted that computing power would essentially double roughly every year (1). This prediction, although amended in 1975 to state doubling every two years (2), is still a goal and driving force for electrical engineers. Biologists have adopted Moore's Law as a tongue-in-cheek benchmark in reference to the decrease in cost of genome sequencing since the completion of the Human Genome Project in 2003. The drastic drop in sequencing cost since 2007 has continued to fuel the 'genomic revolution'; thousands of complete genomes are now publicly available in online databases (3). An important goal for scientists ever since has been to gain as many new biological insights from these collected genome sequences as possible. A subset of model species have garnered a certain priority in terms of annotation efforts. From DNA replication and repair, all the way to primary metabolic pathways, new genes have been characterized and annotated. Yet, a sizable proportion of the coding part of even well studied model organisms remains, up to the present day, without assigned molecular and/or biological functions.

With millions of protein sequences having been identified, it is impossible to experimentally characterize each protein, leaving computational annotation methods as the only reasonable means for systematic functional predictions. Routinely, homology is inferred, and then annotations are transferred, often leading to inaccurate or wrong predictions. Misannotation in public databases has progressed from <5% in 1998 to 40% in 2005, error propagation being suggested as a primary cause of such a dramatic increase (4). More sophisticated annotation strategies have been utilized for obtaining more solid *in silico* functional predictions (5,6), the most common ones relying on the use of genomic context information (e.g. gene clustering in prokaryotes) and of post-genomic resources (e.g. co-expression of related genes). Ultimately, skepticism for a given annotation lacking experimental evidence is prudent.

Progress in gene function annotation in the model organism *Saccharomyces cerevisiae* was reviewed by Hughes *et al.* in 2004 (7). Based on information contained in the Yeast Proteome Database (YPD; (8)), it was projected that

*To whom correspondence should be addressed. Tel: +352 46 66 44 6231; Fax: +352 46 66 44 36231; Email: carole.linster@uni.lu
Present addresses:
Nils Christian, Information Technology for Translational Medicine, 9 avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg.
Charandeep Singh, Cole Eye Institute, Cleveland Clinic, Cleveland, OH 44195, USA.

all the yeast proteins would be functionally characterized by 2007 (7). The authors admitted, however, that this was an over-optimistic prediction, mainly based on the fact that of the 80% of yeast proteins annotated as 'known' at the time in YPD, many were actually still very poorly understood on a functional level. A different analysis in the *Saccharomyces* Genome Database (SGD) by the same authors revealed that of the protein entries with a Gene Ontology (GO) annotation, 40% were of unknown molecular function and 30% were of unknown biological function (7). This nicely illustrates that defining criteria that qualify a protein as being 'known' is indispensable before even starting to address the question as to how well we understand the protein coding portions of sequenced genomes. In our opinion, protein function identification requires the gathering of experimental evidence to support a defined molecular dimension (e.g. catalytic reaction for enzymes) as well as a biological dimension (e.g. role in a metabolic pathway) (9,10). Until both dimensions are discovered and, importantly, can be reconciled with each other, the protein's function remains, according to this maybe more stringent definition, unknown. Less detailed and preliminary functional predictions, such as assignment to a more general enzyme class or subclass (e.g. 'hydrolase' or 'phosphatase'), may assist with hypothesis generation, but for most of the analyses in this paper will lead to the classification of corresponding proteins into the 'unknown' category (note that in this article 'unknown' in association with the terms 'gene', 'protein' or 'enzyme' is always to be understood as 'of unknown function'). According to our own GO term enrichment analysis in Yeastmine (data updated: 16 May 2016; (11)), 36% and 25% of the 6604 proteins in SGD remain unknown at the molecular and biological levels, respectively, demonstrating that progress in protein function identification since 2004 has been much slower than initially anticipated. In addition, given the misannotation problem, as well as the often limited understanding of the role of proteins that are annotated as known, the real proportion of unknown proteins in the yeast proteome is most certainly even higher. Often when the remaining unknown proteins are examined by studying strains overexpressing or deleted for the corresponding genes, the altered strains lack a strong phenotype (12). Functional redundancy with other genes or dispensability of the gene under standard laboratory conditions are possible reasons for lack of detectable phenotypic alterations in mutant strains (12), rendering functional elucidation of a number of the remaining genes of unknown function a more challenging task.

A similar proportion of the proteome of other well-studied organisms remains functionally less well characterized; various estimates state that about 30–50% of proteins encoded by the *Escherichia coli, Arabidopsis thaliana* and human genomes are unknown (13–15). Corroborating these estimates, ∼25% (∼5000) of the human proteins present in UniProt (16) have not been studied experimentally and for many of the remaining proteins, only sporadic experimental details have been reported (17). Additionally, upwards of 20% of protein domains contained in the Pfam database (18) are listed as 'domains of unknown function' (DUFs) (19) (also, ∼10–25% of the UniProt database protein entries are not associated with a Pfam family based on our own analyses in the *E. coli*, *Saccharomyces cerevisiae*, and human proteomes). Work to characterize these DUFs is important as a recent study emphasized that many essential proteins in model bacterial species contain such domains (20). This study also stated that about 9% of all the DUFs in Pfam (release 23) were found in all domains of life (20).

The fraction of DUFs, incompletely annotated proteins, and the disturbingly high error rates in biological database annotations in each given proteome reveal a gap in knowledge that has to be addressed with an earnest fervor. In this review, we focus on the budding yeast *S. cerevisiae* and humans as model systems as they rank among the most well annotated and thoroughly studied organisms. As a simple eukaryotic cell, yeast is genetically malleable, with many molecular and cellular tools to be exploited in addition to the feasibility of large-scale studies and the existence of extensive publicly available datasets. Also, given the high genetic conservation between *S. cerevisiae* and humans, this yeast has long been used as a model organism to progress in our understanding of human biology and disease (21,22). By providing updated estimations of the fraction of unknown proteins in the yeast and human proteomes, we provide a benchmark that can be compared to other model species. We also provide a classification of these unknown proteins into predicted functional categories.

Of particular relevance to fundamental metabolic research, metabolic modeling and engineering as well as rare disease research is the fraction of unknown genes predicted to code for enzymes. Given our interest in those fields, we applied systematic, bioinformatics-based approaches, described here, to more accurately estimate how many enzymes of uncertain function remain in proteomes of interest. We also critically review the methodologies that can be envisaged for enzyme function discovery. We conclude by discussing possible roles of the many players left in the catalytic dark matter; confronting it will be key to completing our understanding of an essential component of the cell machinery that many consider as fully elucidated, namely metabolism.

UPDATE ON THE STATUS OF THE UNKNOWN PROTEIN PROBLEM IN YEAST AND HUMANS

As discussed above, current estimates state that about 30-50% of the proteins in well-studied genomes are unknown (13–15). Starting from reviewed protein entries in the Swiss-Prot database, we aimed at updating these estimates more specifically for the yeast and human proteomes (6721 entries for *S. cerevisiae* and 20 201 entries for *Homo sapiens*; Figure 1A; UniProtKB/Swiss-Prot 2016_05). This was done according to instructions listed on the UniProt website where specific terms are suggested while others are to be avoided for functional annotations. Based on these recommendations, querying the reviewed human and yeast proteomes with the terms 'uncharacterized', 'putative', 'probable', 'containing', or 'like', should retrieve a high proportion of all the unknown or ambiguously annotated proteins. Very similar results were obtained for *S. cerevisiae* and for *H. sapiens*, with an estimated 1936 (i.e. 29%) yeast proteins and 6612 (i.e. 33%) human proteins of unknown function,

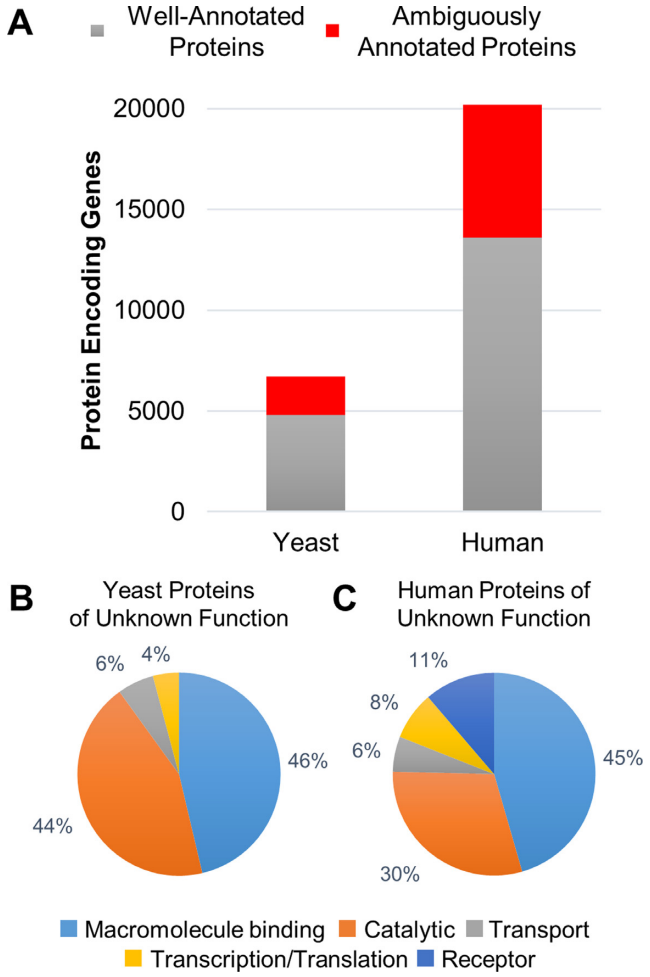

Figure 1. General statistics on proteins of unknown function in *S. cerevisiae and H. sapiens*. (**A**) Numbers of protein entries which have some certainty associated with their annotation (gray) and of ambiguously annotated protein entries (red). (**B** and **C**) Rough functional category predictions for yeast and human proteins of unknown function, respectively. Data retrieved from the UniProt database (http://www.uniprot.org/). Functional category predictions were made using the bioCompendium tool (http://biocompendium.embl.de/).

respectively (Figure 1A), showing that the numbers stated in above cited references are not yet outdated.

These yeast and human unknown proteins were further analyzed using bioCompendium (http://biocompendium.embl.de/; see Supplementary Material and Supplementary Tables S1 and S2 for details on the analysis), a publicly available, high-throughput experimental data analysis platform that allows to collect biological information for large protein or gene lists based on existing annotations in various databases, for functional category prediction (Figure 1B and C). It should be noted that this analysis allowed to make preliminary functional predictions for only 631 and 3741 proteins out of the initial 1936 yeast and 6612 human proteins of unknown function, respectively, which is not surprising given that we started out with poorly annotated protein sets. Despite the limitations of this approach, it allowed to quickly obtain a rough first idea about the relative distribution of the remaining yeast and human proteins of unknown function among the major functional categories 'Macromolecule binding', 'Catalytic', 'Transport', 'Transcription/Translation' and 'Receptor'. Strikingly, this preliminary analysis indicated that a third or more of the analyzed proteins of unknown function possess a catalytic activity, with considerably less of them being involved in signaling, transcription, translation or transport activities. To gain a more accurate assessment of the number and potential roles of enzymes of unknown function remaining in the yeast and human proteomes, more elaborate bioinformatics methods were used as described in the next section.

ENHANCED APPROACHES TO RETRIEVE ENZYMES OF UNKNOWN FUNCTION IN KNOWN PROTEOMES

First we used two more advanced searches in the UniProt database (Figure 2, blue background). Using information available in this database, one can readily find out that of the 1746 yeast proteins and 4207 human proteins annotated with Enzyme Commission (EC) numbers, 369 and 1156, respectively, are linked to (only) incomplete (i.e. <4-digit) EC numbers (Figure 2; note that for proteins with multiple EC number associations, hits linked to at least one 4-digit EC number were excluded and that preliminary EC numbers, containing an 'n' as part of the fourth digit, were treated as 3-digit EC numbers). This indicates that an enzymatic activity has been associated with the latter proteins, but that this molecular function remains only partially characterized. In an alternative query, taking as a starting point not the EC number-associated proteins, but the 1936 and 6612 unknown yeast and human proteins retrieved via the keyword-based search in UniProt described in section 2, we found that 242 yeast and 1207 human unknown proteins are associated with at least one EC digit (Figure 2). This second approach has the advantage of retaining proteins as hypothetical enzymes, despite their association with a complete EC number. Specifically, the protein lists generated via this second method contained 117 yeast and 663 human proteins annotated with 4-digit EC numbers, because the associated UniProt protein names contained one of our 'unknown' categorizing terms ('uncharacterized', 'putative', 'probable', 'containing', or 'like'). Upon manual inspection, we noticed, however, that 4-digit EC associated proteins for which 'unknown' terms are only contained in one or several of their alternative UniProt protein names (and not in their 'recommended' UniProt name) tend to be well or fully characterized functionally whereas this characterization is usually more incomplete when the 'unknown' term is comprised in the recommended UniProt name (two examples of the latter type are specified in Supplementary Material). Protein entries of the first type are therefore flagged as possible false positives in the database query lists in Supplementary Table S3. A disadvantage of the second approach is that the search of putative enzymes is restricted to a probably incomplete list of proteins of unknown function and therefore excludes potentially interesting targets from the start (although our keyword-based method aimed at maximizing unknown protein retrieval, it heavily depends on annotation quality and a fraction of unknown proteins were most likely missed). It seems that crossing the resulting lists of putative enzymes retrieved by

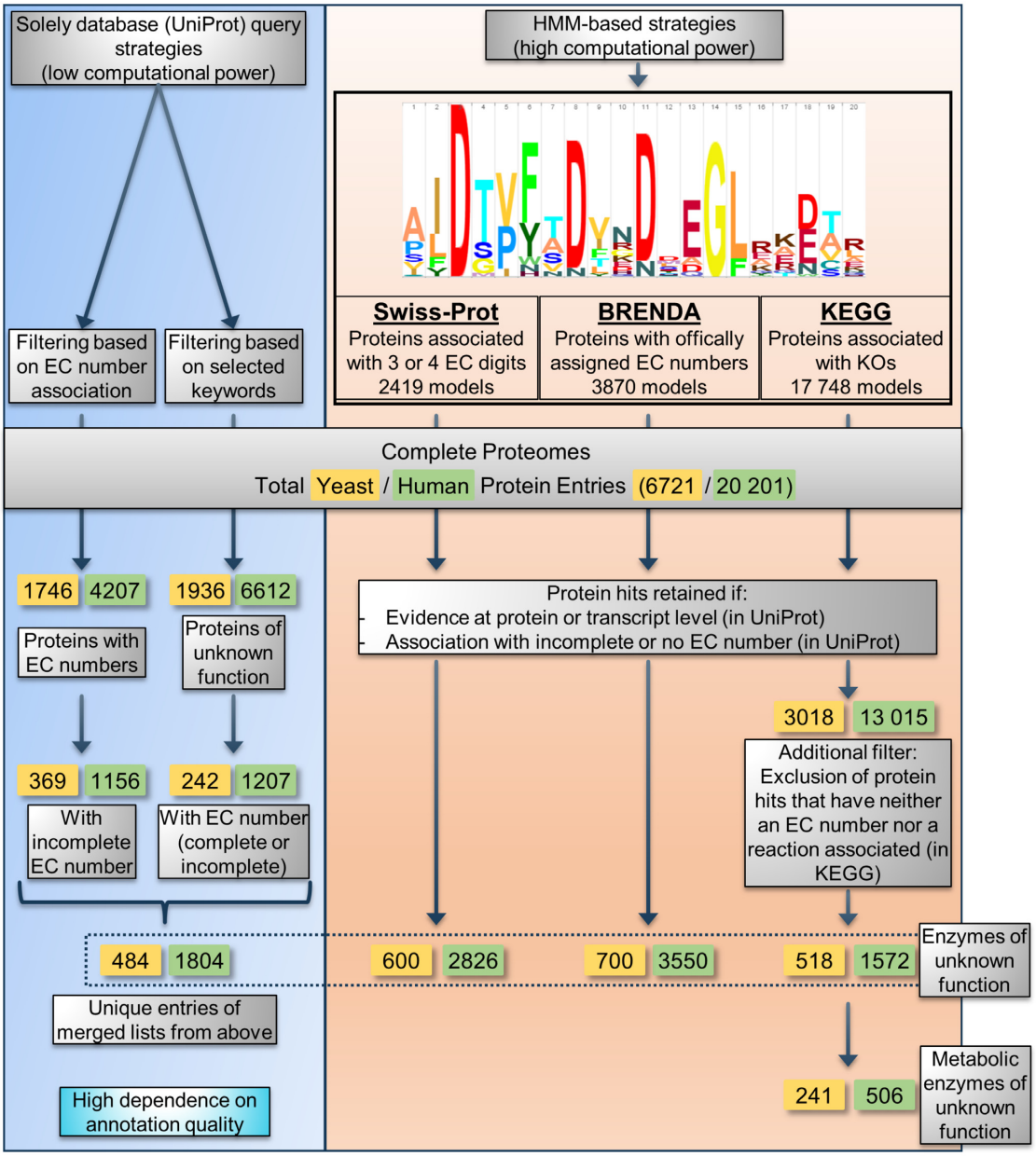

Figure 2. Solely database query- and HMM-based workflows described here to systematically retrieve enzymes of unknown function for an organism of interest. The strategies based only on database queries are represented on a blue background whereas the HMM-based strategies are represented on an orange background. Numerical estimations obtained for the *S. cerevisiae* and *H. sapiens* proteomes are shown in yellow and green boxes, respectively.

these slightly more advanced database searches should lead to a quite comprehensive dataset. 242 yeast and 597 human proteins were only found by the first approach, whereas 115 yeast and 648 human proteins were only found by the second approach. 127 yeast and 559 human proteins were identified by the two strategies. The total number of unique proteins found by the combination of both approaches to correspond to putative enzymes amounted thus to 484 and 1804 for the yeast and human proteomes, respectively (Figure 2 and Supplementary Table S3).

We next went on to adapt a method from Christian *et al.* (23) based on the use of Hidden Markov Models (HMMs) (24) to reach our aim of estimating the proportion of putative enzymes in known proteomes by yet another approach. HMMs are commonly used for functional prediction of protein sequences (25,26). Based on high-quality multiple sequence alignments (MSAs), they are able to capture a broader range of sequence similarity than standard alignment methods like BLAST or PSI-BLAST (27–29). Using sequence information available in the Swiss-Prot (Release: 2015_10), BRENDA (Release: 2015.2; (30)) and KEGG (Release 76.0; (31)) databases, several sets of HMMs were generated (as described in Supplementary Material). We chose to use Swiss-Prot and BRENDA as the manual curation of these databases should minimize the amount of misannotation in the corresponding HMM sets. KEGG was

selected because of the possibility to use the reactions associated with the KEGG Orthology (KO) entries to estimate the fraction of enzymes of unknown function that act solely on small molecules, as will be described below. It should be noted here that while the created Swiss-Prot and BRENDA HMMs (2419 and 3870 models, respectively) represent 'enzyme-specific' HMMs, the KEGG HMMs (17 748 models) also represent non-catalytic protein functions (see Supplementary Material). Hits obtained in the proteomes of interest with the KEGG HMMs were therefore filtered more extensively (as explained in more detail below) than those obtained with the Swiss-Prot and BRENDA HMMs to enrich the generated protein lists in putative enzymes (Figure 2). For the sake of completeness, we also used precomputed Pfam HMM hits for *S. cerevisiae* and *H. sapiens* to identify putative yeast and human enzymes (see Supplementary Figure S4 and method description in Supplementary Material). Given, however, the considerably lower Pfam HMM to EC associations as compared to for example KEGG HMM to EC or reaction associations (329 versus 7788), the Pfam HMM approach yielded much lower numbers of putative yeast and human enzyme hits (see Supplementary Table S3) and these results were not analyzed further in this study.

The complete yeast (Swiss-Prot TaxonomyID: 559292; *Saccharomyces cerevisiae* S288c) and human (Swiss-Prot TaxonomyID: 9606; *H. sapiens*) proteomes were scanned against the Swiss-Prot, BRENDA, and KEGG HMMs using the `hmmscan` command in HMMER (32). Starting from the generated datasets (yeast and human proteins presenting significant matches with one or several of those models; HMM *E*-value cutoff of 0.00005), different lists of proteins of unknown function or with incomplete functional annotation were built (Supplementary Table S3). Independently of which HMM set was used (Swiss-Prot, BRENDA or KEGG), only yeast and human protein hits with evidence level 1 or 2 in UniProt were retained (Figure 2). For the hits obtained with the Swiss-Prot and BRENDA HMMs, a protein was grouped according to whether it associated (in UniProt) with EC numbers containing 3, 2 or 1 digit(s) or was not associated with an EC number; proteins associated with complete EC numbers (four digits) were not retained as they were considered as of known function, although this does certainly not always hold true as illustrated above. For the hits obtained with the KEGG HMMs, entries with four EC digits were again excluded while the remaining entries were divided into four categories: (i) non-catalytic—no EC number and no reaction associated; (ii) ECbutNoReaction—an EC number is associated with the KO, but no reaction is associated with either the KO directly or with the EC number; (iii) DNARNAPeptideChain—entries with associated reactions (direct or via EC) in which at least one of the substrates or products corresponds to DNA, RNA, a peptide, or a repeated chemical subgroup (list presented in Supplementary Table S4); (iv) other catalytic—the remaining entries which should be enriched in enzymes acting on small molecule substrates, i.e. metabolic enzymes. The protein hits in categories (ii), (iii) and (iv) were retained as putative enzymes according to the KEGG HMM strategy.

Based on the 'enzyme-specific' Swiss-Prot and BRENDA HMMs, 600 and 700 yeast proteins as well as 2826 and 3550 human proteins were predicted to correspond to putative enzymes, respectively; KEGG HMMs identified 518 yeast proteins and 1572 human proteins as putative enzymes (Figure 2 and Supplementary Table S3). A significant advantage of the HMM-based strategies as compared to even advanced database searches is the ability to retrieve proteins that are enzyme candidates, but that have not yet been linked to EC numbers in the protein databases. The numbers of such 'no EC-associated' hits were very high in our analysis, corresponding to 381 yeast and 1952 human putative enzymes retrieved through the Swiss-Prot HMMs, 492 yeast and 2818 human putative enzymes retrieved through the BRENDA HMMs, and 248 yeast and 637 human putative enzymes retrieved through the KEGG HMMs. Logically, these putative enzymes would not have been retrieved by searching the databases for proteins with associated EC numbers, but as they match 'enzyme-specific' HMMs with high scores, there is a strong probability that they possess catalytic activities. Another important advantage of the HMM-based strategy is that it inherently generates predictions on possible catalyzed reactions for each of the protein hits. It is important to keep in mind, however, that the HMMs used here only capture pre-existing knowledge on sequence–function associations. There might well be protein sequences with catalytic properties that have not yet been identified as such in any sequenced organism; those are not represented by our sequence models and will rely on other (mostly experimental) strategies to be discovered. An interesting perspective could be to use alignment-free prediction methods to complement our HMM-based strategy, which relies on sequence similarity, for the identification of putative enzymes in proteomes of interest. Homology-free methods operating through neural network or support vector machine approaches have indeed been used to predict global enzyme classes (first or second level of the EC enzyme classification system) based on sequence (33,34) or structural (35) attributes, independently of alignment algorithms. These methods may allow for identification of additional putative enzymes in the yeast and human proteomes, even if they share no significant or low sequence similarity with any enzyme of known function (and thus are unlikely to be retrieved by the HMM-based strategy).

In total, fewer putative enzymes were retrieved using the KEGG HMM method as compared to the Swiss-Prot and BRENDA HMM methods, for both the yeast and human systems. An advantage of the KEGG HMM approach was, however, that the number of protein hits contained in the 'other catalytic' category (as defined above) could be directly used to estimate the proportion of putative metabolic enzymes, i.e. enzymes predicted to act on low molecular weight substrates as opposed to high molecular weight substrates like DNA, RNA, or proteins. From this analysis, it was estimated that at least 241 yeast and 506 human enzymes of unknown function act on small molecules (Figure 2 and Supplementary Table S3). To provide some context, the most comprehensive human metabolic reconstruction, Recon 2, includes 2626 unique metabolites involved in 7440 reactions, annotated with 1789 unique genes (36). The most recently released Yeast metabolic network, Yeast

7.6 (http://yeast.sourceforge.net/) (37), contains 2344 reactions, annotated with 910 yeast genes encoding the catalyzing enzymes. Comparison of the number of genes included in those reconstructions (910 and 1789 for yeast and human, respectively) with the number of enzymes of unknown function acting on small molecules estimated here (241 and 506 for yeast and human, respectively), indicates that we are far from a complete knowledge, and certainly even further from a complete understanding, of the metabolism of these two model organisms.

We next quantified the overlaps (Supplementary Figures S1 and S2) between the unique entries predicted as putative enzymes by the combined Uniprot query-based approaches (Figure 2, blue background) and the putative enzymes identified via the Swiss-Prot, BRENDA or KEGG HMM-based strategies (Figure 2, orange background). Clearly, the query- and HMM-based strategies are complementary to one another, with ∼40–50% of the putative enzymes found by the query-based method also being retrieved by the HMM-based methods. Similarly, when comparing the different variants of HMM-based methods used, none of them was completely redundant with each other. The highest overlap was found between the Swiss-Prot HMM- and BRENDA HMM-based methods, which may be explained by the high manual curation of both the Swiss-Prot and BRENDA databases. While at this stage it is difficult to tell which of the generated putative enzyme lists is the most highly enriched in 'true' enzymes, one may conclude from these comparisons that, because neither of the described strategies is perfect, it is recommendable to use more than one of them in combination if one strives to obtain the most complete lists possible. Given the higher curation of the Swiss-Prot and BRENDA databases compared to the KEGG database and considering the overlap in methods shown in Supplementary Figures S1 and S2, combining the query-based method with either the Swiss-Prot HMM- or the BRENDA HMM-based method seems like a good strategy to maximize the number of putative enzymes to be retrieved from the proteome of interest. It should be noted, however, that the query-based strategy highly depends on the annotation quality of the target proteome and that the HMM-based methods are to be favored when working with poorly annotated proteomes.

For the unknown enzyme 'hits' obtained by the HMM-based approaches and associated with an incomplete EC number, we also compared their distribution among the six major EC classes (Supplementary Figure S3). The overall distribution of enzymes of unknown or ambiguous function in yeast and humans, found by using the Swiss-Prot, BRENDA or KEGG HMMs, were generally similar. Hydrolases (EC 3) made up the largest fraction of unknown enzymes, followed by transferases (EC 2), oxidoreductases (EC 1) and ligases (EC 6), respectively. The only exception to this trend was found for the hits retrieved from KEGG HMMs scanned against the human proteome, where the ligase class had the second most hits, followed by transferases and oxidoreductases. Relatively few of the unknown enzyme hits belonged to the lyase (EC 4) or isomerase (EC 5) classes. A comparison of these unknown enzyme EC category distributions to the one of all the EC-number associated proteins in UniProt (including known and unknown enzymes), reveals as a major difference the enrichment of the unknown enzymes in hydrolases at the expense of the transferase class. It should be noted here, however, that many (KEGG HMMs) if not a majority (Swiss-Prot and BRENDA HMMs) of putative unknown enzyme hits retrieved by the HMM-based methods are not yet associated with an EC number (see above and also legend to Supplementary Figure S3), meaning that the pie-charts in Supplementary Figure S3 may not be representative of the catalytic dark matter that persists in the yeast and human proteomes as a whole.

As essentiality and evolutionary conservation are important parameters in the field of functional genomics, we also analyzed the yeast enzymes of unknown function identified by the various approaches described above to estimate the fraction of proteins encoded by essential or orphan genes among those putative enzymes (the corresponding methods are described in Supplementary Material). Depending on the HMM method used (Swiss-Prot, BRENDA or KEGG), the fraction of orphan genes (defined here as genes having only fungal homologs) among the yeast unknown enzymes corresponded to 12–25% while the fraction of essential genes ranged from 15% to 23% (Supplementary Table S5). The relatively low orphan gene fraction further underlines the importance of investigating the functions of these putative yeast enzymes as the conservation of the majority of them suggests fundamental roles in cellular processes. Regarding the fraction of essential genes among the yeast unknown enzymes, it seems unexpectedly high as strong phenotypes should in principle favor functional identification. It is interesting to note that the subset of 241 yeast 'metabolic' unknown enzymes predicted by the KEGG HMM method contains a lower fraction of essential genes (5%), which indicates that most of them are good candidates to be functionally investigated by experimental strategies involving the analysis of knockout mutants (see for example, the *ex-vivo* metabolomic profiling approach described in the section below).

All of our automated bioinformatic approaches (database queries and HMM-based methods) used to approximate the number of remaining unknown enzymes in the yeast and human proteomes heavily relied on EC number annotation. While it seemed like the most suitable annotation parameter to consider for reaching our specific objective, it is clear that the number of digits in an EC number does not always accurately reflect the degree of functional characterization of associated proteins. This, as well as the more general misannotation problems in protein databases mentioned in the Introduction, are reasons why the protein lists generated here (Supplementary Table S3), while enriched in unknown enzymes, also contain a number of 'false positives', i.e. proteins whose enzymatic functions are actually reasonably well or sometimes even fully understood. While extensive manual curation would have gone beyond the scope of this article, we attempted to analyze and roughly quantify the 'false positive' problem. Based on systematic criteria explained in more detail in Supplementary Material and given the ambiguous meaning of especially 3-digit EC numbers (38) in terms of degree of functional characterization of the associated enzymatic activities, a subset of the 3-digit EC number linked entries

in Supplementary Table S3 have been flagged as 'possible false positives'. In addition, some of the proteins associated with 4-digit EC numbers (only present in the database-query lists in Supplementary Table S3) have been highlighted as 'possible false positives' as described at the beginning of this section. Depending on the approach used for automated unknown enzyme prediction, the estimated possible false positive percentages amounted to 43% (database-query), 22% (SwissProt HMMs), 15% (BRENDA HMMs) and 29% (KEGG HMMs) for the yeast proteins and 50% (database-query), 21% (SwissProt HMMs), 13% (BRENDA HMMs) and 40% (KEGG HMMs) for the human proteins. Manual inspection of the functional UniProt annotations of the flagged entries revealed that some of them are clear false positives (i.e. enzymes whose function can be considered as known, even according to stringent criteria), while others remain functionally underdetermined; only manual assessment of each individual entry would allow to filter out the clear false positives. For the sake of comprehensiveness, all the possible false positive entries have therefore been retained in our enzyme of unknown function lists; especially for those flagged proteins, however, a careful assessment of the 'unknown' status should be performed before including them into any functional identification projects. Overall, the apparent overestimation of our total unknown enzyme numbers due to possible false positives is, however, limited to a certain extent by other drawbacks of our automated bioinformatics analyses which rather lead to underestimations of these numbers (e.g. dependency on previously identified sequence-enzyme function associations for the HMM methods, difficulty to properly account for multi-functionality of proteins).

EXPERIMENTAL AND COMPUTATIONAL STRATEGIES TO IDENTIFY NEW METABOLIC ENZYMES

Several different methods have proved to be effective in addressing the unknown enzyme problem. They can be hypothesis based or non-hypothesis based and the starting point can be a gene of unknown function, an enzyme activity with no associated gene (orphan enzyme) or a metabolite that remains unconnected to the known metabolic pathways. Given our primary research interests, the focus is mostly (although not exclusively) on model organisms and intermediary metabolism related enzymes, but it should be noted here that specialized metabolism or natural product biosynthesis in plants, fungi and/or marine organisms, although less studied and often considered as esoteric, may even constitute more fertile grounds for enzyme discovery (39).

The 'classical biochemical' approach

Here the starting point is a known or hypothesized enzyme activity and the approach is inevitably hypothesis-driven. It can only be implemented if the reaction catalyzed by the putative native enzyme can be assayed in cell or tissue extracts. The crude starting material is then enriched for the new or orphan enzymatic activity through multiple purification steps to acquire a protein preparation that is pure enough for protein sequence identification via tandem MS/MS (Figure 3). Selection of the best protein candidate among the identified sequences can be assisted by (i) determination of the molecular weight of the putative enzyme based on SDS-PAGE gel bands co-eluting with the enzymatic activity during the chromatographic procedure, (ii) transcriptomic data obtained under conditions, or from strains/species, where the enzyme is known to be differentially expressed, (iii) information on the tissue distribution, subcellular localization and/or other properties such as metal-dependency (obtained experimentally for the enzymatic activity of interest) and (iv) sequence information on proteins catalyzing similar reactions to the one investigated. If the strategy results in the identification of a reasonable number of plausible candidate protein sequences, they can then be produced recombinantly for final enzyme function assignment. While viewed, somewhat accurately, as tedious and time consuming (40), this approach continues to bear fruit and leads to discoveries where other more systematic approaches fail. This and some other biochemical approaches described below also have no substitute when working with organisms that are not amenable to genetic manipulation, which still represent a majority. In recent years, it has led to the molecular identifications of carnosine synthase (41), β-citrylglutamate hydrolase (42), and a lysoplasmalogenase (43) in animals, a glucuronokinase in plants (44), the lyase responsible for forming dimethyl sulfide from dimethylsulfoniopropionate in algae (45), hemocyanin with potential lignin-modification activities in a wood-feeding termite (46), and four enzymes involved in a process called metabolite repair that will be described in more detail in the last section (ethylmalonyl-CoA decarboxylase (47), NAD(P)HX dehydratase (48), β-alanyl-lysine dipeptidase (49), and a 4-phosphoerythronate/2-phospho-L-lactate phosphatase (50)), to name only a few.

In the case of ethylmalonyl-CoA decarboxylase, the compound ethylmalonyl-CoA was hypothesized to be formed intracellularly by a known side activity of propionyl-CoA carboxylase and acetyl-CoA carboxylase on butyryl-CoA (47). Mouse tissue extracts were assayed with radioactive ethylmalonyl-CoA to determine if a decarboxylation reaction could be detected. After activity was shown, an enzyme was partially purified from rat liver by successive chromatographic steps (47). LC-MS/MS performed on the most purified ethylmalonyl-CoA decarboxylase activity-containing fraction provided a list of 75 protein hits from which candidates were selected based on sequence similarity with proteins known to catalyze a similar reaction. ECHDC1 emerged as the top protein candidate for ethylmalonyl-CoA decarboxylase, which was then confirmed by characterization of the recombinant protein. Understandably, this approach cannot easily be adapted for a proteome scale project, making it seem less desirable to implement, but for groups with the specialized biochemical expertise and skills, it remains a powerful technique for original findings in the field of enzyme discovery (51). As with any targeted approach, it will always depend on the functional hypotheses that can be generated by investigators and completely unexpected enzymatic reactions may be missed.

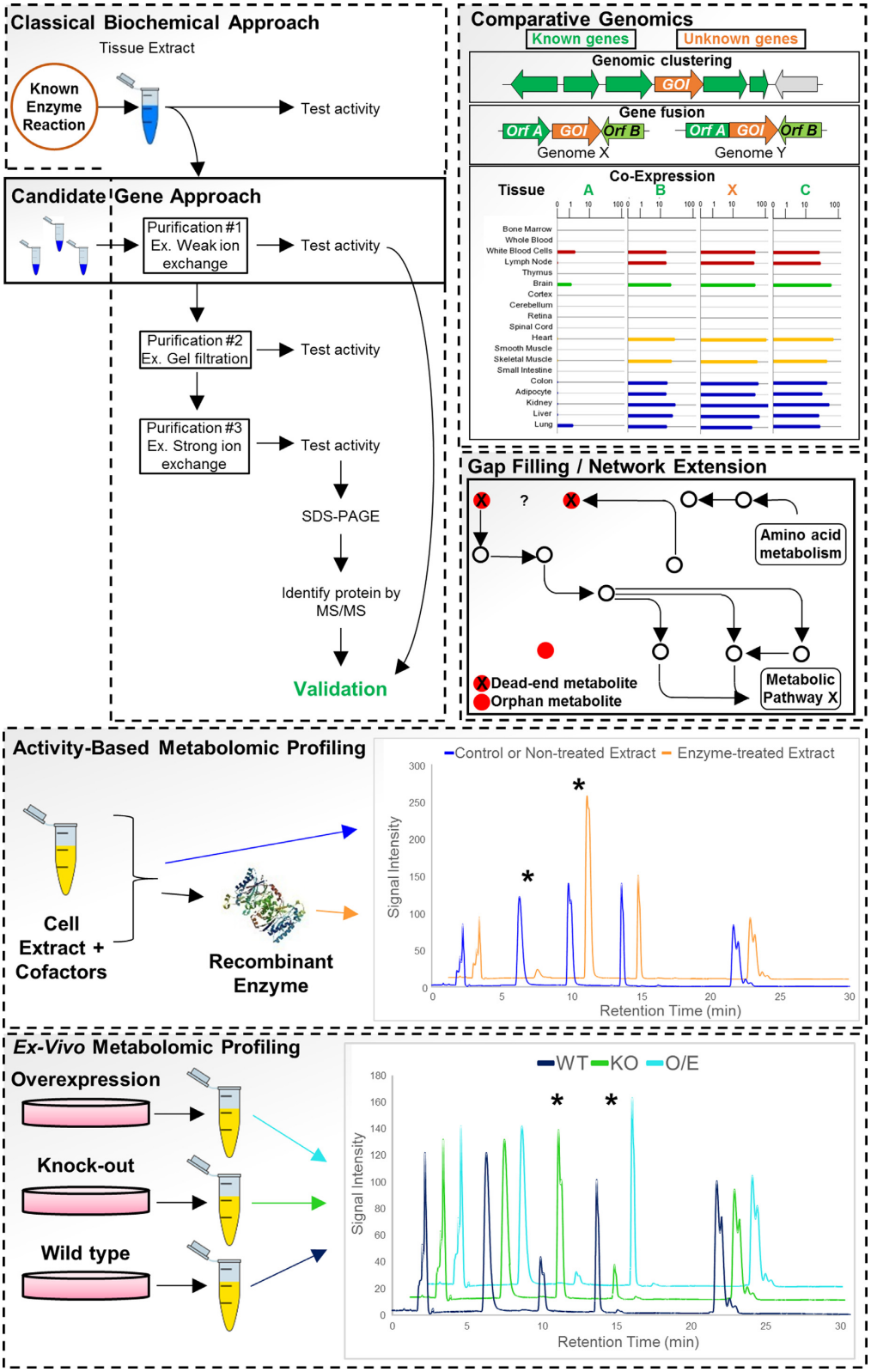

Figure 3. Experimental and computational strategies to find new metabolic enzymes. Only a subset of the strategies described in this review are illustrated here. Asterisks in the chromatograms denote peaks of interest. GOI, gene of interest; Orf, open reading frame; WT, wild-type; KO, knock-out; O/E, overexpression.

The 'candidate gene' approach

This approach also starts from a known or hypothetical enzymatic activity, but then directly proceeds to functional validation by testing one or several candidate genes that have been carefully selected based on a number of known or hypothesized properties of the orphan or putative enzyme (Figure 3). It replaces the classical biochemical approach when, for instance, the activity of the native enzyme cannot be detected in tissue extracts (because of low endogenous expression and/or the presence of interfering activities for example) or when the native enzyme cannot be purified in an active form (because of stability problems and/or integral membrane localization for example). The success of this approach depends largely on the availability of a specific and sensitive enzymatic assay, especially if enzymes cannot be purified from cultured cells manipulated to overexpress the candidate genes, and the amount of prior information available for the enzymatic activity of interest. Knowledge of the substrates and cofactors involved, enzymatic properties determined in crude extracts, tissue distribution, subcellular localization, and conservation across species can all guide the candidate gene selection process. In principle, a screening approach based on detection of the enzymatic activity of interest in cultured cells transfected with plasmid pools of a cDNA expression library should allow one to skip the prior gene selection process. While such a screen has allowed the identification of an endothelin receptor (52), it has not yet been successfully used, in our knowledge, for the molecular identification of an enzyme. Examples of enzymes for which the encoding genes have been more recently found by the candidate gene approach include the aspartate N-acetyltransferase NAT8L, which synthesizes one of the most abundant brain metabolites (N-acetylaspartate) (53), the mammalian *cis*-aconitate decarboxylase IRG1, which produces the antimicrobial metabolite itaconic acid (54), alkylglycerol monooxygenase, a tetrahydrobiopterin-dependent enzyme responsible for the cleavage of the O-alkyl bond in ether lipids (55), the GDP-L-galactose phosphorylase VTC2, which catalyses the last step of the main plant vitamin C synthesis pathway that remained genetically unidentified (56), and the omega-amidase NIT2, which hydrolyzes the amide group of the alpha-ketoglutaramate product formed by glutamine transaminases (57). Additional examples are N-acetylaspartylglutamate synthase (58), acetyl-CoA:lyso-PAF (platelet-activating factor) acetyltransferase (59) and two other enzymes involved in ether lipid metabolism (60).

In a variant of this approach, one can aim to identify the molecular function of a particular gene, predicted to encode an enzyme. The starting point here is a gene, and not an enzymatic activity. This can lead to the discovery of a new enzymatic activity by characterization of the properties of the recombinant enzyme *in vitro*, for which the biological role then needs to be determined. The latter task is often not straightforward and can be initiated by confirming that the proposed substrate of the newly identified enzymatic activity accumulates in knockdown or knockout cell lines of the corresponding gene. This approach has for example led to the identification of the human C15orf58 gene as a GDP-glucose phosphorylase (61), of a malate/β-methylmalate synthase activity for the human CLYBL protein (62), and most recently of the mammalian Nit1 protein and its yeast ortholog (encoded by the *NIT2* gene) as amidases that hydrolyze deaminated glutathione, a side product of several transaminases (63). Both the human Nit1 protein and its yeast ortholog were still retrieved in the current study as enzymes of unknown function by the database query as well as the SwissProt, Brenda, and KEGG HMM-based approaches (Supplementary Table S3).

When starting an enzyme discovery project from a gene of interest, bioinformatics tools for functional prediction, typically based on sequence similarity, are often used for hypothesis generation. Caution needs to be taken with these predictions, however, as small changes in amino acid sequence can lead to drastic changes in enzyme function. The *Streptomyces purpurascens* RdmB protein for example is homologous to classical SAM-dependent methyltransferases, but turned out to catalyze a SAM-dependent hydroxylation reaction in the synthesis pathway of the anticancer antibiotic rhodomycin (64). Another extreme example to illustrate this are melamine deaminase and atrazine chlorohydrolase, two enzymes encoded by two different *Pseudomonas* strains and that differ only by nine amino acids, but that carry out different reactions on different substrates (65). These considerations again emphasize the importance of experimental validation in the context of enzyme identification studies, when starting from sequence similarity-based predictions.

Activity-based metabolomic profiling

The improved sensitivity and decreasing cost of mass spectrometry techniques have encouraged researchers to move away from solely hypothesis-driven approaches for enzyme discovery to untargeted approaches using metabolomics techniques. Often these metabolomics projects begin with little to no prior information available about potential substrate(s) or product(s) of the putative enzyme of interest. Therefore it is advantageous for the metabolomics methods used to be as comprehensive as possible in terms of metabolite coverage (40). One metabolomics-based approach, designated activity-based metabolomic profiling (ABMP), involves incubating a recombinant enzyme with a cell extract from the homologous organism (or another organism) enriched in potential cofactors by supplementation (Figure 3) (66). Enzymatic activities are revealed by analyzing the consumption and production of metabolites in a time- and protein-dependent manner using mass spectrometry-based approaches.

Initial work on the *E. coli* protein YihU can be used to illustrate the approach in action (67). A small molecule extract, derived from yeast, was supplemented with additional cofactors and incubated with and without recombinant YihU. Using capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS), two anionic compounds were found at increased levels (assumed products) in the YihU incubated sample compared to the non-enzyme control sample. In addition, these compounds were found to be produced in an NAD^+/NADH-dependent manner, strongly suggesting an oxidoreductase-type reaction. The accurate mass of these compounds was compared to the theoretical

mass of compounds in the KEGG LIGAND database, ultimately leading the authors to the conclusion that *E. coli* YihU is a dehydrogenase reducing succinic semialdehyde to γ-hydroxybutyrate using NADH (67). However, five years later the reported γ-hydroxybutyrate dehydrogenase activity was refuted and an alternative activity was found (68). Using a combination of genomic context and enzyme assays in a hypothesis-driven approach, YihU was shown to act as a 3-sulpholactaldehyde reductase in sulphoglycolysis in *E. coli*. This example underlines the importance of validating the *in vivo* biological relevance of a molecular function of an enzyme supported by experimental evidence obtained in *in vitro* systems, even more so when using a non-hypothesis driven approach.

ABMP was also used to investigate further into the function of the *Mycobacterium tuberculosis* protein Rv1248c. An early report had described an activity where α-ketoglutarate (α-KG) was decarboxylated to form succinic semialdehyde (SSA) (69). However, another group realized that the rate at which Rv1248c produced SSA could not match the metabolic role that had been proposed (66). In order to identify the more physiologically relevant role, the authors used ABMP by incubating recombinant Rv1248c with a mycobacterial small molecule extract. There was a time-dependent change in the abundance of only two metabolites; α-KG decreased in an enzyme-dependent manner, while there was an increase of a different feature, identified as 5-hydroxylevulinic acid. It was determined that Rv1248c catalyzes the formation of a C–C bond between α-KG and glyoxylate to produce 2-hydroxy-3-oxoadipate that subsequently decomposes to 5-hydroxylevulinic acid (66). This finding was further corroborated a year later, as the two previously reported functions were confirmed and a third activity was identified, namely the succinyl-transferring α-KG dehydrogenase activity (70).

An asset of the ABMP approach is the use of the cell metabolome as a highly relevant chemical library to screen for potential enzyme functions. This is important because some reactions of interest will involve substrates that are not commercially available. The ABMP approach can also allow one to determine the function of enzymes encoded by essential genes; this is not the case for the '*ex-vivo*' metabolomics strategy described below (see also (66)). Potential bottlenecks are that neither the cell extracts used as substrate source, nor the supplemented cofactor mix are necessarily comprehensive; the substrate of interest may be unstable and lost upon cell extraction or may even be absent from the start if it is only produced by the cell under specific growth conditions for example. Another limitation of this approach is the dependence on obtaining pure recombinant protein. Also, as the approach involves an *in vitro* reconstituted system, it does not necessarily reveal the physiological function of the investigated protein. Finally, with metabolomics being at the center of this strategy, it suffers from the limitations that this young 'omics' technique still has to tackle, notably in terms of metabolite coverage, sensitivity and especially compound identification, as described in more detail in the next subsection.

Ex-vivo metabolomic profiling

The *ex-vivo* metabolomics approach consists in comparing tissue or cell extracts (or spent media) derived from organisms/cells deficient in or overexpressing the enzyme of unknown function to extracts of wild-type organisms/control cells to identify differences in their metabolic profiles (Figure 3) (71). Using untargeted metabolomics, the expectation is that the hypothesis for substrate identity can be significantly narrowed down to a class of compounds or even a single compound that changes between the different cell extracts (e.g. compound accumulating in a knockout cell line compared to the control cell line). While this idea may seem simplistic, systematic comparative metabolomics profiling of yeast strains knocked out for central metabolic enzymes showed that for almost half of the analysed strains, enzyme deletion led to very localized changes in the metabolome, and actually mostly accumulations of the substrate immediately upstream of the lesion (72). For most of the remaining deletion strains, no significant metabolite level changes could be detected and for only very few of the strains more systematic changes across the metabolic network were measured. Similar observations had been reported previously for *E.coli* enzyme deletion strains (73). *Ex-vivo* metabolomic profiling can also assist with enzyme function identification using the guilt by association principle as mentioned in a dedicated subsection below.

The *ex-vivo* metabolomics approach is more likely to reveal a physiologically relevant role for an unknown enzyme than a non-hypothesis driven *in vitro* method, as observed metabolic differences result from processes that occurred within the living cells and as intracellular regulations that may influence the enzyme activity (e.g. allosteric interactions and post-translational modifications) have a better chance of being present in the cell or whole organism models used. However, for non-genetically tractable organisms where methods to produce knockout or overexpression strains are burdensome or currently impossible, alternative methods such as the above mentioned ABMP should or have to be used (71). Reciprocally, in cases where it is challenging to purify an enzyme of interest, the *ex-vivo* approach has to be chosen over the ABMP approach to discover the natural substrates and/or physiological function (74).

Ex-vivo metabolomics profiling led for example to the identification of the α/β-hydrolase domain-containing three protein ABHD3 as a lipase selectively cleaving certain medium-chain phospholipids. The investigators transiently overexpressed 12 uncharacterized enzymes belonging to the serine hydrolase class in HEK293T cells and extracted the organic-soluble metabolites for subsequent LC–MS analysis (74). Compared to a transfection control, extracts of cells overexpressing ABHD3 displayed an increased peak (m/z = 524) identified through tandem MS and the co-elution with a synthetic standard as C18-lysophosphatidylcholine (74). These cell culture-based findings were corroborated with observations from a whole-organism model when tissue extracts of homozygous ABHD3 knockout mice were compared with extracts of WT mice, and C14-lysophosphatidylcholine was

found to accumulate in addition to three other phosphatidylcholines (74). *Ex-vivo* metabolomics profiling also led to the discovery and characterization of sedoheptulose-1,7-bisphosphatase, which is involved in riboneogenesis in yeast (75), a radical SAM dehydratase as well as an NADPH-dependent reductase involved in the biosynthesis of the aminoglycoside antibiotic apramycin in *Streptomyces tenebrarius* (76), and the yeast DLD3 protein as a D-2-hydroxyglutarate-pyruvate transhydrogenase potentially involved in the shuttling of reducing equivalents from cytosolic NADH to the mitochondrial electron transfer chain (77). Most recently, the human FGGY protein and its yeast ortholog were also identified as D-ribulokinases using this experimental approach as a starting point (78). Those two proteins are still contained in the unknown enzyme lists generated in this study (Supplementary Table S3) and were correctly predicted to act as D-ribulokinases by the BRENDA and KEGG HMM methods.

The success of this metabolomics-based approach is mainly constrained by the technical limitations that still distinguishes this younger 'omics' technique from the more well established genomics and proteomics methods (79). Although advanced metabolomics techniques can be very sensitive, metabolites involved in the reaction of interest may be present below detection levels in the analyzed extracts if they are unstable for example. Even if a metabolite is reasonably stable, it can be missed depending on the extraction method used. Comprehensive studies will include the use of multiple extraction procedures to balance this risk of extraction bias. If partial knowledge of substrate identity and properties is available, the extraction and analytical protocols should be selected and optimized accordingly. Furthermore, even if a given metabolite is detectable, it may not be identified based on retention time and mass spectrum or even accurate mass. Hyphenated techniques like MS/NMR are partially solving the unknown compound identification problem (80) and thereby may increase the chances of successful enzyme function discovery via metabolomics-based approaches. As each separation and detection method has particular strengths and weaknesses, the use of different separation techniques with different detection methods broadens the spectrum of metabolites that can be measured (81), which is obviously an additional determining factor for success, especially for non-hypothesis driven projects. Finally, the outcomes of the *ex-vivo* metabolomics approach can be heavily influenced by media composition and other environmental conditions that prevail during the '*in-vivo*' part of the experiment. Growth conditions may have to be varied from standard conditions to detect a metabolic phenotype caused by the investigated enzyme deficiency.

It should be noted here that the combination of metabolomics techniques with stable isotope labeling can be extremely useful in bridging the gap between the identification of an enzyme's molecular function and the understanding of its role in cellular metabolism. Experiments with living cells using carefully chosen labeled metabolic precursors can help to uncover or confirm the metabolic pathway in which the newly identified enzyme participates. In addition, stable isotope labeling-assisted metabolomics experiments enable metabolic flux studies which can become indispensable when chemical or genetic inactivation of target enzymes do not lead to changes in substrate or product pool sizes, but only affect the flux through the metabolic pathway in which they are involved. These approaches are not discussed in more detail here as they have been recently reviewed in the context of metabolic pathway discovery (71).

Activity-based protein profiling

The Activity-Based Protein Profiling (ABPP) approach uses small-molecule probes to determine the functional state of enzymes in diverse biological systems, including cell/tissue extracts, living cells or even whole organisms (82–84). An ABPP probe includes three major components: the first is a reactive group that forms a covalent bond with a catalytically active site of enzymes that have common structural and/or reactive properties. The second is a binding group that often resembles the natural substrates of the enzymes and the third is a reporter tag that is commonly a fluorophore or biotin and allows for the detection and enrichment/identification of probe-labeled enzymes. Detection of labeled enzymes is achieved by gel electrophoresis and in-gel fluorescence scanning or LC–MS (85). The major advantage of ABPP is that only the catalytically active forms of the targeted enzymes are detected. This provides a better representation of the physiological activity of these enzymes, especially in contrast to protein abundance where the assumption is that abundance is directly proportional to activity and events such as posttranslational modifications are ignored. Often this method is performed using cell extracts (e.g. control versus treated or WT versus KO) where the extracted proteins are treated with a specific probe of interest. The probed extracts are next run on a gel or purified chromatographically for detection, quantification and/or enrichment. The proteins that are covalently bound to the detected probe can be further analyzed by LC–MS/MS for sequence identification. The major disadvantages of ABPP include the limited (yet growing) number of probes available and the requirement of advanced chemistry skills to design and prepare the probes. There are ABPP probes available for more than a dozen enzyme classes including serine hydrolases, cysteine/threonine proteases, protein tyrosine phosphatases, monooxygenases and monoamine oxidases (82,86). It should be kept in mind that each of these probes targets a more or less large number of different enzymes belonging to the same enzyme class and not individual enzymes.

An example of a recent successful application of ABPP for enzyme function identification is the characterization of a poorly understood member of the PLA2G4 group of cytosolic phospholipases (87). For over 30 years, an orphan calcium-dependent N-acyltransferase (Ca-NAT) activity producing N-acyl phosphatidylethanolamines (NAPEs) from dog heart and brain tissue was known, but the encoding gene had remained mysterious. NAPEs are precursors for N-acyl ethanolamine (NAE) lipid transmitters; the biosynthesis of these compounds is not well understood (87). In order to identify the protein sequence responsible for Ca-NAT activity, detergent-solubilized mouse brain membrane lysate fractions were assayed for this activity and serine hydrolase content was measured in these

same fractions by ABPP using a fluorophosphonate probe. Subsequent LC-MS/MS analysis identified 58 serine hydrolases; the candidate with the strongest correlation coefficient with Ca-NAT activity (PLA2G4E) was expressed transiently in HEK293T cells. PLA2G4E-transfected cell extracts were found to demonstrate much higher Ca-NAT activity compared to control extracts. Additionally, NAPEs were found to be produced in high amounts in PLA2G4E-transfected cells compared to control cells (87). Taken together, these results, initiated by ABPP, provide strong evidence for PLA2G4E being responsible for the elusive Ca-NAT activity.

Structural biology-based approaches

Members of the Enzyme Function Initiative (EFI) have assembled to address the issue of unknown and misannotated proteins (88) based on a sequence/structure-based strategy. With a focus on enzymes, they highlight the fact that commonplace bioinformatics techniques often provide broad clues of functionality but rarely give information about the specific reaction that is catalyzed (88). The EFI advocates for the use of computational strategies to guide experimental verification. Their process begins with the collection of sequences that are clustered into probable isofunctional groups by bioinformatics analyses where a putative function can be investigated by structural determination, structural modeling and docking, and biochemical experimentation. For proteins without experimentally determined structures, homology-based modeling methods can be used if appropriate template structures are available. Computational docking methods provide functional clues by suggesting substrates and ligands for subsequent biochemical experimentation (88,89). This approach has led to numerous enzyme function identifications, including a pterin deaminase from *Agrobacterium radiobacter K84* (90), the discovery of a catabolic pathway for proline betaine in *Paracoccus denitrificans* and *Rhodobacter sphaeroides* (91), and a sesquiterpene synthase from *Streptomyces clavuligerus* (92); more examples can be found on the EFI website (http://enzymefunction.org/). Intrinsic limitations of this approach range from inadequate algorithms for *in silico* substrate docking and homology modeling to incomplete metabolite libraries for virtual screening and incomplete structural coverage of putative enzyme families (88,89).

An alternative structure-based approach involves incubating protein crystals with metabolite cocktails. This strategy relies upon the principle that proteins have evolved to interact proficiently with their partner compounds at physiological concentrations while ignoring unrelated metabolites (93). In practice, this approach comprises two consecutive phases. The first involves screening metabolite cocktails containing structurally related metabolites with the protein crystal of interest. Positive hits are determined using X-ray diffraction analysis of the protein crystals soaked with these multicompound cocktails. This step allows for detection of low affinity binding of ligand analogs. Once a particular class of compounds is found to interact with the protein crystal, additional metabolites within that compound class are screened individually in a second phase to identify derivatives that bind with the highest affinity (93). This metabolite cocktail screening approach led Shumilin *et al.* (93) to identify a bacterial carbohydrate kinase domain containing protein of unknown function as an ATP/ADP-dependent NAD(P)H-hydrate dehydratase, a discovery made independently by Marbaix et al. (48) using the classical biochemical approach (purification of the enzyme activity from yeast extracts followed by protein sequence identification by tandem MS/MS).

Comparative genomics (guilt by association) methods

The comparative genomics-based approach to determine the function of an unknown gene relies on the integration of various types of genomic and post-genomic information that can be extracted from biological databases (13). The guilt by association (GBA) principle is used to infer the function of an unknown gene through its association (by genomic and/or post-genomic evidence) with known genes (94). It should be noted that GBA can be applied on a gene-by-gene or high-throughput basis. In this section we focus on gene-by-gene, as the manual involvement to determine functional associations (13) often still yields better hypotheses than those where GBA is used in a high-throughput manner (95). One of the most powerful means to accumulate genomic evidence for a gene function is through cross-kingdom comparative genomics (Figure 3). Particularly useful is prokaryotic gene clustering, where functionally related genes often occur in operons. Observing cases where an unknown gene of interest is found in an operon of a known pathway can greatly assist in hypothesis generation (96). Gene fusions can imply functional relatedness, conferring potentially a selective advantage to the cell by decreasing the regulatory load for a common biological process (97); the fusion of a gene of unknown function to a gene of known function can in certain cases lead to strong functional predictions for the unknown gene (98).

Post-genomic evidence such as co-expression data and 'metabolic snapshots' can also be useful to assist with functional hypothesis generation (99,100). Genes involved in similar biological processes have a higher probability of being expressed at the same time and/or in the same subcellular compartments or tissues. Already >15 years ago, it was suggested that metabolomic profiling of yeast deletion mutants (designated above as '*ex-vivo* metabolomic profiling') could assist with functional assignments based on the GBA principle, by comparing metabolic snapshots of strains deleted for unknown genes with the ones of strains deleted for known genes (100). Databases such as the STRING database (101) include known and predicted protein-protein interactions, which can also be used to derive functional associations. The comparative analysis of phenotypes constitutes another type of post-genomic information that has recently been proposed as a guide for gene function identification (more particularly in the context of disease research) through the definition of phenologs, i.e. orthologous phenotypes between organisms that are based on overlapping sets of orthologous genes associated with each phenotype (102–104). Finally, promiscuous activities of a known enzyme can assist with inferring function of homologous enzymes. Within a superfamily of enzymes, promiscuous activities that are detected in one family can

correspond to the native activity of enzymes in a related family, and vice versa (105).

The comparative genomics-based predictions represent all probabilistic lines of evidence pointing to the function of a candidate gene, which may or may not be correct. Gathering multiple lines of evidence converging on the same prediction increases the strength of this prediction. The fact remains that experimental validation is eventually required to confirm or reject the functional hypothesis. Some recent examples of comparative genomics-assisted enzyme function identifications include vertebrate hydroxylysine kinase and ammoniophospholyases that act on 5-phosphohydroxy-L-lysine and phosphoethanolamine (106), a specific glutamine transaminase and ω-amidase acting in the methionine salvage pathway in bacteria and plants (107), and bacterial and plant thiamin monophosphate phosphatases which catalyze the penultimate step in thiamin diphosphate synthesis (108).

Gap-filling or network extension methods based on metabolic network reconstructions

This approach relies on the combined knowledge of biochemical activity and metabolic pathways for a given organism or cell type. Such information is also reconciled with biological context through knowledge of transporters and subcellular compartmentalization. Specifically, a genome-scale metabolic network is a representation of the metabolic reactions that a given organism is known to exhibit and is one of several computational approaches in biological network analysis (109). One main goal of metabolic reconstructions is to enable detailed investigations of genotype-phenotype relationships. Such genome-scale metabolic networks have been constructed for several model organisms, notably *S. cerevisiae* (110) and *H. sapiens* (36). In addition, various multi-organism databases (MetaCyc (111), KEGG), but also meta-networks unifying the metabolic reactions from multiple sources (BKM-react (112), MetRxn (113), MNXRef (114)) facilitate the reconstruction of metabolic networks also for other organisms of interest.

Genome-scale metabolic networks allow for a systematic evaluation of missing enzymes by comparing experimentally observed metabolites to *in silico* predictions of their production. Metabolites that are disconnected from the rest of the network and dead-end metabolites (metabolites either only produced or only consumed) are strong indicators for unknown enzymes or missing annotations (Figure 3). Several gap-filling algorithms were proposed to find pathways connecting these metabolites with the rest of the network (23,115,116). Starting points for these algorithms are the organism-specific metabolic networks, the growth conditions defining the uptake of metabolites and the aforementioned multi-organism reaction networks. Reactions from multi-organism databases are added to the network such that the observed metabolites can be produced. Since the solutions are not necessarily unique, different heuristics are employed to pick biologically meaningful reactions. The network extension method described in Christian *et al.* (23) directly uses the enzyme-specific HMMs, introduced in the previous section, to prioritize solutions containing enzymes likely to be encoded by the genome of the organism of interest.

Recon 2 is a genome scale human metabolic reconstruction (36). It's predecessor, Recon 1, was used to identify metabolic gaps in the form of blocked reactions (i.e. reactions that do not carry flux) (117). Such reactions involve dead-end metabolites, whose anabolic or catabolic route within human metabolism is not clear. The gap-filling algorithm SMILEY (115) was used to generate hypotheses on possible human metabolic reactions that would reconcile these dead-end metabolites with the reconstruction (117). By focusing on the orphan (i.e. detected in humans, but absent from the human metabolic network reconstruction) metabolite gluconate, the investigators were able to identify a human candidate gene which was then experimentally validated to encode a gluconokinase (118).

All the gap-filling algorithms are constrained by the knowledge contained in the multi-organism networks, i.e. reactions that have not been described in the literature and included in these networks cannot be identified. Thus, gaps that cannot be filled with these algorithms hint at unknown enzymes that are likely to escape homology-based identifications and purely computational approaches.

Metabolic quantitative trait loci

Another method that can be envisaged to progress in the field of enzyme function discovery is based on combined advances in genomics and metabolomics and involves the identification of metabolic quantitative trait loci (mQTLs). The goal of QTL analysis is to associate phenotypic traits that vary quantitatively across members of a species with genetic loci that govern this variation. Briefly, linkage mapping provides a relative distance between the genetic loci attributed with a particular trait and known molecular markers. In the case of mQTL studies, the phenotypic traits of interest correspond to specific metabolite levels. In practical terms, studies in yeast and plants have exploited mapping populations and metabolomics techniques to study metabolism and metabolic regulations on a broader scale (119–122). In addition, mQTL analysis has been used to gain new insights into the genetic architecture underlying specific metabolic traits such as corn earworm resistance in *Zea mays* (123), seed oil content in *A. thaliana* (124), and the production of wine aroma compounds by yeast (125).

One of the major bottlenecks encountered during mQTL analysis is causal gene identification and validation, which are the key steps in tying genotype and phenotypic variation together. Depending on the size of the mapping population and the number of genetic markers, identified QTLs can contain a more or less long list of candidate genes in the confidence interval, among which a reasonable number of genes then have to be chosen for validation. Selection criteria are here often heavily based on gene function annotation. Not surprisingly, given the unknown protein/enzyme problem discussed as a main topic of this review, many of the candidate genes found in mQTL studies are annotated with putative or unknown functions (121). However, as mQTL studies are powerful enough to locate genes associated with specific metabolite level changes, they represent an interesting tool to address the unknown enzyme problem. If this is

actually the objective, genes of unknown function with putative metabolic function, as suggested by the linkage mapping analysis, can become very interesting candidate genes to follow up on (using aforementioned approaches) for enzyme function discovery, the mQTL study serving as a first step in functional hypothesis generation.

In addition to critically reviewing approaches currently used for enzyme function identification, a goal of this section was to provide a list of tools at the disposal of the scientist interested in this task. Each of the nine approaches described here has its strengths and weaknesses. However, be it alone or in combination, these strategies will allow for most enzyme identification projects to be addressed (Figure 4), each doing their part to confront the unknown enzyme problem. While some of these strategies rely at least partially on 'omics' or systems biology approaches, none of them currently allows to overcome the problem that each enzyme is more or less unique and will require a dedicated path of discovery; on this path a researcher's intuition is often key and fortune will continue to favor 'the prepared mind'. In other words, while enzyme function identification at an accelerated parallel scale seems like a desirable aim in light of the number of persisting unknown enzymes, it remains currently an unrealistic one.

POSSIBLE ROLES FOR THE REMAINING ENZYMES OF UNKNOWN FUNCTION AND IMPLICATIONS OF THE UNKNOWN ENZYME PROBLEM FOR MODERN BIOLOGICAL RESEARCH

We have shown that more than 30% of the proteins of unknown function in yeast and humans appear to act as enzymes (Figures 1 and 2). In this final part of the review, we will discuss the possible roles hidden within this catalytic dark matter. A fraction of the remaining unknown enzymes has to correspond to current orphan enzymes. A recent estimate is that 22% of enzyme activities across all living species (and 26% of enzyme activities in eukaryotes) classified with Enzyme Commission (EC) numbers—as maintained by the International Union of Biochemistry and Molecular Biology—are not associated with a protein sequence (126). These are referred to as orphan enzymes (in contrast to orphan genes, which are defined as genes that lack homologs in other lineages), and although the proportion of orphan enzymes has decreased from 38% in 2003 to 22% in 2014, that still leaves >1100 known enzyme activities to which a protein sequence must be assigned (126).

As indicated in the previous section, the hundreds of proteins identified through our bioinformatics analyses as putative 'metabolic' enzymes in *S. cerevisiae* and humans suggest that there must be facets of cellular metabolism that we still ignore. While secondary metabolism in plants and bacteria certainly represents a gold mine for enzyme discovery (39), functions for the putative metabolic enzymes in *S. cerevisiae* and humans, typically considered as having a limited secondary metabolism, have to be searched for elsewhere. As it is very unlikely that all these catalytic unknowns correspond to remaining gaps in the by now quite extensively characterized primary metabolism, we hypothesize that a significant proportion of them are involved in metabolite damage control. The latter process can be considered as a sort of support system that is required for primary metabolism to function correctly, despite interfering side reactions catalyzed by many core metabolic enzymes (79,127,128). Indeed, the paradigm that enzymes are 'perfect' catalysts is shifting as chemical biology increasingly recognizes that most enzymes are promiscuous (129–131) and can form useless or toxic side products (50,128,132). Evidence also exists for metabolites being damaged purely by chemical reactions (133–135). It seems that enzyme errors and chemical damage are ancient metabolic problems and the evolutionary driving force for metabolite repair systems (128,136). Metabolite damage appears to be so ubiquitous and pervasive that it has also recently been associated with aging theory (137). Finally, evidence is mounting that metabolite damage and/or repair deficiencies play important roles in inborn metabolic diseases (2-hydroxyglutaric acidurias (127,138), NAXE deficiency (139)) and even cancer (128,140). Based on a survey of the enzyme function identifications in *Escherichia coli* that occurred between 1998 and 2015, it was recently estimated that ∼15% of the remaining enzymes of unknown function may correspond to metabolite repair enzymes (79). Given the wide conservation of many metabolite damage control systems, this 15% value can tentatively be extrapolated to other organisms, thus indicating that more than 35 and 75 metabolite repair enzymes remain to be discovered in yeast and humans (based on the at least 241 and 506 metabolic enzymes of unknown function estimated to persist in these organisms in the present review).

Although enzyme discoveries in core metabolic pathways *per se* are becoming more exceptional as the genetic mapping of the reactions involved is getting more and more complete, recent examples show that such findings are still possible. GDP-L-galactose phosphorylase, a key enzyme in a primary plant metabolic pathway, namely the vitamin C synthesis pathway, was only identified in 2007 (56); in fact this whole metabolic pathway had only been proposed in 1998 (141), followed by the successive molecular identifications of the participating enzymes (142). Very recently, the human ISPD protein, which is mutated in a form of congenital muscular dystrophy, was identified as a CDP-ribitol pyrophosphorylase (143,144). ISPD thus forms CDP-ribitol, a nucleotide-alditol that had not been known before to exist in human cells and that serves as a donor of ribitol phosphate in the glycosylation of alpha-dystroglycan. It remains unclear how ribitol-5-P, the substrate of ISPD for CDP-ribitol formation, is produced in human cells, leaving room for another enzyme identification in primary metabolism.

Moving away from enzymes with probable roles in metabolism, another fraction of the unknown enzymes are likely responsible for posttranslational modifications (PTMs) of proteins and the removal thereof. A landmark review in 2005 (145) described that of the five major categories of PTMs (phosphorylation, acylation, glycosylation, thiol-disulfide chemistry, and alkylation), one of the largest is protein phosphorylation which relies on over 500 protein kinases in the human proteome. Perhaps many of these protein kinases are considered known, however, the enzymes responsible for more newly discovered PTMs often remain enigmatic (146). As an example, the enzyme catalyzing the recently discovered 2-hydroxyisobutyrylation of lysine

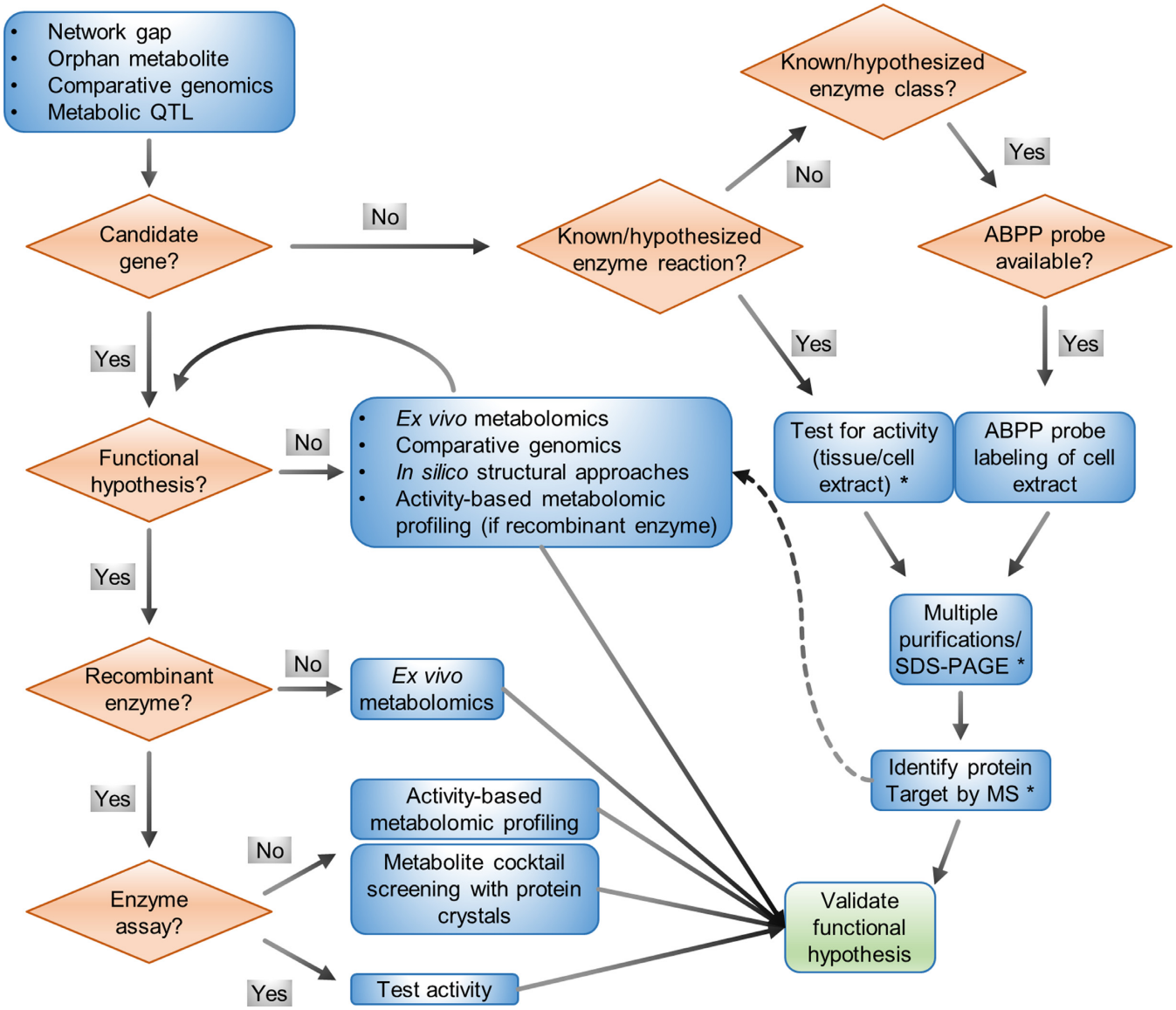

Figure 4. Decision tree to assist enzyme function discovery research. The flowchart illustrates how the different experimental and computational strategies described in this review can be used to find new enzymes. While this chart is not exhaustive, it shows possible paths for enzyme function identification from common starting points, using all the strategies described. The final validation strategy depends on the chosen path, but often consists in demonstrating that the purified enzyme, if available in recombinant or native form, displays the hypothesized activity *in vitro*. Blue boxes represent techniques/approaches or starting points and orange diamonds represent decision points. Asterisks denote intermediate steps of the classical biochemical approach. ABPP, activity-based protein profiling; MS, mass spectrometry; QTL, quantitative trait locus.

residues (K_{hib}) in histones, possibly using hydroxyisobutryl-CoA as a donor, has not been identified yet (147). In the context of enzymatic PTM reversal, *in vitro* experiments have recently suggested that a subset of the JmjC lysine demethylases can also act as methylarginine demethylases (148). Also, bioinformatics approaches have been used to estimate that the human proteome contains 86 deubiquitinating enzymes, with some considered known and others yet to be validated (149).

Functions for the remaining unknown enzymes will probably also continue to be found in the context of posttranscriptional modifications. As an illustration, the roles of A-to-I RNA editing enzymes (or ADARs), which act by converting adenosine to inosine by hydrolytic deamination at the C6 position in coding and non-coding regions of RNAs in vertebrates, remain only partially understood (150,151). Once converted, the translation machinery perceives the inosine as a guanosine, thus pairing it with cytosine. Such editing can result in changes to codons that were not directly encoded in the genome, but the significance of the non-coding region editing remains largely unknown. While ADARs are absent in protozoa, yeast, and plants (150), tRNA modification enzymes have been found to occur in all domains of life (152). The tRNA editing field has seen the discovery of many new modifications in recent years accompanied by the identification of many enzymes responsible for the modifications (152–154). Most recently, it has been shown that bacterial RNAs can be 5′ capped with

NAD, NADH and dephospho-CoA, leading to increased transcript stability (155). The cofactor caps are added when cellular RNA polymerases use these non-canonical nucleotides as initiating molecules during transcription initiation (156). The Nudix phosphohydrolase NudC, whose physiological role had remained unclear, was found to selectively decap NAD(H)-RNA (157).

Additional roles for the unknown enzymes will likely be found to be involved in other aspects of enzyme complexity, including enzymes that have lost their catalytic activity and act as allosteric regulatory proteins (158). Related to this enzyme complexity, it should be reminded here as well that many enzymes and proteins in general have more than one molecular and biological function (moonlighting activities) (159–162), which further increases the functional dark matter encoded by genomes and consequently the discovery potential in this field.

Coming back to metabolism, the assembly of advanced metabolic reconstructions and models (36,110,163) is giving researchers a larger picture of this complex cellular process as a whole. While these reconstructions are not claiming to be complete, some gaps or pathway holes can be problematic. The large amount of unknown proteins represent gaps in the parts lists being used in systems biology and more particularly, the hundreds or thousands of unknown enzymes in sequenced genomes constitute a challenge for genome-wide metabolic models to generate accurate predictions (164). This is a common issue also for metabolic engineering where some of the main goals include the optimized production of value-added chemicals via microbes or the improvement of food supply by either yield or nutritional value. The use of rational design to make such improvements requires the knowledge of the individual proteins that regulate the pathways of interest and the enzymes that participate in them. When there are gaps in knowledge (metabolic regulations, transport reactions, enzymatic reactions, but also enzymatic side reactions), there will be roadblocks on the way to achieve the metabolic engineering goals. Finally, clinical geneticists studying inborn errors of metabolism have been able to take advantage of the decreased cost of sequencing to determine the genetic abnormalities underlying the conditions of their patients. However, when a causal gene is identified and has an unknown or ambiguous function, there may still be no insight into the disease mechanism and into how to develop or improve treatment options for the patient.

While primary metabolism has been studied in depth, and the majority of the low hanging fruit has been picked, enzymes involved in biological fitness will play a leading role in the next wave of metabolic discoveries. Human ingenuity and industriousness are the driving forces behind the sustained success of Moore's Law. These same attributes will be necessary to confront and eventually solve the unknown enzyme problem.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We would like to thank Dr. Paul P. Jung for critical reading of the manuscript and Prof. Rudi Balling for helpful comments. Most of the *in silico* analysis results presented in this article were obtained using the high-performance computing facilities of the University of Luxembourg (http://hpc.uni.lu) (165).

FUNDING

Fonds National de la Recherche Luxembourg [9180195 to K.W.E., 6885527 to C.S.]; Fondation du Pélican de Mie et Pierre Hippert-Faber scholarship (to C.S.); knowledge transfer program through the health transfer initiative of the Luxembourg government (to N.C.); 'le plan Technologies de la Santé par le Gouvernment du Grand-Duché de Luxembourg' through the Luxembourg Centre for Systems Biomedicine, University of Luxembourg (to P.M.). Funding for open access charge: University of Luxembourg.
Conflict of interest statement. None declared.

REFERENCES

1. Moore,G.E. (1965) Cramming more components onto integrated circuits. *Electronics*, **38**, 8.
2. Moore,G.E. (1975) Progress in digitial integrated electronics. *International Electron Devices Meeting*, **21**, 11–13.
3. Reddy,T.B., Thomas,A.D., Stamatis,D., Bertsch,J., Isbandi,M., Jansson,J., Mallajosyula,J., Pagani,I., Lobos,E.A. and Kyrpides,N.C. (2015) The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.*, **43**, D1099–D1106.
4. Schnoes,A.M., Brown,S.D., Dodevski,I. and Babbitt,P.C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.
5. Saghatelian,A. and Cravatt,B.F. (2005) Assignment of protein function in the postgenomic era. *Nat. Chem. Biol.*, **1**, 130–142.
6. Niehaus,T.D., Thamm,A.M., de Crecy-Lagard,V. and Hanson,A.D. (2015) Proteins of unknown biochemical function: a persistent problem and a roadmap to help overcome it. *Plant Physiol.*, **169**, 1436–1442.
7. Hughes,T.R., Robinson,M.D., Mitsakakis,N. and Johnston,M. (2004) The promise of functional genomics: completing the encyclopedia of a cell. *Curr. Opin. Microbiol.*, **7**, 546–554.
8. Hodges,P.E., McKee,A.H., Davis,B.P., Payne,W.E. and Garrels,J.I. (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.*, **27**, 69–73.
9. Osterman,A. and Overbeek,R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.*, **7**, 238–251.
10. Blaby-Haas,C.E. and de Crecy-Lagard,V. (2011) Mining high-throughput experimental data to link gene and function. *Trends Biotechnol.*, **29**, 174–182.
11. Balakrishnan,R., Park,J., Karra,K., Hitz,B.C., Binkley,G., Hong,E.L., Sullivan,J., Micklem,G. and Cherry,J.M. (2012) YeastMine–an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database (Oxford)*, **2012**, bar062.
12. Pena-Castillo,L. and Hughes,T.R. (2007) Why are there still over 1000 uncharacterized yeast genes? *Genetics*, **176**, 7–14.
13. Hanson,A.D., Pribat,A., Waller,J.C. and de Crecy-Lagard,V. (2010) 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list–and how to find it. *Biochem J.*, **425**, 1–11.
14. Gerdes,S., El Yacoubi,B., Bailly,M., Blaby,I.K., Blaby-Haas,C.E., Jeanguenin,L., Lara-Nunez,A., Pribat,A., Waller,J.C., Wilke,A. *et al.* (2011) Synergistic use of plant-prokaryote comparative genomics for functional annotations. *BMC Genomics*, **12**(Suppl. 1), S2.

15. Galperin,M.Y. and Koonin,E.V. (2004) 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.*, **32**, 5452–5463.
16. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
17. Lane,L., Argoud-Puy,G., Britan,A., Cusin,I., Duek,P.D., Evalet,O., Gateau,A., Gaudet,P., Gleizes,A., Masselot,A. *et al.* (2012) neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res.*, **40**, D76–D83.
18. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
19. Brown,S.D. and Babbitt,P.C. (2014) New insights about enzyme evolution from large scale studies of sequence and structure relationships. *J. Biol. Chem.*, **289**, 30221–30228.
20. Goodacre,N.F., Gerloff,D.L. and Uetz,P. (2014) Protein domains of unknown function are essential in bacteria. *MBio*, **5**, doi:10.1128/mBio.00744-13.
21. Kachroo,A.H., Laurent,J.M., Yellman,C.M., Meyer,A.G., Wilke,C.O. and Marcotte,E.M. (2015) Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, **348**, 921–925.
22. Laurent,J.M., Young,J.H., Kachroo,A.H. and Marcotte,E.M. (2016) Efforts to make and apply humanized yeast. *Brief. Funct. Genomics*, **15**, 155–163.
23. Christian,N., May,P., Kempa,S., Handorf,T. and Ebenhoh,O. (2009) An integrative approach towards completing genome-scale metabolic networks. *Mol. Biosyst.*, **5**, 1889–1903.
24. Eddy,S.R. (2004) What is a hidden Markov model? *Nat Biotechnol*, **22**, 1315–1316.
25. Tian,W., Arakaki,A.K. and Skolnick,J. (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.*, **32**, 6226–6239.
26. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
27. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
29. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
30. Chang,A., Schomburg,I., Placzek,S., Jeske,L., Ulbrich,M., Xiao,M., Sensen,C.W. and Schomburg,D. (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.*, **43**, D439–D446.
31. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
32. Finn,R.D., Clements,J., Arndt,W., Miller,B.L., Wheeler,T.J., Schreiber,F., Bateman,A. and Eddy,S.R. (2015) HMMER web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.
33. Jensen,L.J., Gupta,R., Blom,N., Devos,D., Tamames,J., Kesmir,C., Nielsen,H., Staerfeldt,H.H., Rapacki,K., Workman,C. *et al.* (2002) Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, **319**, 1257–1265.
34. Cai,C.Z., Han,L.Y., Ji,Z.L. and Chen,Y.Z. (2004) Enzyme family classification by support vector machines. *Proteins*, **55**, 66–76.
35. Dobson,P.D. and Doig,A.J. (2005) Predicting enzyme class from protein structure without alignments. *J. Mol. Biol.*, **345**, 187–199.
36. Thiele,I., Swainston,N., Fleming,R.M., Hoppe,A., Sahoo,S., Aurich,M.K., Haraldsdottir,H., Mo,M.L., Rolfsson,O., Stobbe,M.D. *et al.* (2013) A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, **31**, 419–425.
37. Aung,H.W., Henry,S.A. and Walker,L.P. (2013) Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind. Biotechnol.*, **9**, 215–228.
38. Green,M.L. and Karp,P.D. (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res.*, **33**, 4035–4039.
39. Tawfik,D.S. and van der Donk,W.A. (2016) Editorial overview: Biocatalysis and biotransformation: esoteric, niche enzymology. *Curr. Opin. Chem. Biol.*, **31**, v–vii.
40. Saito,N., Robert,M., Kitamura,S., Baran,R., Soga,T., Mori,H., Nishioka,T. and Tomita,M. (2006) Metabolomics approach for enzyme discovery. *J. Proteome Res.*, **5**, 1979–1987.
41. Drozak,J., Veiga-da-Cunha,M., Vertommen,D., Stroobant,V. and Van Schaftingen,E. (2010) Molecular identification of carnosine synthase as ATP-grasp domain-containing protein 1 (ATPGD1). *J. Biol. Chem.*, **285**, 9346–9356.
42. Collard,F., Vertommen,D., Constantinescu,S., Buts,L. and Van Schaftingen,E. (2011) Molecular identification of beta-citrylglutamate hydrolase as glutamate carboxypeptidase 3. *J. Biol. Chem.*, **286**, 38220–38230.
43. Wu,L.C., Pfeiffer,D.R., Calhoon,E.A., Madiai,F., Marcucci,G., Liu,S. and Jurkowitz,M.S. (2011) Purification, identification, and cloning of lysoplasmalogenase, the enzyme that catalyzes hydrolysis of the vinyl ether bond of lysoplasmalogen. *J. Biol. Chem.*, **286**, 24916–24930.
44. Pieslinger,A.M., Hoepflinger,M.C. and Tenhaken,R. (2010) Cloning of Glucuronokinase from Arabidopsis thaliana, the last missing enzyme of the myo-inositol oxygenase pathway to nucleotide sugars. *J. Biol. Chem.*, **285**, 2902–2910.
45. Alcolombri,U., Ben-Dor,S., Feldmesser,E., Levin,Y., Tawfik,D.S. and Vardi,A. (2015) MARINE SULFUR CYCLE. Identification of the algal dimethyl sulfide-releasing enzyme: A missing link in the marine sulfur cycle. *Science*, **348**, 1466–1469.
46. Qiu,H., Geng,A., Zhu,D., Le,Y., Wu,J., Chow,N., Wu,J.H. and Sun,J. (2015) Purification and characterization of a hemocyanin (Hemo1) with potential lignin-modification activities from the wood-feeding termite, Coptotermes formosanus Shiraki. *Appl. Biochem. Biotechnol.*, **175**, 687–697.
47. Linster,C.L., Noel,G., Stroobant,V., Vertommen,D., Vincent,M.F., Bommer,G.T., Veiga-da-Cunha,M. and Van Schaftingen,E. (2011) Ethylmalonyl-CoA decarboxylase, a new enzyme involved in metabolite proofreading. *J. Biol. Chem.*, **286**, 42992–43003.
48. Marbaix,A.Y., Noel,G., Detroux,A.M., Vertommen,D., Van Schaftingen,E. and Linster,C.L. (2011) Extremely conserved ATP- or ADP-dependent enzymatic system for nicotinamide nucleotide repair. *J. Biol. Chem.*, **286**, 41246–41252.
49. Veiga-da-Cunha,M., Chevalier,N., Stroobant,V., Vertommen,D. and Van Schaftingen,E. (2014) Metabolite proofreading in carnosine and homocarnosine synthesis: molecular identification of PM20D2 as beta-alanyl-lysine dipeptidase. *J. Biol. Chem.*, **289**, 19726–19736.
50. Collard,F., Baldin,F., Gerin,I., Bolsee,J., Noel,G., Graff,J., Veiga-da-Cunha,M., Stroobant,V., Vertommen,D., Houddane,A. *et al.* (2016) A conserved phosphatase destroys toxic glycolytic side products in mammals and yeast. *Nat. Chem. Biol.*, **12**, 601–607.
51. Earnshaw,W.C. (2013) Deducing protein function by forensic integrative cell biology. *PLoS Biol.*, **11**, e1001742.
52. Elshourbagy,N.A., Lee,J.A., Korman,D.R., Nuthalaganti,P., Sylvester,D.R., Dilella,A.G., Sutiphong,J.A. and Kumar,C.S. (1992) Molecular cloning and characterization of the major endothelin receptor subtype in porcine cerebellum. *Mol. Pharmacol.*, **41**, 465–473.
53. Wiame,E., Tyteca,D., Pierrot,N., Collard,F., Amyere,M., Noel,G., Desmedt,J., Nassogne,M.C., Vikkula,M., Octave,J.N. *et al.* (2009) Molecular identification of aspartate N-acetyltransferase and its mutation in hypoacetylaspartia. *Biochem. J.*, **425**, 127–136.
54. Michelucci,A., Cordes,T., Ghelfi,J., Pailot,A., Reiling,N., Goldmann,O., Binz,T., Wegner,A., Tallam,A., Rausell,A. *et al.* (2013) Immune-responsive gene 1 protein links metabolism to immunity by catalyzing itaconic acid production. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 7820–7825.
55. Watschinger,K., Keller,M.A., Golderer,G., Hermann,M., Maglione,M., Sarg,B., Lindner,H.H., Hermetter,A., Werner-Felmayer,G., Konrat,R. *et al.* (2010) Identification of the gene encoding alkylglycerol monooxygenase defines a third class of tetrahydrobiopterin-dependent enzymes. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 13672–13677.

56. Linster,C.L., Gomez,T.A., Christensen,K.C., Adler,L.N., Young,B.D., Brenner,C. and Clarke,S.G. (2007) Arabidopsis VTC2 encodes a GDP-L-galactose phosphorylase, the last unknown enzyme in the Smirnoff-Wheeler pathway to ascorbic acid in plants. *J. Biol. Chem.*, **282**, 18879–18885.
57. Jaisson,S., Veiga-da-Cunha,M. and Van Schaftingen,E. (2009) Molecular identification of omega-amidase, the enzyme that is functionally coupled with glutamine transaminases, as the putative tumor suppressor Nit2. *Biochimie*, **91**, 1066–1071.
58. Collard,F., Stroobant,V., Lamosa,P., Kapanda,C.N., Lambert,D.M., Muccioli,G.G., Poupaert,J.H., Opperdoes,F. and Van Schaftingen,E. (2010) Molecular identification of N-acetylaspartylglutamate synthase and beta-citrylglutamate synthase. *J. Biol. Chem.*, **285**, 29826–29833.
59. Shindou,H., Hishikawa,D., Nakanishi,H., Harayama,T., Ishii,S., Taguchi,R. and Shimizu,T. (2007) A single enzyme catalyzes both platelet-activating factor production and membrane biogenesis of inflammatory cells. Cloning and characterization of acetyl-CoA:LYSO-PAF acetyltransferase. *J. Biol. Chem.*, **282**, 6532–6539.
60. Watschinger,K. and Werner,E.R. (2013) Orphan enzymes in ether lipid metabolism. *Biochimie*, **95**, 59–65.
61. Adler,L.N., Gomez,T.A., Clarke,S.G. and Linster,C.L. (2011) A novel GDP-D-glucose phosphorylase involved in quality control of the nucleoside diphosphate sugar pool in Caenorhabditis elegans and mammals. *J. Biol. Chem.*, **286**, 21511–21523.
62. Strittmatter,L., Li,Y., Nakatsuka,N.J., Calvo,S.E., Grabarek,Z. and Mootha,V.K. (2014) CLYBL is a polymorphic human enzyme with malate synthase and beta-methylmalate synthase activity. *Hum. Mol. Genet.*, **23**, 2313–2323.
63. Peracchi,A., Veiga-da-Cunha,M., Kuhara,T., Ellens,K.W., Paczia,N., Stroobant,V., Seliga,A.K., Marlaire,S., Jaisson,S., Bommer,G.T. *et al.* (2017) Nit1 is a metabolite repair enzyme that hydrolyzes deaminated glutathione. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E3233–E3242.
64. Grocholski,T., Dinis,P., Niiranen,L., Niemi,J. and Metsa-Ketela,M. (2015) Divergent evolution of an atypical S-adenosyl-l-methionine-dependent monooxygenase involved in anthracycline biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 9866–9871.
65. Seffernick,J.L., de Souza,M.L., Sadowsky,M.J. and Wackett,L.P. (2001) Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. *J. Bacteriol.*, **183**, 2405–2410.
66. de Carvalho,L.P., Zhao,H., Dickinson,C.E., Arango,N.M., Lima,C.D., Fischer,S.M., Ouerfelli,O., Nathan,C. and Rhee,K.Y. (2010) Activity-based metabolomic profiling of enzymatic function: identification of Rv1248c as a mycobacterial 2-hydroxy-3-oxoadipate synthase. *Chem. Biol.*, **17**, 323–332.
67. Saito,N., Robert,M., Kochi,H., Matsuo,G., Kakazu,Y., Soga,T. and Tomita,M. (2009) Metabolite profiling reveals YihU as a novel hydroxybutyrate dehydrogenase for alternative succinic semialdehyde metabolism in Escherichia coli. *J. Biol. Chem.*, **284**, 16442–16451.
68. Denger,K., Weiss,M., Felux,A.K., Schneider,A., Mayer,C., Spiteller,D., Huhn,T., Cook,A.M. and Schleheck,D. (2014) Sulphoglycolysis in Escherichia coli K-12 closes a gap in the biogeochemical sulphur cycle. *Nature*, **507**, 114–117.
69. Tian,J., Bryk,R., Itoh,M., Suematsu,M. and Nathan,C. (2005) Variant tricarboxylic acid cycle in Mycobacterium tuberculosis: identification of alpha-ketoglutarate decarboxylase. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 10670–10675.
70. Wagner,T., Bellinzoni,M., Wehenkel,A., O'Hare,H.M. and Alzari,P.M. (2011) Functional plasticity and allosteric regulation of alpha-ketoglutarate decarboxylase in central mycobacterial metabolism. *Chem. Biol.*, **18**, 1011–1020.
71. Prosser,G.A., Larrouy-Maumus,G. and de Carvalho,L.P. (2014) Metabolomic strategies for the identification of new enzyme functions and metabolic pathways. *EMBO Rep.*, **15**, 657–669.
72. Ewald,J.C., Matt,T. and Zamboni,N. (2013) The integrated response of primary metabolites to gene deletions and the environment. *Mol. Biosyst.*, **9**, 440–446.
73. Ishii,N., Nakahigashi,K., Baba,T., Robert,M., Soga,T., Kanai,A., Hirasawa,T., Naba,M., Hirai,K., Hoque,A. *et al.* (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science*, **316**, 593–597.
74. Long,J.Z., Cisar,J.S., Milliken,D., Niessen,S., Wang,C., Trauger,S.A., Siuzdak,G. and Cravatt,B.F. (2011) Metabolomics annotates ABHD3 as a physiologic regulator of medium-chain phospholipids. *Nat. Chem. Biol.*, **7**, 763–765.
75. Clasquin,M.F., Melamud,E., Singer,A., Gooding,J.R., Xu,X., Dong,A., Cui,H., Campagna,S.R., Savchenko,A., Yakunin,A.F. *et al.* (2011) Riboneogenesis in yeast. *Cell*, **145**, 969–980.
76. Lv,M., Ji,X., Zhao,J., Li,Y., Zhang,C., Su,L., Ding,W., Deng,Z., Yu,Y. and Zhang,Q. (2016) Characterization of a C3 deoxygenation pathway reveals a key branch point in aminoglycoside biosynthesis. *J. Am. Chem. Soc.*, **138**, 6427–6435.
77. Becker-Kettern,J., Paczia,N., Conrotte,J.F., Kay,D.P., Guignard,C., Jung,P.P. and Linster,C.L. (2016) Saccharomyces cerevisiae Forms D-2-Hydroxyglutarate and Couples Its Degradation to D-Lactate Formation via a Cytosolic Transhydrogenase. *J. Biol. Chem.*, **291**, 6036–6058.
78. Singh,C., Glaab,E. and Linster,C.L. (2017) Molecular identification of d-ribulokinase in budding yeast and mammals. *J. Biol. Chem.*, **292**, 1005–1028.
79. Hanson,A.D., Henry,C.S., Fiehn,O. and de Crecy-Lagard,V. (2016) Metabolite damage and metabolite damage control in plants. *Annu. Rev. Plant Biol.*, **67**, 131–152.
80. Bingol,K., Bruschweiler-Li,L., Yu,C., Somogyi,A., Zhang,F. and Bruschweiler,R. (2015) Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. *Anal. Chem.*, **87**, 3864–3870.
81. Quanbeck,S.M., Brachova,L., Campbell,A.A., Guan,X., Perera,A., He,K., Rhee,S.Y., Bais,P., Dickerson,J.A., Dixon,P. *et al.* (2012) Metabolomics as a hypothesis-generating functional genomics tool for the annotation of Arabidopsis thaliana genes of "Unknown Function". *Front. Plant Sci.*, **3**, 15.
82. Cravatt,B.F., Wright,A.T. and Kozarich,J.W. (2008) Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. *Annu. Rev. Biochem.*, **77**, 383–414.
83. Heal,W.P., Dang,T.H. and Tate,E.W. (2011) Activity-based probes: discovering new biology and new drug targets. *Chem. Soc. Rev.*, **40**, 246–257.
84. Niphakis,M.J. and Cravatt,B.F. (2014) Enzyme inhibitor discovery by activity-based protein profiling. *Annu. Rev. Biochem.*, **83**, 341–377.
85. Galmozzi,A., Dominguez,E., Cravatt,B.F. and Saez,E. (2014) Application of activity-based protein profiling to study enzyme function in adipocytes. *Methods Enzymol.*, **538**, 151–169.
86. Willems,L.I., Overkleeft,H.S. and van Kasteren,S.I. (2014) Current developments in activity-based protein profiling. *Bioconjug. Chem.*, **25**, 1181–1191.
87. Ogura,Y., Parsons,W.H., Kamat,S.S. and Cravatt,B.F. (2016) A calcium-dependent acyltransferase that produces N-acyl phosphatidylethanolamines. *Nat. Chem. Biol.*, **12**, 669–671.
88. Gerlt,J.A., Allen,K.N., Almo,S.C., Armstrong,R.N., Babbitt,P.C., Cronan,J.E., Dunaway-Mariano,D., Imker,H.J., Jacobson,M.P., Minor,W. *et al.* (2011) The enzyme function initiative. *Biochemistry*, **50**, 9950–9962.
89. Jacobson,M.P., Kalyanaraman,C., Zhao,S. and Tian,B. (2014) Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. *Trends Biochem Sci.*, **39**, 363–371.
90. Fan,H., Hitchcock,D.S., Seidel,R.D. 2nd, Hillerich,B., Lin,H., Almo,S.C., Sali,A., Shoichet,B.K. and Raushel,F.M. (2013) Assignment of pterin deaminase activity to an enzyme of unknown function guided by homology modeling and docking. *J. Am. Chem. Soc.*, **135**, 795–803.
91. Kumar,R., Zhao,S., Vetting,M.W., Wood,B.M., Sakai,A., Cho,K., Solbiati,J., Almo,S.C., Sweedler,J.V., Jacobson,M.P. *et al.* (2014) Prediction and biochemical demonstration of a catabolic pathway for the osmoprotectant proline betaine. *MBio*, **5**, doi:10.1128/mBio.00933-13.
92. Chow,J.Y., Tian,B.X., Ramamoorthy,G., Hillerich,B.S., Seidel,R.D., Almo,S.C., Jacobson,M.P. and Poulter,C.D. (2015) Computational-guided discovery and characterization of a sesquiterpene synthase from Streptomyces clavuligerus. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 5661–5666.

93. Shumilin,I.A., Cymborowski,M., Chertihin,O., Jha,K.N., Herr,J.C., Lesley,S.A., Joachimiak,A. and Minor,W. (2012) Identification of unknown protein function using metabolite cocktail screening. *Structure*, **20**, 1715–1725.
94. Aravind,L. (2000) Guilt by association: contextual information in genome analysis. *Genome Res.*, **10**, 1074–1077.
95. Gillis,J. and Pavlidis,P. (2012) "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Comput. Biol.*, **8**, e1002444.
96. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 2896–2901.
97. Enright,A.J. and Ouzounis,C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.*, **2**, RESEARCH0034.
98. Henry,C.S., Lerma-Ortiz,C., Gerdes,S.Y., Mullen,J.D., Colasanti,R., Zhukov,A., Frelin,O., Thiaville,J.J., Zallot,R., Niehaus,T.D. *et al.* (2016) Systematic identification and analysis of frequent gene fusion events in metabolic pathways. *BMC Genomics*, **17**, 473.
99. Okamura,Y., Aoki,Y., Obayashi,T., Tadaka,S., Ito,S., Narise,T. and Kinoshita,K. (2015) COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.*, **43**, D82–D86.
100. Raamsdonk,L.M., Teusink,B., Broadhurst,D., Zhang,N., Hayes,A., Walsh,M.C., Berden,J.A., Brindle,K.M., Kell,D.B., Rowland,J.J. *et al.* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.*, **19**, 45–50.
101. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
102. Bubier,J.A., Wilcox,T.D., Jay,J.J., Langston,M.A., Baker,E.J. and Chesler,E.J. (2016) Cross-species integrative functional genomics in GeneWeaver reveals a role for Pafah1b1 in altered response to alcohol. *Front. Behav. Neurosci.*, **10**, 1.
103. McGary,K.L., Park,T.J., Woods,J.O., Cha,H.J., Wallingford,J.B. and Marcotte,E.M. (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6544–6549.
104. Kochinke,K., Zweier,C., Nijhof,B., Fenckova,M., Cizek,P., Honti,F., Keerthikumar,S., Oortveld,M.A., Kleefstra,T., Kramer,J.M. *et al.* (2016) Systematic phenomics analysis deconvolutes genes mutated in intellectual disability into biologically coherent modules. *Am. J. Hum. Genet.*, **98**, 149–164.
105. Afriat,L., Roodveldt,C., Manco,G. and Tawfik,D.S. (2006) The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. *Biochemistry*, **45**, 13677–13686.
106. Veiga-da-Cunha,M., Hadi,F., Balligand,T., Stroobant,V. and Van Schaftingen,E. (2012) Molecular identification of hydroxylysine kinase and of ammoniophospholyases acting on 5-phosphohydroxy-L-lysine and phosphoethanolamine. *J. Biol. Chem.*, **287**, 7246–7255.
107. Ellens,K.W., Richardson,L.G., Frelin,O., Collins,J., Ribeiro,C.L., Hsieh,Y.F., Mullen,R.T. and Hanson,A.D. (2015) Evidence that glutamine transaminase and omega-amidase potentially act in tandem to close the methionine salvage cycle in bacteria and plants. *Phytochemistry*, **113**, 160–169.
108. Hasnain,G., Roje,S., Sa,N., Zallot,R., Ziemak,M.J., de Crecy-Lagard,V., Gregory,J.F. 3rd and Hanson,A.D. (2016) Bacterial and plant HAD enzymes catalyse a missing phosphatase step in thiamin diphosphate biosynthesis. *Biochem. J.*, **473**, 157–166.
109. Pfau,T., Christian,N. and Ebenhoh,O. (2011) Systems approaches to modelling pathways and networks. *Brief. Funct. Genomics*, **10**, 266–279.
110. Heavner,B.D., Smallbone,K., Price,N.D. and Walker,L.P. (2013) Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database (Oxford)*, **2013**, bat059.
111. Caspi,R., Altman,T., Billington,R., Dreher,K., Foerster,H., Fulcher,C.A., Holland,T.A., Keseler,I.M., Kothari,A., Kubo,A. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**, D459–D471.
112. Lang,M., Stelzer,M. and Schomburg,D. (2011) BKM-react, an integrated biochemical reaction database. *BMC Biochem.*, **12**, 42.
113. Kumar,A., Suthers,P.F. and Maranas,C.D. (2012) MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics*, **13**, 6.
114. Moretti,S., Martin,O., Van Du Tran,T., Bridge,A., Morgat,A. and Pagni,M. (2016) MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.*, **44**, D523–D526.
115. Reed,J.L., Patel,T.R., Chen,K.H., Joyce,A.R., Applebee,M.K., Herring,C.D., Bui,O.T., Knight,E.M., Fong,S.S. and Palsson,B.O. (2006) Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 17480–17484.
116. Satish Kumar,V., Dasika,M.S. and Maranas,C.D. (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, **8**, 212.
117. Rolfsson,O., Palsson,B.O. and Thiele,I. (2011) The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst. Biol.*, **5**, 155.
118. Rolfsson,O., Paglia,G., Magnusdottir,M., Palsson,B.O. and Thiele,I. (2013) Inferring the metabolism of human orphan metabolites from their metabolic network context affirms human gluconokinase activity. *Biochem. J.*, **449**, 427–435.
119. Breunig,J.S., Hackett,S.R., Rabinowitz,J.D. and Kruglyak,L. (2014) Genetic basis of metabolome variation in yeast. *PLoS Genet.*, **10**, e1004142.
120. Hill,C.B., Taylor,J.D., Edwards,J., Mather,D., Langridge,P., Bacic,A. and Roessner,U. (2015) Detection of QTL for metabolic and agronomic traits in wheat with adjustments for variation at genetic loci that affect plant phenology. *Plant Sci.*, **233**, 143–154.
121. Wen,W., Li,K., Alseekh,S., Omranian,N., Zhao,L., Zhou,Y., Xiao,Y., Jin,M., Yang,N., Liu,H. *et al.* (2015) Genetic determinants of the network of primary metabolism and their relationships to plant performance in a maize recombinant inbred line population. *Plant Cell*, **27**, 1839–1856.
122. Keurentjes,J.J., Fu,J., de Vos,C.H., Lommen,A., Hall,R.D., Bino,R.J., van der Plas,L.H., Jansen,R.C., Vreugdenhil,D. and Koornneef,M. (2006) The genetics of plant metabolism. *Nat. Genet.*, **38**, 842–849.
123. McMullen,M.D., Byrne,P.F., Snook,M.E., Wiseman,B.R., Lee,E.A., Widstrom,N.W. and Coe,E.H. (1998) Quantitative trait loci and metabolic pathways. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 1996–2000.
124. Hobbs,D.H., Flintham,J.E. and Hills,M.J. (2004) Genetic control of storage oil synthesis in seeds of Arabidopsis. *Plant Physiol.*, **136**, 3341–3349.
125. Steyer,D., Ambroset,C., Brion,C., Claudel,P., Delobel,P., Sanchez,I., Erny,C., Blondin,B., Karst,F. and Legras,J.L. (2012) QTL mapping of the production of wine aroma compounds by yeast. *BMC Genomics*, **13**, 573.
126. Sorokina,M., Stam,M., Medigue,C., Lespinet,O. and Vallenet,D. (2014) Profiling the orphan enzymes. *Biol. Direct.*, **9**, 10.
127. Van Schaftingen,E., Rzem,R., Marbaix,A., Collard,F., Veiga-da-Cunha,M. and Linster,C.L. (2013) Metabolite proofreading, a neglected aspect of intermediary metabolism. *J. Inherit. Metab. Dis.*, **36**, 427–434.
128. Linster,C.L., Van Schaftingen,E. and Hanson,A.D. (2013) Metabolite damage and its repair or pre-emption. *Nat. Chem. Biol.*, **9**, 72–80.
129. Tawfik,D.S. (2010) Messy biology and the origins of evolutionary innovations. *Nat. Chem. Biol.*, **6**, 692–696.
130. Khersonsky,O. and Tawfik,D.S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.*, **79**, 471–505.
131. D'Ari,R. and Casadesus,J. (1998) Underground metabolism. *Bioessays*, **20**, 181–186.
132. Beaudoin,G.A. and Hanson,A.D. (2016) A Guardian Angel Phosphatase for Mainline Carbon Metabolism. *Trends Biochem. Sci.*, **41**, 893–894.
133. Golubev,A.G. (1996) [The other side of metabolism]. *Biokhimiia*, **61**, 2018–2039.
134. Piedrafita,G., Keller,M.A. and Ralser,M. (2015) The impact of non-enzymatic reactions and enzyme promiscuity on cellular

metabolism during (oxidative) stress conditions. *Biomolecules*, **5**, 2101–2122.
135. Lerma-Ortiz,C., Jeffryes,J.G., Cooper,A.J., Niehaus,T.D., Thamm,A.M., Frelin,O., Aunins,T., Fiehn,O., de Crecy-Lagard,V., Henry,C.S. *et al.* (2016) 'Nothing of chemistry disappears in biology': the Top 30 damage-prone endogenous metabolites. *Biochem Soc. Trans.*, **44**, 961–971.
136. Galperin,M.Y., Moroz,O.V., Wilson,K.S. and Murzin,A.G. (2006) House cleaning, a part of good housekeeping. *Mol. Microbiol.*, **59**, 5–19.
137. Gladyshev,V.N. (2014) The free radical theory of aging is dead. Long live the damage theory! *Antioxid. Redox Signal.*, **20**, 727–731.
138. Van Schaftingen,E., Rzem,R. and Veiga-da-Cunha,M. (2009) L: -2-Hydroxyglutaric aciduria, a disorder of metabolite repair. *J. Inherit. Metab. Dis.*, **32**, 135–142.
139. Kremer,L.S., Danhauser,K., Herebian,D., Petkovic Ramadza,D., Piekutowska-Abramczuk,D., Seibt,A., Muller-Felber,W., Haack,T.B., Ploski,R., Lohmeier,K. *et al.* (2016) NAXE mutations disrupt the cellular NAD(P)HX repair system and cause a lethal neurometabolic disorder of early childhood. *Am. J. Hum. Genet.*, **99**, 894–902.
140. Losman,J.A. and Kaelin,W.G. Jr (2013) What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer. *Genes Dev.*, **27**, 836–852.
141. Wheeler,G.L., Jones,M.A. and Smirnoff,N. (1998) The biosynthetic pathway of vitamin C in higher plants. *Nature*, **393**, 365–369.
142. Linster,C.L. and Clarke,S.G. (2008) L-Ascorbate biosynthesis in higher plants: the role of VTC2. *Trends Plant Sci.*, **13**, 567–573.
143. Riemersma,M., Froese,D.S., van Tol,W., Engelke,U.F., Kopec,J., van Scherpenzeel,M., Ashikov,A., Krojer,T., von Delft,F., Tessari,M. *et al.* (2015) Human ISPD is a cytidyltransferase required for dystroglycan O-mannosylation. *Chem. Biol.*, **22**, 1643–1652.
144. Gerin,I., Ury,B., Breloy,I., Bouchet-Seraphin,C., Bolsee,J., Halbout,M., Graff,J., Vertommen,D., Muccioli,G.G., Seta,N. *et al.* (2016) ISPD produces CDP-ribitol used by FKTN and FKRP to transfer ribitol phosphate onto alpha-dystroglycan. *Nat. Commun.*, **7**, 11534.
145. Walsh,C.T., Garneau-Tsodikova,S. and Gatto,G.J. Jr (2005) Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Ed. Engl.*, **44**, 7342–7372.
146. Choudhary,C., Weinert,B.T., Nishida,Y., Verdin,E. and Mann,M. (2014) The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat. Rev. Mol. Cell Biol.*, **15**, 536–550.
147. Dai,L., Peng,C., Montellier,E., Lu,Z., Chen,Y., Ishii,H., Debernardi,A., Buchou,T., Rousseaux,S., Jin,F. *et al.* (2014) Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. *Nat. Chem. Biol.*, **10**, 365–370.
148. Walport,L.J., Hopkinson,R.J., Chowdhury,R., Schiller,R., Ge,W., Kawamura,A. and Schofield,C.J. (2016) Arginine demethylation is catalysed by a subset of JmjC histone lysine demethylases. *Nat. Commun.*, **7**, 11974.
149. Eletr,Z.M. and Wilkinson,K.D. (2014) Regulation of proteolysis by human deubiquitinating enzymes. *Biochim. Biophys. Acta*, **1843**, 114–128.
150. Nishikura,K. (2016) A-to-I editing of coding and non-coding RNAs by ADARs. *Nat. Rev. Mol. Cell Biol.*, **17**, 83–96.
151. Tomaselli,S., Locatelli,F. and Gallo,A. (2014) The RNA editing enzymes ADARs: mechanism of action and human disease. *Cell Tissue Res.*, **356**, 527–532.
152. Cantara,W.A., Crain,P.F., Rozenski,J., McCloskey,J.A., Harris,K.A., Zhang,X., Vendeix,F.A., Fabris,D. and Agris,P.F. (2011) The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.*, **39**, D195–D201.
153. Phizicky,E.M. and Hopper,A.K. (2010) tRNA biology charges to the front. *Genes Dev.*, **24**, 1832–1860.
154. El Yacoubi,B., Bailly,M. and de Crecy-Lagard,V. (2012) Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annu. Rev. Genet.*, **46**, 69–95.
155. Jaschke,A., Hofer,K., Nubel,G. and Frindert,J. (2016) Cap-like structures in bacterial RNA and epitranscriptomic modification. *Curr. Opin. Microbiol.*, **30**, 44–49.
156. Bird,J.G., Zhang,Y., Tian,Y., Panova,N., Barvik,I., Greene,L., Liu,M., Buckley,B., Krasny,L., Lee,J.K. *et al.* (2016) The mechanism of RNA 5' capping with NAD+, NADH and desphospho-CoA. *Nature*, **535**, 444–447.
157. Cahova,H., Winz,M.L., Hofer,K., Nubel,G. and Jaschke,A. (2015) NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs. *Nature*, **519**, 374–377.
158. Van Schaftingen,E., Veiga-da-Cunha,M. and Linster,C.L. (2015) Enzyme complexity in intermediary metabolism. *J. Inherit. Metab. Dis.*, **38**, 721–727.
159. Copley,S.D. (2012) Moonlighting is mainstream: paradigm adjustment required. *Bioessays*, **34**, 578–588.
160. Copley,S.D. (2014) An evolutionary perspective on protein moonlighting. *Biochem. Soc. Trans.*, **42**, 1684–1691.
161. Khan,I., Chen,Y., Dong,T., Hong,X., Takeuchi,R., Mori,H. and Kihara,D. (2014) Genome-scale identification and characterization of moonlighting proteins. *Biol. Direct.*, **9**, 30.
162. Espinosa-Cantu,A., Ascencio,D., Barona-Gomez,F. and DeLuna,A. (2015) Gene duplication and the evolution of moonlighting proteins. *Front. Genet.*, **6**, 227.
163. Seaver,S.M., Gerdes,S., Frelin,O., Lerma-Ortiz,C., Bradbury,L.M., Zallot,R., Hasnain,G., Niehaus,T.D., El Yacoubi,B., Pasternak,S. *et al.* (2014) High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 9645–9650.
164. Seaver,S.M., Henry,C.S. and Hanson,A.D. (2012) Frontiers in metabolic reconstruction and modeling of plant genomes. *J. Exp. Bot.*, **63**, 2247–2258.
165. Varrette,S., Bouvry,P., Cartiaux,H. and Georgatos,F. (2014) Management of an academic HPC cluster The UL experience. *IEEE*, 959–967.