This is a preprint of the following chapter: Arghya Mukherjee and Paul D. Cotter, Metagenomic Approaches For The Discovery of Pollutants Remediating Enzymes: Recent Trends and Challenges, published in Omics Insights in Environmental Bioremediation, edited by Vineet Kumar and Indu Shekhar Thakur, 2022, Springer, reproduced with permission of Springer. The final authenticated version is available online at: https://doi.org/10.1007/978-981-19-4320-1_24.

Metagenomic Approaches For The Discovery of Pollutants Remediating Enzymes: Recent Trends and Challenges

Arghya Mukherjee [1]* and Paul D. Cotter [1,2,3]

[1] Department of Food Biosciences, Teagasc Food Research Centre, Moorepark, Fermoy, Ireland

[2] APC Microbiome Ireland, University College Cork, Cork, Ireland

[3] VistaMilk, Cork, Ireland

*Corresponding author Emails: Arghya Mukherjee (arghya.mukherjee@teagasc.ie)

Abstract

Metagenomic studies in diverse environments have generated petabytes of sequencing data, allowing biologists to peer into the uncultivated microbial majority with unprecedented clarity. Such advances have heralded a new era of enzyme discovery where proteins of interest can be directly extracted from metagenomic sequencing data, although this remains a challenging task. Traditionally, metagenomic enzyme discovery, including those involved in xenobiotic degradation, have largely relied on functional insights derived from activity-guided or PCR-based metagenomics. Due to its untargeted and holistic nature, metagenomics allows us to probe the unknown and underexplored microbial diversity, which represents a key resource for novel biocatalyst discovery. Metagenomic shotgun sequencing-based enzyme discovery additionally avoids common biases introduced through the PCR-based or activity-guided functional genomics methods. In this chapter, we have provided an overview of metagenomics in novel enzyme discovery with discussions on both experimental and computational aspects of the same. We discuss in detail the computational strategies for identifying possible enzyme candidates from shotgun sequencing data and experimental strategies to characterize candidate enzymes once they have been identified. Finally, we review emerging methods for metagenomic enzyme discovery as well as future goals and challenges with an emphasis on metagenomics-based approaches.

Keywords: Metagenomics, xenobiotics, pollution, enzyme discovery, shotgun sequencing, protein function

1. Introduction

Increasing industrial, agricultural and domestic activities coupled with economic globalisation has led to large amounts of diverse pollutants being released in the environment. These include diverse pollutants such as pesticides, herbicides, dyes, petroleum hydrocarbons, and plastics, among others. Most pollutants are polymeric aromatic molecules with cyclic or planar units, which contribute to their stability and persistence in the environment. Indeed, the most frequently persistent and recalcitrant pollutants in natural habitats include polycyclic aromatic hydrocarbons (PAHs), steroids and dyes. Other persistent pollutants, which are not necessarily polymeric aromatic compounds, include long chain aliphatic hydrocarbons, organocyanides, organophosphates, polyurethanes, and pyrethroid herbicides and pesticides. Their recalcitrance to natural degradation, a direct result of their highly stable structures, contributes to the toxicity of these compounds in the environment and in animals. These compounds are often categorised as persistent organic pollutants (POP), with some of them being potent carcinogens, mutagens and/or capable of endocrine disruption (Ballschmite et al., 2002).

Restoration of polluted soil, lacustrine, marine and groundwater environments can be a major challenge. Different methods to clean up such contaminated habitats include physical and chemical methods such as pollutant adsorption, electrokinetic coagulation, ion exchange, electrochemical treatments, as well as biological techniques. Most physical or chemical methods for remediation of polluted sites have major drawbacks including the generation of undesirable by-products, economic infeasibility and high sludge production (Robinson et al., 2001). Biological remediation methods, in contrast, employ bacteria, fungi, algae or enzymes derived from such microbes for degradation and mineralization of pollutants in the environment to stable, innocuous end products (Cerniglia, 1997, Sutherland et al., 2004). Additionally, as bioremediation can attain complete removal of contaminants from the environment, such methods are considered more

effective and environment-friendly compared to more intrusive physical and/or chemical methods (Colleran, 1997). Bioremediation exploits the metabolic versatility of microorganisms to degrade a wide variety of organic pollutants in the environment as opposed to directly applied physical or chemical methods. Indeed, since most organic pollutants resemble certain naturally occurring compounds, most contaminated habitats would have certain endemic microorganisms that may be capable of metabolising these pollutants as their primary carbon source.

Microbial degradation of pollutants is mediated by specific microbial enzymes and are key components of bioremediation strategies. In most cases, such enzymes have broad specificity for a range of compounds having similar structures, thereby contributing to the metabolic plasticity of microbes in the environment. Upon detection and isolation, directed evolution-based strategies can be employed to improve the stability and catalytic efficiency of such enzymes (Theerachat et al., 2012). Most major advances in the recent years in such work has however been derived from 'omics'-based techniques, more specifically, next-generation sequencing (NGS)-based methods. Indeed, NGS-based methods have provided several insights into the catabolism of contaminants in the environment by bacteria and fungi through elucidation of relevant genes, operons, and metabolic pathways. This has helped us gain a better understanding of the genetic and molecular determinants involved in microbial degradative processes in the environment and in turn, aided the development of better bioremediation strategies. Most of such knowledge has been gathered from functional genomics studies, where functional characterization of microbial genomes after genome sequencing and/or activity-based screening of genomic libraries have led to identification of genes directly involved in the degradation of pollutants of interest (Pieper et al., 2004). However, given that a majority of microbial species in environmental microbiomes are not detectable through culture-dependent methods (Bakken, 1997, Pham and Kim, 2012), it can be reasonably speculated that most microorganisms capable of xenobiotic degradation remain uncharacterized.

Functional metagenomics, which endeavours to assign function to proteins in all genomes in a microbial community, in contrast to functional genomics that focuses on a single microbial genome, is a highly efficient method to access the uncharacterized microbiome. Since NGS-based functional metagenomic studies involve no isolation or cultivation steps, they represent a high throughput method for accelerated discovery of novel biocatalysts from the plethora of uncharacterized and uncultivated microbes in the environment. In this chapter, we discuss the various considerations, tools and strategies for functional metagenomics-based enzyme discovery, including those that involve metagenomic cloning and heterologous expression. Additionally, we discuss emerging techniques, methods and approaches in functional metagenomics with special emphasis on sequencing-based ones. Finally, we review in brief the upcoming goals and future challenges in the fast-changing discipline of enzyme discovery. We have endeavoured to focus on the aspects of functional metagenomics that represent the frontiers of the discipline and limit discussions of classical approaches, which are elaborated elsewhere in this book. Ultimately, we present a largely generalized approach to metagenomic enzyme discovery that can be applied in most investigations of the nature, with discussions specific to pollutant degrading enzymes wherever appropriate.

2. Metagenomics and functional characterization of microbiomes

The term metagenomics was first coined by Handelsman *et al* (Handelsman et al., 1998) and involves the study of environmental nucleic acids, specifically DNA. The first direct-sequencing-based study of microbial communities in the environment can however be attributed to phylogenetic studies carried out by Pace *et al* (Pace et al., 1985). Metagenomic studies not only include natural environments, but also investigation of microbial communities in human and animal guts (Almeida et al., 2019, Xia et al., 2017), built-up spaces (Danko et al., 2021) and indeed, the International Space Station (Mora et al., 2019), among others. While early metagenomic

studies followed a low throughput, cloning-based strategy, the advent of next-generation sequencing (NGS) technologies have transformed metagenomics into a high-throughput, holistic method to investigate microbiomes. Combined with continually decreasing sequencing costs, this has contributed to an exponential growth of sequencing repositories and has propelled metagenomics into becoming a scientific discipline in its own right. Indeed, sequencing repositories are estimated to accommodate approximately 10^{18} bp of nucleotide data within the next five years (Stevens et al., 2015).

Towards the end of the twentieth it became increasingly obvious that culture-dependent techniques were inadequate in recovering the majority of microbial diversity from environmental microbiomes. Most microorganisms in the environment do not survive in isolation and often exhibit a fastidious dependence on specific nutrients or metabolic fluxes to thrive. Indeed, only 1-5% of microbes in the highly studied rhizosphere could be recovered in the laboratory using culture-based methods (Bakken, 1997). The inability of culture-based methods in the laboratory to replicate such environmental conditions therefore meant that the majority of genetic information in environmental microbiomes remained elusive. Due to its culture-independent nature, metagenomics, more specifically NGS-based metagenomics, can provide researchers access to the genetic information of a greater proportion of the microbiome and therefore holds a major advantage over culture-based methods. Indeed, the 'rare biosphere', i.e. microbes occupying specific habitat niches and present in low abundance, can only be accessed through culture-independent, metagenomic techniques. The capability of metagenomic approaches to elucidate the functional characteristics of a largely uncultivated microbial majority has been demonstrated in the identification of metabolically-diverse lineages including candidate phyla radiation and DPANN archaea (Lomakina et al., 2018, Nicolas et al., 2020). Other examples include the elaboration of the biosynthetic potential of several uncultivated microbial phyla such as Eelbacter, *Candidatus* Tectomicrobia and Angelobacter, among others (Wilson et al., 2014, Crits-Christoph et

al., 2018). Indeed, heterologous expression of genes of interest from *Candidatus* Tectobacteria has allowed characterization of novel biosynthetic products and metabolic pathways (Wilson et al., 2014, Crits-Christoph et al., 2018, Mori et al., 2018). Although promising, the discovery of novel enzymes or products from metagenomes present unique challenges that can primarily be attributed to the uncultivable nature of microbes/DNA being through metagenomic approaches.

3. Targeted metagenomic sampling

Several microbial communities have been demonstrated to have xenobiotic degradation potential. However, the ability to degrade and mineralize xenobiotic compounds is not exclusive to microbes in contaminated ecosystems. Sufficient taxonomic and functional diversity along with the ability to metabolise compounds similar to the pollutant of interest as the primary carbon source can allow microbial communities from non-polluted environments to successfully negotiate contamination events. In recent years, several such natural habitats, which have never or rarely been exposed to pollution events, have been investigated for xenobiotic degrading enzymes. Indeed, such studies have revealed the presence of genes approximating to nearly 15% of the entire annotated human gut microbiome catalogue to be involved in the degradation and mineralization of xenobiotic compounds (Qin et al., 2010). However, due to the inherently contaminated nature of polluted environments, such ecosystems represent genetic sinks enriched in microbes able to degrade and thrive on xenobiotic compounds. Understandably, most metagenomic studies aimed at xenobiotic degrading enzyme discovery have been carried out in polluted habitats including contaminated soils and groundwater, activated sludge, and wastewater, among others. Several such studies have employed stable-isotope probing (SIP) to deconvolute the degradation of a pollutant of interest by microbial communities, where the contaminant has been radiolabelled. For example, to understand petroleum hydrocarbon degradation by microbial communities in contaminated groundwater, Wang et al (Wang et al., 2012) supplemented contaminated groundwater

microcosms with ^{13}C-labelled naphthalene. The heavy, radiolabelled DNA was subsequently separated from the unlabelled genetic material, sequenced and cloned to screen for desirable degradative properties. Such an enrichment-based approach is adequate for investigating active degraders of primarily simpler xenobiotic compounds. However, more complex pollutants, such as PAHs and complex aliphatics, often require the participation of multiple microbial species to accomplish complete degradation into simple products. Indeed, such approaches may overlook or miss microbes involved in the degradation of the most recalcitrant pollutant molecules and only detect microbes catalysing the metabolism of simpler products in the final steps of degradation.

4. Experimental designs for metagenomic enzyme discovery

In this section, we briefly discuss the various approaches, both *in silico* and experimental, in metagenomic enzyme discovery. While our primary emphasis will be on enzyme discovery from shotgun sequencing data, we will briefly outline both activity-guided and PCR-based functional metagenomics approaches (Figure 1). Comprehensive reviews on the latter are available elsewhere (Katz et al., 2016).

4.1. Activity-guided functional metagenomics

Functional metagenomic library screening through identification of desirable activities was one of the earliest methods developed in metagenomic enzyme discovery. In contrast to other methods, activity-based metagenomic methods are capable of simultaneously screening for novel enzymes as well as obtaining biochemical data. The approach relies on the observation of desirable phenotypes, in this case the degradation of pollutants of interest, by metagenomic clones. There are three main steps in such an approach: (i) cloning of environmental DNA or cDNA ranging between 2-200 kb into appropriate expression vectors such as fosmids, cosmids, or bacterial artificial chromosomes to create metagenomic or metatranscriptomic clone libraries; (ii) heterologous expression of the cloned fragments in an appropriate microbial host; and, (iii)

implementation of appropriately sensitive phenotypic screening strategies to identify 'hits', i.e., clones of interest exhibiting desirable properties. Metagenomic clones identified as 'hits' are then subjected to sequencing to reveal the genes or coding regions in the environmental DNA fragment and further investigate the protein of interest. Since such functional screening does not depend on any prior information, such as sequence homology with known genes, activity-guided metagenomics is particularly effective in identifying novel protein families, i.e. *de novo* enzyme discovery, as well as to assign new functions to previously identified protein families. Indeed, such screening has been widely used to identify additional properties of known enzymes, especially those considered commercially relevant, such as lipases, amylases, and cellulases, among others (Berini et al., 2017). Activity-based metagenomic studies have been instrumental in providing proof of function for xenobiotic degrading enzymes too. Two protein families involved in pollutant degradation and identified through activity-guided screening are oxidoreductases and hydrolases; these have been reviewed in detail elsewhere (Ufarté et al., 2015). Activity-based metagenomic screening, however, suffers from a number of disadvantages. Firstly, many enzymes require specific substrates or cofactors to function, and so may evade detection through general screening assays developed for primary metabolic enzymes. Additionally, incompatibility in codon usage, low levels of expression, and differential metabolic requirements, among others, in metagenomic library hosts may lead to detection of a lower number of 'hits'. Activity-based metagenomic studies have been more productive compared to sequence-based ones in providing evidence for enzyme function with respect to xenobiotic degrading enzymes. The primary two families of enzymes involved in pollutant degradation that have been unravelled through the use of activity-guided screening are oxidoreductases and hydrolases. *E. coli* was the most frequently used host for enzyme validation, although non-conventional hosts such as *Bacillus, Thermus, Pseudomonas* and *Sphingomonas*, among others, were also used in limited capacity (Ufarté et al., 2015). The transformation efficiency of E. coli for various plasmid types was indeed significantly higher than

other hosts, with the ability to appropriately express genes from taxonomically-distant species as well (Tasse et al., 2010).

4.2. PCR and microarray-based functional metagenomics

Microbiomes can be directly explored for known enzyme families using polymerase chain reaction (PCR)-based methods or through DNA/DNA and DNA/cDNA hybridisation methods. This can be achieved through the design of degenerate primers or probes from conserved sequences in the genes of interest. PCR-based screening strategies are highly specific and sensitive as such characteristics are inherent to PCR itself, with upscaling and of throughput possible through pooling and other deconvolution methods (Ferrer et al., 2016). PCR-based screening for pollutant degrading enzymes are fairly common and numerous examples of the same are available. For instance, ammonia monooxygenases, which are key enzymes involved in denitrification, were amplified and identified from soil, marine and laboratory microcosm samples contaminated with nitrates (Nogales et al., 2002, Treusch et al., 2005) using PCR-based methods. Similarly, PCR-based screening techniques were also employed to understand the diversity and abundance of 2,4-D/α-ketoglutarate and Fe^{2+}-dependent dioxygenases that are involved in the degradation of phenoxyalkanoic acid herbicides (Zaprasis et al., 2010). Interestingly, in most PCR-based screening studies, the emphasis is on understanding the genetic diversity of the protein of interest and the taxa contributing to the same in a given habitat. Consequently, reports of identification of full-length gene sequences and subsequent characterization of the enzyme products aimed at capture of novel catalysts are rare. One such example is the identification and validation of polychlorinated biphenyl degrading enzymes by Sul *et al* (Sul et al., 2009). In this study, microcosms of river sediments contaminated with polychlorinated biphenyl were amended with ^{13}C-biphenyl and the labelled DNA was later extracted and used to construct a cosmid library with 30-40 kb inserts. These inserts were subsequently screened using primers designed against aromatic dioxygenases

leading to the identification of a biphenyl dioxygenase encoding gene, which was subcloned, expressed and validated for its biphenyl degradation activity. Another such example is the identification of novel bacterial laccases through PCR-based screening of a marine metagenomic library and subsequent experimental validation of their activity against azo-dyes, catechol, syringaldazine, and 2,6-dimethoxyphenol, among others, by Fang *et al* (Fang et al., 2012). Probe-based metagenomic enzyme discoveries involve high density arrays designed to detect sequences specific to certain genes of interest and allows investigators to gain a detailed understanding of the diversity and abundance of such sequences in the habitat under study (Eyers et al., 2004). For instance, the GeoChip 4.0, which accommodates 83,992 50-mer oligonucleotide probes targeting 152,414 genes, was used by Lu *et al* to elucidate the differences in hydrocarbon degradation processes when comparing deep sea water samples and non-contaminated ones (Lu et al., 2012). This microarray-based method successfully highlighted the functional differences between the two microbial communities as well as quantified the hydrocarbonoclastic genes involved in aerobic degradation of petroleum hydrocarbons.

In the last few decades, metagenomic investigations have established that the actual enzyme diversity in the environment is much higher than that predicted from functional genomic data. To access the enzyme diversity unavailable using conventional primers, certain innovative strategies have been designed in recent years. For instance, Iwai *et al* have developed a hybrid approach where PCR-based screening is combined with next-generation sequencing of generated amplicons to iteratively construct new primers from other conserved regions in the gene of interest (Iwai et al., 2010). Such an approach allows the detection of structurally diverse and divergent proteins belonging the same functional group. Indeed, such approaches have allowed the identification of thousands of full-length dioxygenases from novel, unknown clusters of sequences recovered from metagenomes of polychlorinated biphenyl and 3-chlorobenzoate contaminated soils (Iwai et al., 2010, Morimoto and Fujii, 2009). In the latter study by Morimoto *et al* (Morimoto and Fujii, 2009),

genes that became dominant after addition of pollutants of interest were identified through PCR-denaturing gradient gel electrophoresis (DGGE) based methods. These dominant DNA sequences were subsequently used as anchor fragments to facilitate PCR-metagenome walking that led to the retrieval of full-length genes as well as elucidation of upstream and downstream genes/sequences.

The major drawback of PCR-based screening approaches is, understandably, its dependence on sequence homology, i.e. prior information on the gene or protein family of interest. This severely limits the usefulness of PCR-based methods in probing the unknown enzyme diversity. Additionally, being based on PCR, such methods have inherent amplification biases against GC-rich sequences and the less abundant members of microbiomes. Furthermore, short amplicons from genes do not provide reliable and/or complete information about the taxonomy of the source organisms or the co-occurrence information regarding other neighbouring genes. While this can be somewhat overcome through PCR-metagenome walking as mentioned above, the process is tedious and low-throughput. CONKAT-Seq, an innovative method recently developed by Libis *et al* (Libis et al., 2019) uses a co-occurrence network analysis-based approach to identify rare gene clusters. It employs position-barcoded domain amplification in an array of highly partitioned metagenome library, following which co-occurrence networks are computed for various protein domains and statistical analysis carried out to elucidate highly linked but rare gene clusters. Amplicon sequencing remains a favoured choice among researchers due to its low cost, but with decreasing sequencing prices, it is anticipated to be superseded by metagenomic shotgun sequencing in functional metagenomics studies aimed at enzyme discovery.

4.3. Metagenomic shotgun sequencing-based approaches

As mentioned above, metagenomic shotgun sequencing refers to the direct, untargeted sequencing of environmental DNA. Methods for library preparation in shotgun sequencing and

bioinformatic analysis have been discussed elsewhere (Ayling et al., 2019, Almeida and De Martinis, 2019). Metagenomic shotgun sequencing-based enzymatic discovery has several advantages over traditional functional metagenomics-based approaches. For example, shotgun sequencing-guided methods typically introduce less bias compared to functional metagenomics as PCR amplification and microbial hosts such as *E. coli* are not required. Additionally, shotgun sequencing is a high-throughput, highly scalable method that is less labour intensive compared to fosmid/cosmid library based functional metagenomics that take much longer to produce sequencing/gene level data. Furthermore, since sequencing data are now typically deposited in public repositories, analysis of these metagenomes can be carried out by different research groups around the world to provide further insights into genes and gene clusters of interest, their regulation, and other relevant aspects; this could never be possible with large-insert library based functional metagenomics. However, while shotgun sequencing is capable of providing information at an accelerated rate compared to functional metagenomics, downstream cloning and heterologous expression of the genes of interest are still necessary for biochemical validation of enzymes identified through both methods.

One of the major challenges in shotgun sequencing-guided enzyme discovery is the retrieval of environmental DNA Of sufficient quality and quantity from complex habitats/microbiomes and achievement of adequate sequencing depth that may allow error correction. This can be partly resolved through targeted enrichment methods in sequencing, which are be particularly effective in extracting longer reads from specific taxa or gene clusters of interest from low amounts of environmental DNA. For instance, recent advances from Samplix Technologies (Kvist et al., 2014) relies on microdroplet multiple displacement amplification of unknown sequences flanking short, desired sequences to enable indirect capture and sequence enrichment of the former. Other disadvantages in shotgun sequencing-based approaches include computational costs attached to such projects, limitations of bioinformatic approaches, as well as the inherent biases and

inaccuracies of metagenomic assembly, annotation, and binning. Shotgun sequencing projects frequently employ the Illumina sequencing platform, which produces short reads that can be an impediment to recovery of full-length genes/gene clusters, even with the significant advances made in metagenomic assembly in recent years. Short-read sequencing can however be complemented with other sequencing methods to improve metagenomic assemblies. For instance, techniques such as Hi-C chromosome capture for proximity-guided assembly of short reads have successfully demonstrated improvements in the genome-level resolution of human gut and cow rumen microbiomes (Yaffe and Relman, 2020, Stewart et al., 2018). Additionally, long-read sequencing technologies such as Oxford Nanopore (Jain et al., 2016) or PacBio HiFi (Hon et al., 2020) can be combined with short-read sequencing to dramatically improve metagenomic assemblies. This is particularly important in retrieval of full-length genes/operons of microbial enzymes involved in secondary metabolism which are often clustered together in microbial genomes.

In recent years, several studies have investigated heavily polluted sites using metagenomic methods to understand microbial community structures, functional characteristics, species interactions, and genetic determinants necessary for survival in such habitats. However, discovery of enzymes from direct shotgun sequencing as compared to functional metagenomics approaches remain rare. Indeed, a recent estimate has reported that only seven studies have identified novel enzymes using shotgun sequencing methods compared more 300 such instances using functional metagenomic methods (Berini et al., 2017). Metagenomic analysis of the activated biomass from an effluent plant carried out by Jadeja *et al* allowed the identification and retrieval of novel oxygenase sequences (Jadeja et al., 2014). Additionally, samples of the activated biomass were demonstrated to degrade a range of aromatic hydrocarbons, but the degradation activity was not attributed to specific taxa or enzymatic component. A study by Wang *et al* (Wang et al., 2012) discovered a key operon involved in naphthalene degradation by combining stable isotope probing

and shotgun sequencing. Shotgun sequencing of ^{13}C-labelled DNA was followed by cloning of the identified operon and experimental validation of naphthalene degradation. This remains one of the few examples where massively parallel sequencing has been successfully employed to identify and characterize new pollutant degrading enzymes. With increasing focus on shotgun sequencing studies, a considerable reservoir of metagenomes and annotated genes are now available that can be analysed *in silico* for identification of novel catalysts in bioremediation. Indeed, 1,200 bacterial laccases have been identified in genomic and metagenomic datasets, which remain to be experimentally validated (Ausec et al., 2011).

5. Computational methods in metagenomic enzyme discovery

5.1. Searching metagenomic databases

Due to the petabytes of sequencing data being deposited in public repositories, a wealth of information is available for researchers to mine and identify novel enzymes and predict new enzyme functions in protein families, including those involved in xenobiotic degradation. After sampling, DNA extraction and sequencing, shotgun sequencing-based metagenomic studies deposit their data in public repositories such as the Sequence Read Archive (SRA) (Leinonen et al., 2011), Joint Genome Institute Integrated Microbial Genomes and Microbiomes Resource (JGI IMG/M) (Chen et al., 2019) or MGnify (Mitchell et al., 2020), among others. MGnify is unique compared to other metagenome resources as it allows researchers to probe metagenomes using Hidden Markov Models (HMMs) compared to more commonly implemented sequence-alignment based methods such as BLAST (McGinnis and Madden, 2004) or DIAMOND (Buchfink et al., 2015). This is an important distinction as compared to alignment-based methods, HMM-guided methods are more effective in identification of remote homologs and as an extension, novel enzymes. Sets of aligned sequences are used to build probabilistic models, i.e. HMM profiles, which can sensitively detect distantly related enzymes at the boundaries of protein families. Custom HMM

profiles can be built for defined clades of closely related sequences to investigate metagenomes for protein subfamilies as desired by the user. A complementary approach to HMM-based searches involve the reconstruction of metagenome-assembled genomes (MAGs) from metagenomic shotgun sequencing data. Assembly and binning of metagenomic contigs into MAGs allow researchers to extract full-length genes at a genome level resolution, which in turn enables recovery of full-length genes/operons involved in secondary metabolism from the unknown microbial biodiversity. Indeed, a recent study on ~ 10,000 diverse environmental metagenomes by Nayfach *et al* (Nayfach et al., 2021) using MAG-based approaches led to the identification of > 100,000 biosynthetic gene clusters, thereby underlining both the challenges and opportunities in metagenomic enzyme discovery from metagenomic shotgun sequencing data. Given the exponential rise in shotgun sequencing data being generated, there is a distinct need for improved tools and platforms, HMM-based, MAG-based or otherwise, to facilitate and accelerate metagenomic novel enzyme discovery.

5.2. Phylogenetic methods

Phylogenetics, which aims to understand the links between shared functional traits including proteins with similar functions, can be a particularly useful tool in enzyme discovery. Most enzymes tend to group by their shared preference of substrates or similar functionality and not their taxonomic origin, unlike other genes such as the 16S rRNA. This makes phylogenetics a powerful tool in reference-based enzyme discovery, where sequences of genes of interest can be aligned with uncharacterized, metagenomic sequences to identify distantly related enzymes. Reference sequences for characterized enzymes of interest can be obtained from Swiss-Prot (Bairoch and Apweiler, 1996), the Protein Data Bank (Berman et al., 2000) or through literature search. In such an approach, reference sequences act as anchors or seeds around which the phylogenetic tree is expanded, with distantly related, uncharacterized sequences forming distinct

clades with no seed sequences. This provides easy, visual inspection of distantly related, possibly novel sequences and is often a good place to start investigations for enzymes preferring different substrates or with novel functionalities. A key drawback of phylogenetic trees is that their accuracy is dependent on the quality of the multiple sequence alignments they are built on. While there are several multiple sequence alignment tools available, such as MAFFT (Katoh et al., 2002), MUSCLE (Edgar, 2004) and Clustal Omega (Sievers and Higgins, 2018), manual inspection of the alignment to trim regions outside the core alignment as well as large gaps is essential to ensure high-quality trees downstream. Tools such as Gblocks (Castresana, 2000) and trimAl (Capella-Gutiérrez et al., 2009) are well suited to carry out such intermediate steps before treeing. An additional limitation of phylogenetic methods is the computational cost of creating trees from large multiple sequence alignments. While rigorous maximum-likelihood based treeing tools such as RaxML (Stamatakis, 2014) or PhyML (Guindon et al., 2005) make few assumptions and require significant investment in time and resources, newer algorithms such as FastTree (Price et al., 2010) and RaxML-NG (Kozlov et al., 2019) faster, less resource-hungry and nearly as accurate. FastTree employs heuristic methods to shrink the tree search space and make approximations of maximum-likelihood probabilities, in turn achieving drastically reduced treeing times. RaxML-NG, in contrast, allows computational scalability for large metagenomic datasets while retaining the improved accuracy of RaxML. IQ-Tree (Nguyen et al., 2015) is yet another popular treeing tool, and includes automated model selection as well as ultra-fast bootstrapping. The ETE3 toolkit (Huerta-Cepas et al., 2016) or ggtree (Yu et al., 2018) implemented in R can be used for advanced annotation and visualization of phylogenetic trees.

Ancestral state reconstruction (Hochberg and Thornton, 2017) is a related method that analyses contemporary protein sequences to infer their evolutionary history and may help identify evolutionarily distant relatives of reference genes of interest. FastML (Ashkenazy et al., 2012) is a web-based tool that allows ancestral state reconstruction to be explored by non-experts. Indeed,

ancestral state reconstruction of diterpene cyclases by Hendrikse *et al* and subsequent experimental validation of the predicted sequences revealed that ancestral forms of the enzymes had improved thermostability and broader substrate specificity (Hendrikse et al., 2018). Tools like EvoMining (Sélem-Mojica et al., 2019) and CORASON (Navarro-Muñoz et al., 2020) can be implemented in phylogeny-based genome mining to identify new enzymes. CORASON, or CORe Analysis of Syntenic Orthologs to prioritize natural products biosynthetic gene clusters, takes a query gene and genome database to identify gene clusters that share the same genomic core and reconstructs multi-locus phylogenies to explore evolutionary relationships. Although originally created to explore biosynthetic gene clusters, the tool can be used for any gene cluster. EvoMining is based on the assumption that primary metabolic enzymes often undergo duplication or horizontal gene transfer and both events can lead to the rise of secondary metabolic enzymes with novel functions. It has been implemented recently in the discovery of novel enzymes involved in the biosynthesis of arseno-organic molecules (Cruz-Morales et al., 2016). Overall, phylogenetics is a key first step in enzyme bioprospecting from metagenomes and can be highly effective when used in combination with some of the other relevant methods described in this section.

5.3. Genome context-based approaches

Gene neighbourhoods, i.e. genes flanking the gene of interest can often provide important information regarding the substrates, cofactors or partner proteins required for enzyme activity. Indeed, gene context is a critical but frequently overlooked aspect of metagenomic enzyme discovery. Gene neighbourhood analysis can be automated using the Genome Neighbourhood Tool (EFI-GNT) addition for the Enzyme Function Initiative Enzyme Similarity Tool (EFI-EST) (Gerlt et al., 2015, Zallot et al., 2019). The EFI-GNT provides statistical analysis on gene co-occurrence to facilitate identification of possible functional linkages as well as generates genome neighbourhood networks that allows visual inspection of genome context. Other tools such as BiG-SLICE (Kautsar

et al., 2021) and BiG-SCAPE (Navarro-Muñoz et al., 2020) are designed to identify and group biosynthetic gene clusters. The latter is integrated with CORASON thereby allowing simultaneous analysis using neighbourhood clustering as well as phylogenetics. BiG-SLICE dramatically reduces computational resource requirement and run time by clustering gene clusters in a Euclidean space rather than by pairwise comparison. Indeed, BiG-SLICE is well suited for genome-context guided metagenomic enzyme discovery due to its ability to cluster millions of gene clusters from metagenome-assembled genomes within a reasonable run-time. The STRING web resource is another platform that is helpful in identification of gene functions through investigations into gene context (Szklarczyk et al., 2019). STRING employs text mining of scientific literature and gene associations inferred from diverse data including co-expression data and gene orthology to model organisms to facilitate functional analysis of proteins including prediction of protein-protein interactions. Although STRING is not designed specifically for metagenomic investigations, it can nevertheless be used to confidently infer protein-protein interactions. CO-ED is a more specific tool, which can implement network analysis to identify co-occurring domains in multi-domain proteins (de Rond et al., 2020). CO-ED relies on PFAM information as input, which can be extracted from metagenomes using PfamScan (Madeira et al., 2019). Upon completion of analysis, CO-ED highlights the domains in the input data that are already present in public repositories (for e.g. UniPROT (The UniProt, 2019), MIBiG (Kautsar et al., 2020)), and other domains that have not yet been characterized. CO-ED therefore provides insights into unusual domains or combination of protein domains (in multi-domain proteins) in metagenomes and in turn facilitates identification of possible points of initiation for enzyme discovery investigations.

5.4. Sequence similarity networking

Sequence similarity networks (SSNs), are graphs that display relationships between protein families and were first published in 2009 for the purpose of investigations into protein

superfamilies (Atkinson et al., 2009). SSNs use scores from an all-against-all BLAST search for a user-defined custom dataset to generate network graphs where each protein sequence is represented by a node and pairwise sequence similarities define the edges between the nodes. Smaller clusters of protein families of interest can be revealed by pruning the network graph through manipulation of the sequence similarity threshold. Clusters of metagenomic sequences of interest can therefore be analysed relatively quickly to identify protein subfamilies that may yield novel biocatalysts. As mentioned above for phylogenetics, it is useful and indeed, common practice to include representative sequences as seed or anchor points, which in turn helps draw robust inferences for relationships between protein families and subfamilies. A major advantage of SSNs is the ability to visually inspect thousands of sequences and interpret the relationship between proteins. Additionally, SSNs can be computed much faster than bootstrapped maximum-likelihood phylogenetic trees and are computationally less intensive as well. Visual inspection of network graphs can be carried out interactively using network visualisation tools such as Cytoscape (Shannon et al., 2003). However, while the graphical user interface in Cytoscape makes network inspection more accessible to researchers, it may be difficult to reproduce such graphs/networks due to the point-and-click system. Python, R or C/C++ can be used to create reproducible SSN workflows for CyREST API or igraph (Csardi and Nepusz, 2006, Ono et al., 2015); users without coding or programming experience can alternatively use the EFI-EST platform for automated sequence similarity network construction online (Gerlt et al., 2015). A key drawback of SSN analysis are biases introduced during pruning of the network graph using sequence similarity threshold scores. BLAST e-values form an important parameter in such decisions, which can often be misleading since BLAST e-values are derivatives of database size that in turn make comparisons between SSNs infeasible as the size of the sequence databases may vary. Furthermore, a variety of graph layouts may lead to different interpretations of the same data. As a compromise, sequences, codes, network layouts, similarity scores and BLAST e-values can be made publicly

available, which will allow third-party examination of SSN workflows and limit bias introduced through preferences of the involved researchers.

5.5. Structure-based methods

Traditionally, the lack of structural information for most proteins has been a roadblock for inferring enzyme functions from metagenomes. Although in recent years efforts have been made to characterize more proteins at the structural level, these have focused disproportionately on culturable organisms. Indeed, only a small proportion of PDB entries originate from metagenomes (Robinson et al., 2001). Surprisingly, large-scale efforts to structurally characterize proteins, as mentioned above, have yielded much fewer novel protein folds than expected (Jaroszewski et al., 2009). While it is expected that the repertoire of protein folds will be expanded as we further interrogate the unknown microbial majority, it must be noted that given the enormous possibilities of amino acid combinations, only a small fraction of protein conformations are represented in biological macromolecules studied to date. Observations that even the most conserved protein folds can catalyse a variety of reactions underline our nascent understanding of the effects of the multi-dimensional enzyme structure on catalytic diversity. Nevertheless, protein structure can provide important insights into functionality beyond the primary amino acid sequence. Powerful structural prediction tools such as AlphaFold2 (Jumper et al., 2021) can significantly advance our understanding of uncharacterized proteins. However, AlphaFold2 is currently not available publicly and web-based homology modelling tools such as I-TASSER (Roy et al., 2010), Phyre2 (Kelley et al., 2015) and SWISS-MODEL (Schwede et al., 2003) can be used alternatively to predict protein structure. Importantly, structural prediction is often the first step towards detection of active sites as well structurally vital motifs, both of which play critical roles in determining enzymatic/protein function.

Enzyme active sites represent a minute proportion of the full-length, folded protein volume in three-dimensional space. Catalytic residues in enzyme active sites as well as the architecture of the site itself are highly conserved compared to the rest of the enzyme. However, the same catalytic residues can accommodate a wide range of substrates as well as catalyse a diversity of reactions, most famously in case of the Ser-His-Asp catalytic triad (Rauwerdink and Kazlauskas, 2015), thereby indicating the multifunctional capabilities of even highly conserved residues in the enzyme active site. The prediction of enzyme function from examination of the active site alone can therefore be challenging, if not impossible. Alteration of a single residue in the active site can drastically change the substrate specificity or enantioselectivity of an enzyme; protein engineering for enzyme discovery can therefore lead to a lot of dead-ends (May et al., 2000). In this respect, identification of naturally occurring active site variants across genomes/metagenomes can provide an alternative and complementary route in enzyme discovery besides protein engineering. Studies on active site variants targeted at enzyme discovery from metagenomic shotgun sequencing are rare, with unusual active sites only reported upon enzyme characterization. Tools such as CASTp are useful for the *de novo*, automated detection of active sites/active site variants (Tian et al., 2018). Databases such as the Mechanism and Catalytic Site Atlas (M-CSA), which catalogues active site architecture and chemistry, can also be useful in this regard (Ribeiro et al., 2018). Although M-CSA currently contains more than 1000 curated entries representing approximately 73k Swiss-Prot entries and 15k PDB entries, it still represents less than 10% of proteins with solved structures (Robinson et al., 2001). Additionally, predicted active site information is also available in UniProt, which can be useful in identifying variant active sites in structurall similar metagenomic sequences. Other conserved motifs beyond active sites, such as cofactor binding sites that can be detected using tools such as ScanProSite (de Castro et al., 2006), are also generally conserved and can be useful in enzyme discovery. Overall, strategies such as motif searching combined with sequence homology modelling or *de novo* identification of conserved/variant structural features

and cofactor binding sites, can be particularly useful in structure-based enzyme discovery from metagenomic sequences.

5.6. Machine-learning based methods

Machine learning (ML) has made great strides in the past few decades with the continuous advancement of computing power and algorithms, particularly in the life sciences and chemistry. Indeed, more than 35 machine-learning-based methods for protein prediction have been published in the past decade; these approaches provide researchers with the opportunity to apply methods beyond homology modelling to understand unknown links between protein sequence, structure and function. One of the key drawbacks of machine-learning-based methods is their requirement for large amounts data to train. The quality and quantity of data directly influence the accuracy and sensitivity of the resulting machine-learning model, with more complex models requiring greater amounts of training data. Machine-learning models, particularly deep neural networks, frequently suffer from overfitting, where such models become highly specific for the given dataset and cannot be implemented in a generalized manner for similar studies or datasets. Availability of high-quality, curated data in public databases such as MiBIG (Kautsar et al., 2020) and Swiss-Prot (Bairoch and Apweiler, 1996) is therefore essential for adequate training of machine-learning algorithms and in enabling enzyme discovery. Transfer learning is an emerging method that is designed to handle the scarcity of experimentally verified data, where machine-learning models are pre-trained on large quantities of unlabelled data such as metagenomic sequences in order to learn the general characteristics of the data and in turn, improve performance on separate, related tasks such as protein function prediction. Further advances in transfer-learning methods as well as semi-supervised machine-learning will allow researchers to investigate large metagenomic datasets with machine-learning models trained on relatively smaller amounts of labelled data.

Machine-learning-based methods can incorporate a wide variety of features to produce better models. These include protein sequences, structural information, protein-protein interaction data and physicochemical properties for amino acids, among others (Robinson et al., 2001). Certain ML methods have also employed text mining to extract feature information available in text format, particularly in journal articles (You et al., 2018). Autoencoders, which are artificial neural networks that can automate feature extraction, thus removing human biases, have been recently employed in unsupervised encoding of protein features 181. Although such automation is attractive, Autoencoders require even larger datasets and longer compute times, which in turn may limit their usability. A diverse array of machine learning models including simple logistic regression, random forest and multi-layer neural networks, among others, have found use in protein function prediction (Bonetta and Valentino, 2020). However, benchmarking on different machine-learning models can be difficult due to the inconsistent classification systems used for prediction of protein functionality. Most machine-learning models use hierarchically structured protein classification systems such as the EC classification (Webb et al., 1992), Gene Ontology (GO) (Ashburner et al., 2000) or Functional Catalogue (Ruepp et al., 2004), among others, as objectives. Understandably, this creates difficulties in comparing various protein function prediction models, but recent initiatives such as the Critical Assessment of Functional Annotation Challenge have made important efforts to standardize the diverse approaches undertaken (Zhou et al., 2019). Importantly, deep neural networks have been outperformed by simple homology modelling and logistic regression models in certain instances of protein function prediction, and underlines the nascent and developing nature of ML in this field (Bonetta and Valentino, 2020). A primary drawback of currently available ML methods is their limitation of predicting truly novel protein functions. While ML models can be trained on a variety of objectives (i.e., GO or EC), they are unable to predict entirely new enzyme classes. Negative selection algorithms can be employed as an alternative to multi-objective training, where enzymes can be classified based on their ability to

perform a particular function or not (Youngs et al., 2014). In this method, proteins are not forced into previously defined enzyme classes, but can be predicted to not fit into any known classes, or fit into multiple classes, indicating potential substrate promiscuity (Copley, 2003).

6. Characterization of new enzymes

6.1. Primary selection and cloning

Until this point in the chapter, we have discussed various computational methods for prediction of enzyme function and identification of potential candidate enzymes. While such *in silico* techniques might be useful for prioritization of newer areas of the protein space to be explored, they remain predictions at best, and require functional validation through rigorous experimentation. In the following sections we will review and discuss some of the newer methods and approaches to successfully navigate functional validation of novel enzymes, primarily selection, cloning, expression and screening. We will not discuss the classical methods for heterologous expression and screening here; they have been reviewed in detail previously (Katz et al., 2016). An important initial step when selecting proteins for characterization in the lab is to ensure removal of truncated or chimeric sequences that in turn may give rise to dysfunctional proteins. Additional clues for mispredicted start or stop codons may be obtained from visual inspection of multiple sequence alignments, where erroneous sequences will show up as outliers in terms of sequence length. Several tools are currently available to facilitate selection of candidate sequences from hundreds of thousands of metagenomic sequences for enzyme discovery. Clustering tools such as CD-HIT, UCLUST and Linclust (available through the MMSeq2 software suite) (Li and Godzik, 2006, Edgar, 2010, Steinegger and Söding, 2017) can cluster metagenomic sequences by similarity to give automatically selected representative sequences. Representative sequences for protein subfamilies of interest can be subsequently selected through SSN analysis and visualisation in Cytoscape (Shannon et al., 2003) or igraph (Csardi and Nepusz, 2006). The strategy here,

irrespective of the clustering and selection pipeline followed, is that highly similar proteins will almost always perform similar functions. While exceptions to this rule have been reported (North et al., 2020), clustering remains a useful tool in the initial stages for selection of representative candidate sequences from metagenomes.

Further filtering of representative sequences may be necessary depending on the dataset size. In many cases, the absence of high-throughput enzyme activity assays can preclude further investigation of several metagenomic sequences. Most frequently however, metagenomic sequences found in culturable organisms are readily selected since functional characterization in this case would be possible in the native host. Yet another popular strategy is to select proteins from thermophilic organisms that tend to have increased thermostability. It must be noted that this is also a generalization, since proteomes of diverse organisms across the tree of life have exhibited highly variable melting temperatures, even for organisms adapter to extreme temperatures (Jarzab et al., 2020). Alternatively, proteins with higher GC content, transmembrane regions or extended disordered regions may be filtered out to obtain candidate proteins more likely to be stable. Various tools such as XANNPred (Overton et al., 2011), CrystalP2 (Kurgan et al., 2009), XtalPred (Slabinski et al., 2007)and ParCrys (Overton et al., 2008), among others, have been developed to predict crystallization proclivities and can be used to predict protein stability. Multiple tools can be used to assess protein stability and results aggregated and ranked in order to prioritize representative sequences for experimental validation. Such an approach ensures that biases and weaknesses in individual tools are offset through an ensemble-based analytical approach, and in turn enables robust identification of the most promising candidates for enzyme discovery. Removal of signal peptides, i.e. 16-30 amino acids at the end of the N-terminus of many prokaryotic and eukaryotic proteins, is an approach that increases the likelihood of obtaining soluble proteins. Signal proteins are short sequences that direct the export of proteins from the cytosol and can therefore influence the solubility of proteins during heterologous expression

experiments. SignalP (Teufel et al., 2022) has been the most commonly used tool for detection and removal of N-terminal signal peptides in recent years, but more advanced, machine learning based tools for signal peptide detection are now available (Wu et al., 2021). With better understanding of the links between signal peptides and protein function, these can be used as attributes to filter representative sequences in enzyme discovery.

6.2. Heterologous expression

Once genes or enzymes of interest have been identified, constructs for heterologous expression of the same must be created. Most vectors used in construction of metagenomic libraries in functional metagenomics, however, are typically ill-suited for heterologous expression. Additionally, certain secondary metabolism gene clusters, which can be quite extensive in length, may be captured in a truncated form in fosmid or cosmid vectors that have maximum insert sizes of approximately 4.5 kb. New methods, besides classical restriction-based cloning and Gibson assembly methods, have been developed to facilitate efficient cloning of genes/gene clusters of interest in heterologous hosts (Zhang et al., 2019, Bordat et al., 2015). Among these new techniques, transformation-associated cloning relies on natural homologous recombination in yeast to aggregate overlapping cosmid/fosmid clones, while genetic recombineering employs diverse bacteriophage proteins to enable homologous recombination in *E. coli* (Zhang et al., 1998). The latter also allows rapid and efficient cloning of large gene clusters using RecET direct cloning in combination with Redαβ recombination (Wang et al., 2016). The availability of sufficient genetic material for cloning can be an issue in heterologous expression and PCR amplification remains one of the most reliable methods to produce the same if the original environmental sample or the culturable organism is available. In the event of an absence of the genetic material for amplification, gene synthesis can be carried out for the enzyme/gene of interest. Advantages of gene synthesis include codon optimization for the heterologous host of choice, which can be

particularly useful for the expression of metagenomic sequences in non-conventional, distantly related hosts. Heterologous expression of metagenomic sequences often fail even when constructs are prepared properly. One of the reasons behind this may be that some enzymes exhibit activity only when co-expressed with surrounding genes, i.e. the entire gene cluster. Additionally, other factors such as suitable coenzymes, cofactors, or post-translational modification machineries may be absent in the host leading to a failure of heterologous expression or production of dysfunctional proteins.

Non-conventional hosts can be used as an alternative to conventional hosts such as *E. coli* for heterologous expression when the latter fails. To do this, homologous sequences for metagenomic sequences of interest can be mined from genomes and such sequences in turn can be cloned and expressed in *E. coli* or other conventional hosts. In this manner, researchers can obtain information on the sequence/gene/domains of interest by studying similar proteins from related organisms that can be expressed in conventional hosts. Examples include the investigation of FR900359, a Gq protein inhibitor identified from a leaf metagenome but later characterized using a homologous protein from *Chromobacterium vaccinia* expressed in *E. coli* (Hermes et al., 2021). Yet another strategy for heterologous expression can be selection of the closest culturable taxonomic relatives if genetic engineering tools are available for the same. Indeed, such an approach has been frequently implemented in expression of various biosynthetic gene clusters from *Streptomyces* spp. in the model host *Streptomyces coelicolor* A3(2) (Zhang et al., 2016). Protein expression in *E. coli* for sequences from evolutionarily distant organisms, however, can be successful in certain cases. Overall, the choice of a host for heterologous expression of metagenomic sequences of interest are still based on a process trial-and-error. Design-build-test-learn pipelines in synthetic biology may be able to produce optimized hosts for heterologous expression in the future, thereby reducing the trial-and-error process (Carbonell et al., 2018).

6.3. Enzyme activity screening

The final challenge in characterization of enzymes is carrying out assays, *in vitro* or *in vivo*, to assess enzyme activity. Generally, there is a compromise required at this stage between throughput and specificity for enzyme screening methods. Typically, activity-screening assays involve visual confirmation for zones of inhibition around cultured bacterial colonies or production of colour or fluorescence due to enzymatic action on chromophores or fluorophores respectively. However, suitable substrates analogues may not be available for a numerous, especially novel, uncharacterized enzymes; the presence of large chromophore or fluorophores may also interfere with the activity of candidate enzymes. Mass spectrometry (MS) can be used as an alternative here as it is sensitive and allows generalized application without the need for substrate analogues. Although this allows accurate monitoring of enzyme activity, MS suffers from low-throughput due to the requirement for chromatographic separation. Alternative workflows are now available that bypass the chromatographic step to allow MS-based enzyme activity screening with higher throughput. For instance, nanostructure-initiator mass spectrometry (NIMS) employs a nanostructured surface to trap compounds that can be subsequently released by laser irradiation for MS (Northen et al., 2008). A washing step promotes non-covalent fluorous interactions on said surface that in turn enables compound separation. An improvement on NIMS, PECAN (Probing Enzymes with Click-assisted NIMS) relies on click chemistry, thereby allowing screening of enzymes which are not able to accommodate bulky perfluoroalkylated tails in their active site. As it uses click chemistry, substrate analogues with 'clickable' alkynes or azides are still required in PECAN, but being much less bulkier than the perfluoroalkylated tails used in NIMS, a wider range of enzyme classes can be assayed by this method (de Rond et al., 2019). Another alternative to NIMS is SAMDI-MS (Self-assembled Monolayers for matrix-assisted desorption/ionization-MS), where proteins or metabolites are immobilized on self-assembled monolayers of alkanethiolates on gold. This allows significantly higher throughput compared to MS-based method while conserving the

specificity and generalizability of the latter (Mrksich, 2008). Additionally, SAMDI-MS does not require labelling, which makes it ideal for enzymes without established screening assays. Although these methods have only been employed in a limited capacity for metagenomic enzyme discovery, they are well suited for the same, with specialised equipment and expertise needed for these methods being the primary barriers to their successful implementation in the discipline. Besides mass spectrometry-based methods, biosensors may also be used in metagenomic enzyme discovery. Indeed, biosensors have been employed in identifying a novel enzyme capable of cyclizing ω-amino fatty acids from a marine sediment metagenomic library (Yeom et al., 2018). Methods such as mRNA display has been recently used to screen a naturally occurring RiPP maturase (Fleming et al., 2020). It must be noted that in contrast to MS-based methods, biosensor and mRNA display are generally tailored to enzyme classes and therefore are high-throughput but less generalizable. Overall, several emerging techniques have been discussed in this section that have been used in a variety of applications and can be adapted and implemented in metagenomic enzyme discovery from diverse environments.

7. Metagenomic enzyme discovery: Outlook and emerging techniques

In this section, we discuss upcoming methods and techniques that can be used in association with metagenomics to accelerate enzyme discovery. These include meta-omic methods such as metabolomics and metatranscriptomics as well as single-cell genomics, cell-free techniques and sequence independent methods, among others.

7.1. Cell-free methods

Cell-free systems such as filtered lysate from *E. coli* or other expression hosts of choice can be a viable alternative to heterologous expression and purification of proteins of interest. All cellular organelles required for successful expression remains in the lysate, and additional components such as amino acids, cofactors and DNA can be exogenously added to induce expression of

proteins or metabolic pathways of interest (Bogart et al., 2021). In contrast to whole-cell hosts, cell-free platforms allow rapid transcription and translation of sequences of interest unbound to cellular growth. Additionally, cell-free platforms are also more tolerant to toxic metabolite production that may kill whole-cell hosts in heterologous expression. Throughput in such cell-free systems can be further increased using integrated screening methods such as in-droplet reaction microfluidics and mRNA display, among others (Bogart et al., 2021). Low yield however remains a frequently encountered issue for cell-free platforms, particularly when involving exogenous DNA distantly related to *E. coli*. Other commonly encountered issues include rapid degradation of mRNA templates and other critical cellular machinery. Despite these issues, cell-free systems are rapidly rising in popularity and it is anticipated that they will be widely implemented in metagenomic enzyme discovery in future.

7.2. Meta-omics approaches

While we have discussed enzyme discovery in this chapter primarily in terms of metagenomic shotgun sequencing, other meta-omic applications, both sequencing-dependent and independent, can make important, complementary contribution to enzyme. Indeed, integration of omic-methods such as transcriptomics/metatranscriptomics, metabolomics and metaproteomics, among others, can provide a powerful framework to link genetic information with phenotypic data, thereby facilitating hypothesis generation. For instance, transcriptomic studies provides insights into the transcription profile of an organism of interest at any given time and can elucidate differentially expressed genes (of both known and unknown functions) under conditions of interest. Instances of RNA-Seq involvement in enzyme discovery and expansion include the identification of biosynthetic pathways for domoic acid and expansion of the enzymatic repertoire of DadH related enzymes in the gut (Rekdal et al., 2020, Chekan et al., 2020). Metatranscriptomics-based approaches can similarly be used for enzyme discovery from microbiomes. Although

metatranscriptomics-based studies in enzyme discovery are still sparse, ease of sequencing and increasingly accessible bioinformatic tools are driving a steady increase in these numbers. For instance, BiG-MAP is a recently released tool that enables differential expression analysis of biosynthetic gene clusters in transcriptomic/metatranscriptomic datasets (Pascal Andreu et al., 2020). While similar tools are not yet available specifically for pollutant degrading enzymes, most of the tools described in the chapter, frequently developed for biosynthetic gene clusters, can be used for xenobiotic-degrading enzyme discovery. Integration of metatranscriptomic and increasingly available metabolomics data, along with analytical tools such as BiG-MAP, will contribute to an accelerated metagenomic mining for relevant enzymes of interest. Metaproteomics, compared to other omics-methods, remains an under-utilized technique. Recently, a functional metaproteomics workflow has been proposed by Sukul *et al*, where proteins are directly isolated from soil samples, separated using 2D-polyacrylamide gel electrophoresis, and refolded to assay, screen and identify novel lipolytic enzymes using a fluorogenic lipase substrate (Sukul et al., 2017). Proteins exhibiting desired activities are then subjected to mass-spectrometric analysis. Simultaneous metagenomic shotgun sequencing of the same soil sample allowed searching of the lipolytic enzymes identified using mass-spectrometry against a custom environmental database and recover full-length sequences and taxonomic origins for proteins of interest (Sukul et al., 2017). This study employed in-gel assays, which may be challenging to execute for screening enzymes without well-established colorimetric or fluorometric substrates. Thus, low-yields of proteins from environmental samples, refolding of proteins in-gel and limitations regarding available assays, may be some issues that need to be resolved before metaproteomics methods can be better integrated into meta-omic workflows. Nevertheless, integration of meta-omic methods, as discussed, provide a promising avenue towards metagenomic enzyme discovery.

7.3. Single-cell genomics

Increasingly popular, single-cell genomics can be implemented as both an alternative to shotgun sequencing and/or as a complimentary application. In single–cell genomics, microbial cells first sorted, usually using FACS or microfluidics-based method, with subsequent lysis and whole genome multiple displacement amplification using high-fidelity polymerases (Woyke et al., 2017). One of the major drawbacks of single-cell genomics is the low-yield of DNA material from a single cell even after a billion-fold amplification of single-cell genetic material, which in turn results in poor quality of genomes. Additionally, minute amounts of contaminating DNA can result in contributing a significant proportion of the DNA material due to the initial amplification step and in turn destabilise the entire pipeline. Recent advances have contributed to the development of optimized protocols implementing high-fidelity polymerases to improve such issues along with amplification bias from GC-rich genetic material commonly encountered in secondary metabolism related gene clusters from Actinobacteria and other organisms with high GC content (Stepanauskas et al., 2017). A clear advantage of single-cell genomics over shotgun sequencing-based metagenomics is the ability of the former to directly link taxonomic information with genomic and phenotypic data without the requirement for binning and annotation. Ideally, best results would be obtained through simultaneous examination of the samples of interest using both single-cell genomics and shotgun sequencing, where limitations of either can be offset to maximise result quality. Reference genomes from cultivated isolates have been found to rarely overlap with MAGs or single-amplified genomes (SAGs; genomes amplified using single-cell genomics methods) and indicates that diverse sequencing application may contribute to greater genome-level resolution of microbial communities (Paoli et al., 2021). Importantly, since single-cell genomics methods sequence every cell individually, it has contributed to a better understanding of the within-population genomic variability and evolution. Notable here is that gene clusters involved in secondary metabolism are frequently strain-specific and single-cell sequencing therefore provides a high-throughput, reliable and sensitive method to screen for the same in

microbiomes. Single-cell RNA sequencing has been used to show spatial transcription heterogeneity in microbial populations that are genetically identical (Imdahl et al., 2020). Since spatial heterogeneity has been reported in secondary metabolite production, single-cell transcriptomics may be a useful method to probe microbial communities for unknown genes with functions of interest (Tobias and Bode, 2019). Overall, single-cell genomics remains a nascent field of research with wide applications, including metagenomic enzyme discovery.

7.4. Sequence-independent methods

Most techniques to infer protein function described in this chapter have either been sequence-based on structural homology-based. However, such approaches may be insufficient in identifying and characterizing enzymes that do not have any structural or sequence similarity with known protein families. Although limited to certain protein families, few tools are available now to investigate and aid such *de novo* enzyme discovery. decRiPPter (Data-driven Exploratory Class-independent RiPP TrackER) has been recently developed to identify new RiPP classes without consulting homology to known RiPP classes or enzymes (Ziemert et al., 2016). Briefly, decRiPPter employs a filtering step where it implements pan-genomic analyses to infer sparsely distributed operons within taxonomic clades, thereby identifying genes/gene clusters that may be involved in secondary metabolism. While this approach has been used previously to identify a new family of RiPP maturases, a major drawback of the process remains the increased possibility of false positives compared to homology-based methods when searching for novelty. Due to its limited implementation in current research, sequence- and structure-independent methods remain an open and emerging field for metagenomic enzyme discovery.

8. Conclusion and future challenges

Several challenges remain before the rate of enzyme discovery, particularly from the unexplored microbial space, can see significant acceleration. Firstly, studies on metagenomic enzyme

discoveries are heavily skewed towards prokaryotes and eukaryotes in the environment are rarely investigated for enzymatic properties. Eukaryotes, however, play an important role in several ecosystems with some cultivable fungi being frequently used in bioremediation. Moving forward, eukaryotic populations in the environment need to be screened more diligently, with possible expression in new screening hosts, to facilitate discovery of novel biocatalysts from the unknown eukaryotic biodiversity. Similar bias can be noticed in the existing literature for studies involving esterases and oxidases with diverse substrate specificities. While such enzyme discovery studies are common, investigations into proteases, which are likely to be effective in hydrolysis of amide bonds in pollutants, are sparse. Additionally, uncharacterized protein families enriched in polluted habitats are seldom further investigated. Future efforts in metagenomic enzyme discovery may require certain changes in established approaches. For example, a recent meta-analysis revealed that among metagenomic enzymes discovered between January 2014 and March 2017, > 84% were esterases or cellulases and > 82% were discovered through activity-guided screening (Berini et al., 2017). This indicates a bias towards industrially relevant enzymes that can be detected using colorimetric assays. Additionally, hosts besides *E. coli* need to be considered for heterologous expression of candidate enzymes/genes in metagenomic enzyme discovery studies. Indeed, a controlled study reported the successful expression of ~30-40% of environmental bacterial genes and only 7% of high GC-content DNA the *E. coli* model system (Gabor et al., 2004). Given that many organisms classified as high-GC content are important producers of secondary metabolites and represent interesting reservoirs of novel enzymes, alternative and non-conventional hosts should be assessed for heterologous expression. Additional complications in leading to failed heterologous expression in *E. coli* hosts may involve a lack of cofactors, coenzymes, post-translational modification systems, and protein folding factors, among others. Proteins with reduced sequence identity with reference proteins with known functions are more likely to catalyse novel reactions and utilise a broader range of substrates, thereby making them

interesting targets for enzyme discovery. In order to detect more distantly related sequences, iterative HMM based strategies or PSI-BLAST may be used. For the former, an initial HMM model can be supplemented with BLAST searches to identify distant neighbours of the gene, the model rebuilt with the newly identified sequences and the search carried out iteratively until no new homologs are identified. Ultimately, enzymes/genes of interest need to be studied and understood as clusters of genes they are generally present as and not in isolation. Since bacterial genes involved in secondary metabolism are highly linked, identification of coexpression modules from metatranscriptomic datasets can help predict function of unknown genes through coexpression with genes with known functions. It can be anticipated that advances in long-read sequencing methods will allow examination of genetic loci directly from environmental DNA compared to requirements for assembly and binning for short-read sequencing datasets. Overall, the milieu of genetic clusters, transcriptomes and protein interactomes represent underexplored routes in enzyme discovery.

The emergence of meta-omic methodologies in the last decades has enabled researchers to access the functional potential of the hitherto unknown microbial populations in environmental microbiomes. Indeed, activity-guided screening of metagenomes have led to the discovery of numerous enzymes, primarily bacterial in origin, frequently from highly contaminated environments or habitats with other extreme characteristics. Notably, such *de novo* enzyme discovery using shotgun sequencing-based metagenomics is still rare. Thus, even with the continually accelerating accrual of meta-omics data in public repositories, the ease of next-generation sequencing has not necessarily translated to increased functional characterization of novel protein families. Functional metagenomics approaches are currently limited in several ways. These include the logistical impossibility of screening hundreds of thousands of metagenomic clones for desired properties, the requirement to design novel screening assays for new enzymatic reactions, and the need to handle hazardous, toxic materials when genes involved in xenobiotic

degradation are sought, among others. Some of these limits are being overcome by the development of multi-step activity-based screening strategies as well as the use of non-toxic structural analogues of the pollutants of interest in primary screening. Although continual advances in bioinformatic algorithms and next-gen sequencing methodologies show great promise in providing new avenues enzyme discovery, functional metagenomics techniques will remain important pillars for *de novo* enzyme identification and expansion. Importantly, we anticipate shotgun-sequencing based methods in combination with mass spectrometry-based molecular networking methods will facilitate rapid de-replication of candidate enzymes/genes and limit rediscovery, thereby providing proper direction and indirectly expediting truly novel enzyme discovery.

9. References

ALMEIDA, A., MITCHELL, A. L., BOLAND, M., FORSTER, S. C., GLOOR, G. B., TARKOWSKA, A., LAWLEY, T. D. & FINN, R. D. 2019. A new genomic blueprint of the human gut microbiota. *Nature,* 568**,** 499-504.

ALMEIDA, O. G. G. & DE MARTINIS, E. C. P. 2019. Bioinformatics tools to assess metagenomic data for applied microbiology. *Appl Microbiol Biotechnol,* 103**,** 69-82.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

ASHKENAZY, H., PENN, O., DORON-FAIGENBOIM, A., COHEN, O., CANNAROZZI, G., ZOMER, O. & PUPKO, T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research,* 40**,** W580-W584.

ATKINSON, H. J., MORRIS, J. H., FERRIN, T. E. & BABBITT, P. C. 2009. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLOS ONE,* 4**,** e4345.

AUSEC, L., ZAKRZEWSKI, M., GOESMANN, A., SCHLÜTER, A. & MANDIC-MULEC, I. 2011. Bioinformatic Analysis Reveals High Diversity of Bacterial Genes for Laccase-Like Enzymes. *PLOS ONE,* 6**,** e25724.

AYLING, M., CLARK, M. D. & LEGGETT, R. M. 2019. New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics,* 21**,** 584-594.

BAIROCH, A. & APWEILER, R. 1996. The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL. *Nucleic Acids Research,* 24**,** 21-25.

BAKKEN, L. R. 1997. Culturable and nonculturable bacteria in soil. *Modern soil microbiology.***,** 47-61.

BALLSCHMITE, K., HACKENBERG, R., JARMAN, W. M. & LOOSER, R. 2002. Man-made chemicals found in remote areas of the world: the experimental definition for POPs. *Environ Sci Pollut Res Int,* 9**,** 274-88.

BERINI, F., CASCIELLO, C., MARCONE, G. L. & MARINELLI, F. 2017. Metagenomics: novel enzymes from non-culturable microbes. *FEMS Microbiology Letters,* 364**,** fnx211.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic acids research,* 28**,** 235-242.

BOGART, J. W., CABEZAS, M. D., VÖGELI, B., WONG, D. A., KARIM, A. S. & JEWETT, M. C. 2021. Cell-free exploration of the natural product chemical space. *ChemBioChem,* 22**,** 84-91.

BONETTA, R. & VALENTINO, G. 2020. Machine learning techniques for protein function prediction. *Proteins: Structure, Function, and Bioinformatics,* 88**,** 397-413.

BORDAT, A., HOUVENAGHEL, M.-C. & GERMAN-RETANA, S. 2015. Gibson assembly: an easy way to clone potyviral full-length infectious cDNA clones expressing an ectopic VPg. *Virology journal,* 12**,** 89-89.

BUCHFINK, B., XIE, C. & HUSON, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods,* 12**,** 59-60.

CAPELLA-GUTIÉRREZ, S., SILLA-MARTÍNEZ, J. M. & GABALDÓN, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England),* 25**,** 1972-1973.

CARBONELL, P., JERVIS, A. J., ROBINSON, C. J., YAN, C., DUNSTAN, M., SWAINSTON, N., VINAIXA, M., HOLLYWOOD, K. A., CURRIN, A., RATTRAY, N. J. W., TAYLOR, S., SPIESS, R., SUNG, R., WILLIAMS, A. R., FELLOWS, D., STANFORD, N. J., MULHERIN, P., LE FEUVRE, R., BARRAN, P., GOODACRE, R., TURNER, N. J., GOBLE, C., CHEN, G. G., KELL, D. B., MICKLEFIELD, J., BREITLING, R., TAKANO, E., FAULON, J.-L. & SCRUTTON, N. S. 2018. An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Communications Biology,* 1**,** 66.

CASTRESANA, J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution,* 17**,** 540-552.

CERNIGLIA, C. E. 1997. Fungal metabolism of polycyclic aromatic hydrocarbons: past, present and future applications in bioremediation. *Journal of Industrial Microbiology and Biotechnology,* 19**,** 324-333.

CHEKAN, J. R., MCKINNIE, S. M. K., NOEL, J. P. & MOORE, B. S. 2020. Algal neurotoxin biosynthesis repurposes the terpene cyclase structural fold into an N-prenyltransferase. *Proceedings of the National Academy of Sciences,* 117**,** 12799.

CHEN, I. M. A., CHU, K., PALANIAPPAN, K., PILLAY, M., RATNER, A., HUANG, J., HUNTEMANN, M., VARGHESE, N., WHITE, J. R., SESHADRI, R., SMIRNOVA, T., KIRTON, E., JUNGBLUTH, S. P., WOYKE, T., ELOE-FADROSH, E. A., IVANOVA, N. N. & KYRPIDES, N. C. 2019. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research,* 47**,** D666-D677.

COLLERAN, E. 1997. Uses of bacteria in bioremediation. *Bioremediation protocols.* Springer.

COPLEY, S. D. 2003. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr Opin Chem Biol,* 7**,** 265-72.

CRITS-CHRISTOPH, A., DIAMOND, S., BUTTERFIELD, C. N., THOMAS, B. C. & BANFIELD, J. F. 2018. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature,* 558**,** 440-444.

CRUZ-MORALES, P., KOPP, J. F., MARTÍNEZ-GUERRERO, C., YÁÑEZ-GUERRA, L. A., SELEM-MOJICA, N., RAMOS-ABOITES, H., FELDMANN, J. & BARONA-GÓMEZ, F. 2016. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biology and Evolution,* 8**,** 1906-1916.

CSARDI, G. & NEPUSZ, T. 2006. The igraph software package for complex network research. *InterJournal, complex systems,* 1695**,** 1-9.

DANKO, D., BEZDAN, D., AFSHIN, E. E., AHSANUDDIN, S., BHATTACHARYA, C., BUTLER, D. J., CHNG, K. R., DONNELLAN, D., HECHT, J., JACKSON, K., KUCHIN, K., KARASIKOV, M., LYONS, A., MAK, L., MELESHKO, D., MUSTAFA, H., MUTAI, B., NECHES, R. Y., NG, A., NIKOLAYEVA, O., NIKOLAYEVA, T., PNG, E., RYON, K. A., SANCHEZ, J. L., SHAABAN, H., SIERRA, M. A., THOMAS, D., YOUNG, B., ABUDAYYEH, O. O., ALICEA, J., BHATTACHARYYA, M., BLEKHMAN, R., CASTRO-NALLAR, E., CAÑAS, A. M., CHATZIEFTHIMIOU, A. D., CRAWFORD, R. W., DE FILIPPIS, F., DENG, Y., DESNUES, C., DIAS-NETO, E., DYBWAD, M., ELHAIK, E., ERCOLINI, D., FROLOVA, A., GANKIN, D., GOOTENBERG, J. S., GRAF, A. B., GREEN, D. C., HAJIRASOULIHA, I., HASTINGS, J. J. A., HERNANDEZ, M., IRAOLA, G., JANG, S., KAHLES, A., KELLY, F. J., KNIGHTS, K., KYRPIDES, N. C., ŁABAJ, P. P., LEE, P. K. H., LEUNG, M. H. Y., LJUNGDAHL, P. O., MASON-BUCK, G., MCGRATH, K., MEYDAN, C., MONGODIN, E. F., MORAES, M. O., NAGARAJAN, N., NIETO-CABALLERO, M., NOUSHMEHR, H., OLIVEIRA, M., OSSOWSKI, S., OSUOLALE, O. O., ÖZCAN, O., PAEZ-ESPINO, D., RASCOVAN, N., RICHARD, H., RÄTSCH, G., SCHRIML, L. M., SEMMLER, T., SEZERMAN, O. U., SHI, L., SHI, T., SIAM, R., SONG, L. H., SUZUKI, H., COURT, D. S., TIGHE, S. W., TONG, X., UDEKWU, K. I., UGALDE, J. A., VALENTINE, B., VASSILEV, D. I., VAYNDORF,

E. M., VELAVAN, T. P., WU, J., ZAMBRANO, M. M., ZHU, J., ZHU, S., MASON, C. E., ABDULLAH, N., et al. 2021. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell,* 184**,** 3376-3393.e17.

DE CASTRO, E., SIGRIST, C. J., GATTIKER, A., BULLIARD, V., LANGENDIJK-GENEVAUX, P. S., GASTEIGER, E., BAIROCH, A. & HULO, N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res,* 34**,** W362-5.

DE ROND, T., ASAY, J. E. & MOORE, B. S. 2020. Co-Occurrence of Enzyme Domains Guides the Discovery of an Oxazolone Synthetase. *bioRxiv***,** 2020.06.11.147165.

DE ROND, T., GAO, J., ZARGAR, A., DE RAAD, M., CUNHA, J., NORTHEN, T. R. & KEASLING, J. D. 2019. A High-Throughput Mass Spectrometric Enzyme Activity Assay Enabling the Discovery of Cytochrome P450 Biocatalysts. *Angew Chem Int Ed Engl,* 58**,** 10114-10119.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

EDGAR, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26**,** 2460-2461.

EYERS, L., GEORGE, I., SCHULER, L., STENUIT, B., AGATHOS, S. N. & EL FANTROUSSI, S. 2004. Environmental genomics: exploring the unmined richness of microbes to degrade xenobiotics. *Appl Microbiol Biotechnol,* 66**,** 123-30.

FANG, Z.-M., LI, T.-L., CHANG, F., ZHOU, P., FANG, W., HONG, Y.-Z., ZHANG, X.-C., PENG, H. & XIAO, Y.-Z. 2012. A new marine bacterial laccase with chloride-enhancing, alkaline-dependent activity and dye decolorization ability. *Bioresource Technology,* 111**,** 36-41.

FERRER, M., MARTÍNEZ-MARTÍNEZ, M., BARGIELA, R., STREIT, W. R., GOLYSHINA, O. V. & GOLYSHIN, P. N. 2016. Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. *Microb Biotechnol,* 9**,** 22-34.

FLEMING, S. R., HIMES, P. M., GHODGE, S. V., GOTO, Y., SUGA, H. & BOWERS, A. A. 2020. Exploring the Post-translational Enzymology of PaaA by mRNA Display. *J Am Chem Soc,* 142**,** 5024-5028.

GABOR, E. M., ALKEMA, W. B. & JANSSEN, D. B. 2004. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol,* 6**,** 879-86.

GERLT, J. A., BOUVIER, J. T., DAVIDSON, D. B., IMKER, H. J., SADKHIN, B., SLATER, D. R. & WHALEN, K. L. 2015. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim Biophys Acta,* 1854**,** 1019-37.

GUINDON, S., LETHIEC, F., DUROUX, P. & GASCUEL, O. 2005. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research,* 33**,** W557-W559.

HANDELSMAN, J., RONDON, M. R., BRADY, S. F., CLARDY, J. & GOODMAN, R. M. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology,* 5**,** R245-R249.

HENDRIKSE, N. M., CHARPENTIER, G., NORDLING, E. & SYRÉN, P. O. 2018. Ancestral diterpene cyclases show increased thermostability and substrate acceptance. *Febs j,* 285**,** 4660-4673.

HERMES, C., RICHARZ, R., WIRTZ, D. A., PATT, J., HANKE, W., KEHRAUS, S., VOß, J. H., KÜPPERS, J., OHBAYASHI, T., NAMASIVAYAM, V., ALENFELDER, J., INOUE, A., MERGAERT, P., GÜTSCHOW, M., MÜLLER, C. E., KOSTENIS, E., KÖNIG, G. M. & CRÜSEMANN, M. 2021. Thioesterase-mediated side chain transesterification generates potent Gq signaling inhibitor FR900359. *Nature Communications,* 12**,** 144.

HOCHBERG, G. K. A. & THORNTON, J. W. 2017. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu Rev Biophys,* 46**,** 247-269.

HON, T., MARS, K., YOUNG, G., TSAI, Y.-C., KARALIUS, J. W., LANDOLIN, J. M., MAURER, N., KUDRNA, D., HARDIGAN, M. A., STEINER, C. C., KNAPP, S. J., WARE, D., SHAPIRO, B., PELUSO, P. & RANK, D. R. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data,* 7**,** 399.

HUERTA-CEPAS, J., SERRA, F. & BORK, P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular biology and evolution,* 33**,** 1635-1638.

IMDAHL, F., VAFADARNEJAD, E., HOMBERGER, C., SALIBA, A. E. & VOGEL, J. 2020. Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nat Microbiol,* 5**,** 1202-1206.

IWAI, S., CHAI, B., SUL, W. J., COLE, J. R., HASHSHAM, S. A. & TIEDJE, J. M. 2010. Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. *The ISME journal,* 4**,** 279-285.

JADEJA, N. B., MORE, R. P., PUROHIT, H. J. & KAPLEY, A. 2014. Metagenomic analysis of oxygenases from activated sludge. *Bioresource technology,* 165**,** 250-256.

JAIN, M., OLSEN, H. E., PATEN, B. & AKESON, M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology,* 17**,** 239.

JAROSZEWSKI, L., LI, Z., KRISHNA, S. S., BAKOLITSA, C., WOOLEY, J., DEACON, A. M., WILSON, I. A. & GODZIK, A. 2009. Exploration of uncharted regions of the protein universe. *PLoS Biol,* 7**,** e1000205.

JARZAB, A., KURZAWA, N., HOPF, T., MOERCH, M., ZECHA, J., LEIJTEN, N., BIAN, Y., MUSIOL, E., MASCHBERGER, M., STOEHR, G., BECHER, I., DALY, C., SAMARAS, P., MERGNER, J., SPANIER, B., ANGELOV, A., WERNER, T., BANTSCHEFF, M., WILHELM, M., KLINGENSPOR, M., LEMEER, S., LIEBL, W., HAHNE, H., SAVITSKI, M. M. & KUSTER, B. 2020. Meltome atlas—thermal proteome stability across the tree of life. *Nature Methods,* 17**,** 495-503.

JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A., BRIDGLAND, A., MEYER, C., KOHL, S. A. A., BALLARD, A. J., COWIE, A., ROMERA-PAREDES, B., NIKOLOV, S., JAIN, R., ADLER, J., BACK, T., PETERSEN, S., REIMAN, D., CLANCY, E., ZIELINSKI, M., STEINEGGER, M., PACHOLSKA, M., BERGHAMMER, T., BODENSTEIN, S., SILVER, D., VINYALS, O., SENIOR, A. W., KAVUKCUOGLU, K., KOHLI, P. & HASSABIS, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature,* 596**,** 583-589.

KATOH, K., MISAWA, K., KUMA, K. I. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research,* 30**,** 3059-3066.

KATZ, M., HOVER, B. M. & BRADY, S. F. 2016. Culture-independent discovery of natural products from soil metagenomes. *J Ind Microbiol Biotechnol,* 43**,** 129-41.

KAUTSAR, S. A., BLIN, K., SHAW, S., NAVARRO-MUÑOZ, J. C., TERLOUW, B. R., VAN DER HOOFT, J. J. J., VAN SANTEN, J. A., TRACANNA, V., SUAREZ DURAN, H. G., PASCAL ANDREU, V., SELEM-MOJICA, N., ALANJARY, M., ROBINSON, S. L., LUND, G., EPSTEIN, S. C., SISTO, A. C., CHARKOUDIAN, L. K., COLLEMARE, J., LININGTON, R. G., WEBER, T. & MEDEMA, M. H. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res,* 48**,** D454-d458.

KAUTSAR, S. A., VAN DER HOOFT, J. J. J., DE RIDDER, D. & MEDEMA, M. H. 2021. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience,* 10.

KELLEY, L. A., MEZULIS, S., YATES, C. M., WASS, M. N. & STERNBERG, M. J. E. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols,* 10**,** 845-858.

KOZLOV, A. M., DARRIBA, D., FLOURI, T., MOREL, B. & STAMATAKIS, A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics,* 35**,** 4453-4455.

KURGAN, L., RAZIB, A. A., AGHAKHANI, S., DICK, S., MIZIANTY, M. & JAHANDIDEH, S. 2009. CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Struct Biol,* 9**,** 50.

KVIST, T., SONDT-MARCUSSEN, L. & MIKKELSEN, M. J. 2014. Partition Enrichment of Nucleotide Sequences (PINS) - A Generally Applicable, Sequence Based Method for Enrichment of Complex DNA Samples. *PLOS ONE,* 9**,** e106817.

LEINONEN, R., SUGAWARA, H. & SHUMWAY, M. 2011. The sequence read archive. *Nucleic Acids Res,* 39**,** D19-21.

LI, W. & GODZIK, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-1659.

LIBIS, V., ANTONOVSKY, N., ZHANG, M., SHANG, Z., MONTIEL, D., MANIKO, J., TERNEI, M. A., CALLE, P. Y., LEMETRE, C., OWEN, J. G. & BRADY, S. F. 2019. Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. *Nature Communications,* 10**,** 3848.

LOMAKINA, A. V., MAMAEVA, E. V., GALACHYANTS, Y. P., PETROVA, D. P., POGODAEVA, T. V., SHUBENKOVA, O. V., KHABUEV, A. V., MOROZOV, I. V. & ZEMSKAYA, T. I. 2018. Diversity of Archaea in Bottom Sediments of the Discharge Areas With Oil- and Gas-Bearing Fluids in Lake Baikal. *Geomicrobiology Journal,* 35**,** 50-63.

LU, Z., DENG, Y., VAN NOSTRAND, J. D., HE, Z., VOORDECKERS, J., ZHOU, A., LEE, Y.-J., MASON, O. U., DUBINSKY, E. A., CHAVARRIA, K. L., TOM, L. M., FORTNEY, J. L., LAMENDELLA, R., JANSSON, J. K., D'HAESELEER, P., HAZEN, T. C. & ZHOU, J. 2012. Microbial gene functions enriched in the Deepwater Horizon deep-sea oil plume. *The ISME Journal,* 6**,** 451-460.

MADEIRA, F., PARK, Y. M., LEE, J., BUSO, N., GUR, T., MADHUSOODANAN, N., BASUTKAR, P., TIVEY, A. R. N., POTTER, S. C., FINN, R. D. & LOPEZ, R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res,* 47**,** W636-w641.

MAY, O., NGUYEN, P. T. & ARNOLD, F. H. 2000. Inverting enantioselectivity by directed evolution of hydantoinase for improved production of L-methionine. *Nat Biotechnol,* 18**,** 317-20.

MCGINNIS, S. & MADDEN, T. L. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research,* 32**,** W20-W25.

MITCHELL, A. L., ALMEIDA, A., BERACOCHEA, M., BOLAND, M., BURGIN, J., COCHRANE, G., CRUSOE, M. R., KALE, V., POTTER, S. C., RICHARDSON, L. J., SAKHAROVA, E., SCHEREMETJEW, M., KOROBEYNIKOV, A., SHLEMOV, A., KUNYAVSKAYA, O., LAPIDUS, A. & FINN, R. D. 2020. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research,* 48**,** D570-D578.

MORA, M., WINK, L., KÖGLER, I., MAHNERT, A., RETTBERG, P., SCHWENDNER, P., DEMETS, R., COCKELL, C., ALEKHOVA, T., KLINGL, A., KRAUSE, R., ZOLOTARIOF, A., ALEXANDROVA, A. & MOISSL-EICHINGER, C. 2019. Space Station conditions are selective but do not alter microbial characteristics relevant to human health. *Nature Communications,* 10**,** 3990.

MORI, T., CAHN, J. K. B., WILSON, M. C., MEODED, R. A., WIEBACH, V., MARTINEZ, A. F. C., HELFRICH, E. J. N., ALBERSMEIER, A., WIBBERG, D., DÄTWYLER, S., KEREN, R., LAVY, A., RÜCKERT, C., ILAN, M., KALINOWSKI, J., MATSUNAGA, S., TAKEYAMA, H. & PIEL, J. 2018. Single-bacterial genomics validates rich and varied specialized metabolism of uncultivated Entotheonella sponge symbionts. *Proceedings of the National Academy of Sciences,* 115**,** 1718-1723.

MORIMOTO, S. & FUJII, T. 2009. A new approach to retrieve full lengths of functional genes from soil by PCR-DGGE and metagenome walking. *Applied Microbiology and Biotechnology,* 83**,** 389-396.

MRKSICH, M. 2008. Mass spectrometry of self-assembled monolayers: a new tool for molecular surface science. *ACS nano,* 2**,** 7-18.

NAVARRO-MUÑOZ, J. C., SELEM-MOJICA, N., MULLOWNEY, M. W., KAUTSAR, S. A., TRYON, J. H., PARKINSON, E. I., DE LOS SANTOS, E. L. C., YEONG, M., CRUZ-MORALES, P., ABUBUCKER, S., ROETERS, A., LOKHORST, W., FERNANDEZ-GUERRA, A., CAPPELINI, L. T. D., GOERING, A. W., THOMSON, R. J., METCALF, W. W., KELLEHER, N. L., BARONA-GOMEZ, F. & MEDEMA, M. H. 2020. A computational framework to explore large-scale biosynthetic diversity. *Nature chemical biology,* 16**,** 60-68.

NAYFACH, S., ROUX, S., SESHADRI, R., UDWARY, D., VARGHESE, N., SCHULZ, F., WU, D., PAEZ-ESPINO, D., CHEN, I. M., HUNTEMANN, M., PALANIAPPAN, K., LADAU, J., MUKHERJEE, S., REDDY, T. B. K., NIELSEN, T., KIRTON, E., FARIA, J. P., EDIRISINGHE, J. N., HENRY, C. S., JUNGBLUTH, S. P., CHIVIAN, D., DEHAL, P., WOOD-CHARLSON, E. M., ARKIN, A. P., TRINGE, S. G., VISEL, A., ABREU, H., ACINAS, S. G., ALLEN, E., ALLEN, M. A., ALTEIO, L. V., ANDERSEN, G., ANESIO, A. M., ATTWOOD, G., AVILA-MAGAÑA, V., BADIS, Y., BAILEY, J., BAKER, B., BALDRIAN, P., BARTON, H. A., BECK, D. A. C., BECRAFT, E. D., BELLER, H. R., BEMAN, J. M., BERNIER-LATMANI, R., BERRY, T. D., BERTAGNOLLI, A., BERTILSSON, S., BHATNAGAR, J. M., BIRD, J. T., BLANCHARD, J. L., BLUMER-SCHUETTE, S. E., BOHANNAN, B., BORTON, M. A., BRADY, A., BRAWLEY, S. H., BRODIE, J., BROWN, S., BRUM, J. R., BRUNE, A., BRYANT, D. A., BUCHAN, A., BUCKLEY, D. H., BUONGIORNO, J., CADILLO-QUIROZ, H., CAFFREY, S. M., CAMPBELL, A. N., CAMPBELL, B., CARR, S., CARROLL, J., CARY, S. C., CATES, A. M., CATTOLICO, R. A., CAVICCHIOLI, R., CHISTOSERDOVA, L., COLEMAN, M. L., CONSTANT, P., CONWAY, J. M., MAC CORMACK, W. P., CROWE, S., CRUMP, B., CURRIE, C., DALY, R., DEANGELIS, K. M., DENEF, V., DENMAN, S. E., DESTA, A., DIONISI, H., DODSWORTH, J., DOMBROWSKI, N., DONOHUE, T., DOPSON, M., DRISCOLL, T., DUNFIELD, P., DUPONT, C. L., DYNARSKI, K. A., EDGCOMB, V., EDWARDS, E. A., ELSHAHED, M. S., FIGUEROA, I., et al. 2021. A genomic catalog of Earth's microbiomes. *Nature Biotechnology,* 39**,** 499-509.

NGUYEN, L.-T., SCHMIDT, H. A., VON HAESELER, A. & MINH, B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution,* 32**,** 268-274.

NICOLAS, A. M., JAFFE, A. L., NUCCIO, E. E., TAGA, M. E., FIRESTONE, M. K. & BANFIELD, J. F. 2020. Unexpected diversity of CPR bacteria and nanoarchaea in the rare biosphere of rhizosphere-associated grassland soil. *bioRxiv*, 2020.07.13.194282.

NOGALES, B., TIMMIS, K. N., NEDWELL, D. B. & OSBORN, A. M. 2002. Detection and diversity of expressed denitrification genes in estuarine sediments after reverse transcription-PCR amplification from mRNA. *Appl Environ Microbiol,* 68**,** 5017-25.

NORTH, J. A., NARROWE, A. B., XIONG, W., BYERLY, K. M., ZHAO, G., YOUNG, S. J., MURALI, S., WILDENTHAL, J. A., CANNON, W. R., WRIGHTON, K. C., HETTICH, R. L. & TABITA, F. R. 2020. A nitrogenase-like enzyme system catalyzes methionine, ethylene, and methane biogenesis. *Science,* 369**,** 1094-1098.

NORTHEN, T. R., LEE, J. C., HOANG, L., RAYMOND, J., HWANG, D. R., YANNONE, S. M., WONG, C. H. & SIUZDAK, G. 2008. A nanostructure-initiator mass spectrometry-based enzyme activity assay. *Proc Natl Acad Sci U S A,* 105**,** 3678-83.

ONO, K., MUETZE, T., KOLISHOVSKI, G., SHANNON, P. & DEMCHAK, B. 2015. CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API. *F1000Research,* 4**,** 478-478.

OVERTON, I. M., PADOVANI, G., GIROLAMI, M. A. & BARTON, G. J. 2008. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics,* 24**,** 901-907.

OVERTON, I. M., VAN NIEKERK, C. A. & BARTON, G. J. 2011. XANNpred: neural nets that predict the propensity of a protein to yield diffraction-quality crystals. *Proteins,* 79**,** 1027-33.

PACE, G. M., IVANCIC, M. T., EDWARDS, G. L., IWATA, B. A. & PAGE, T. J. 1985. Assessment of stimulus preference and reinforcer value with profoundly retarded individuals. *J Appl Behav Anal,* 18**,** 249-55.

PAOLI, L., RUSCHEWEYH, H.-J., FORNERIS, C. C., KAUTSAR, S., CLAYSSEN, Q., SALAZAR, G., MILANESE, A., GEHRIG, D., LARRALDE, M., CARROLL, L. M., SÁNCHEZ, P., ZAYED, A. A., CRONIN, D. R., ACINAS, S. G., BORK, P., BOWLER, C., DELMONT, T. O., SULLIVAN, M. B., WINCKER, P., ZELLER, G., ROBINSON, S. L., PIEL, J. & SUNAGAWA, S. 2021. Uncharted biosynthetic potential of the ocean microbiome. *bioRxiv*, 2021.03.24.436479.

PASCAL ANDREU, V., AUGUSTIJN, H. E., VAN DEN BERG, K., VAN DER HOOFT, J. J. J., FISCHBACH, M. A. & MEDEMA, M. H. 2020. BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. *bioRxiv*, 2020.12.14.422671.

PHAM, V. H. & KIM, J. 2012. Cultivation of unculturable soil bacteria. *Trends Biotechnol,* 30**,** 475-84.

PIEPER, D. H., MARTINS DOS SANTOS, V. A. & GOLYSHIN, P. N. 2004. Genomic and mechanistic insights into the biodegradation of organic pollutants. *Curr Opin Biotechnol,* 15**,** 215-24.

PRICE, M. N., DEHAL, P. S. & ARKIN, A. P. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE,* 5**,** e9490.

QIN, J., LI, R., RAES, J., ARUMUGAM, M., BURGDORF, K. S., MANICHANH, C., NIELSEN, T., PONS, N., LEVENEZ, F., YAMADA, T., MENDE, D. R., LI, J., XU, J., LI, S., LI, D., CAO, J., WANG, B., LIANG, H., ZHENG, H., XIE, Y., TAP, J., LEPAGE, P., BERTALAN, M., BATTO, J.-M., HANSEN, T., LE PASLIER, D., LINNEBERG, A., NIELSEN, H. B., PELLETIER, E., RENAULT, P., SICHERITZ-PONTEN, T., TURNER, K., ZHU, H., YU, C., LI, S., JIAN, M., ZHOU, Y., LI, Y., ZHANG, X., LI, S., QIN, N., YANG, H., WANG, J., BRUNAK, S., DORÉ, J., GUARNER, F., KRISTIANSEN, K., PEDERSEN, O., PARKHILL, J., WEISSENBACH, J., ANTOLIN, M., ARTIGUENAVE, F., BLOTTIERE, H., BORRUEL, N., BRULS, T., CASELLAS, F., CHERVAUX, C., CULTRONE, A., DELORME, C., DENARIAZ, G., DERVYN, R., FORTE, M., FRISS, C., VAN DE GUCHTE, M., GUEDON, E., HAIMET, F., JAMET, A., JUSTE, C., KACI, G., KLEEREBEZEM, M., KNOL, J., KRISTENSEN, M., LAYEC, S., LE ROUX, K., LECLERC, M., MAGUIN, E., MELO MINARDI, R., OOZEER, R., RESCIGNO, M., SANCHEZ, N., TIMS, S., TORREJON, T., VARELA, E., DE VOS, W., WINOGRADSKY, Y., ZOETENDAL, E., BORK, P., EHRLICH, S. D., WANG, J. & META, H. I. T. C. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature,* 464**,** 59-65.

RAUWERDINK, A. & KAZLAUSKAS, R. J. 2015. How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: the Serine-Histidine-Aspartate Catalytic Triad of α/β-Hydrolase Fold Enzymes. *ACS Catal,* 5**,** 6153-6176.

REKDAL, V. M., BERNADINO, P. N., LUESCHER, M. U., KIAMEHR, S., LE, C., BISANZ, J. E., TURNBAUGH, P. J., BESS, E. N. & BALSKUS, E. P. 2020. A widely distributed metalloenzyme class enables gut microbial metabolism of host-and diet-derived catechols. *Elife,* 9**,** e50845.

RIBEIRO, A. J. M., HOLLIDAY, G. L., FURNHAM, N., TYZACK, J. D., FERRIS, K. & THORNTON, J. M. 2018. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res,* 46**,** D618-d623.

ROBINSON, T., MCMULLAN, G., MARCHANT, R. & NIGAM, P. 2001. Remediation of dyes in textile effluent: a critical review on current treatment technologies with a proposed alternative. *Bioresource Technology,* 77**,** 247-255.

ROY, A., KUCUKURAL, A. & ZHANG, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols,* 5**,** 725-738.

RUEPP, A., ZOLLNER, A., MAIER, D., ALBERMANN, K., HANI, J., MOKREJS, M., TETKO, I., GÜLDENER, U., MANNHAUPT, G., MÜNSTERKÖTTER, M. & MEWES, H. W. 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res,* 32**,** 5539-45.

SCHWEDE, T., KOPP, J., GUEX, N. & PEITSCH, M. C. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic acids research,* 31**,** 3381-3385.

SÉLEM-MOJICA, N., AGUILAR, C., GUTIÉRREZ-GARCÍA, K., MARTÍNEZ-GUERRERO, C. E. & BARONA-GÓMEZ, F. 2019. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microbial genomics,* 5**,** e000260.

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research,* 13**,** 2498-2504.

SIEVERS, F. & HIGGINS, D. G. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci,* 27**,** 135-145.

SLABINSKI, L., JAROSZEWSKI, L., RYCHLEWSKI, L., WILSON, I. A., LESLEY, S. A. & GODZIK, A. 2007. XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics,* 23**,** 3403-3405.

STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England),* 30**,** 1312-1313.

STEINEGGER, M. & SÖDING, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology,* 35**,** 1026-1028.

STEPANAUSKAS, R., FERGUSSON, E. A., BROWN, J., POULTON, N. J., TUPPER, B., LABONTÉ, J. M., BECRAFT, E. D., BROWN, J. M., PACHIADAKI, M. G., POVILAITIS, T., THOMPSON, B. P., MASCENA, C. J., BELLOWS, W. K. & LUBYS, A. 2017. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nature Communications,* 8**,** 84.

STEVENS, F. R., GAUGHAN, A. E., LINARD, C. & TATEM, A. J. 2015. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE,* 10**,** e0107042.

STEWART, R. D., AUFFRET, M. D., WARR, A., WISER, A. H., PRESS, M. O., LANGFORD, K. W., LIACHKO, I., SNELLING, T. J., DEWHURST, R. J., WALKER, A. W., ROEHE, R. & WATSON, M. 2018. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun,* 9**,** 870.

SUKUL, P., SCHÄKERMANN, S., BANDOW, J. E., KUSNEZOWA, A., NOWROUSIAN, M. & LEICHERT, L. I. 2017. Simple discovery of bacterial biocatalysts from environmental samples through functional metaproteomics. *Microbiome,* 5**,** 28.

SUL, W. J., PARK, J., QUENSEN, J. F., 3RD, RODRIGUES, J. L., SELIGER, L., TSOI, T. V., ZYLSTRA, G. J. & TIEDJE, J. M. 2009. DNA-stable isotope probing integrated with metagenomics for retrieval of biphenyl dioxygenase genes from polychlorinated biphenyl-contaminated river sediment. *Appl Environ Microbiol,* 75**,** 5501-6.

SUTHERLAND, T. D., HORNE, I., WEIR, K. M., COPPIN, C. W., WILLIAMS, M. R., SELLECK, M., RUSSELL, R. J. & OAKESHOTT, J. G. 2004. Enzymatic bioremediation: from enzyme discovery to applications. *Clin Exp Pharmacol Physiol,* 31**,** 817-21.

SZKLARCZYK, D., GABLE, A. L., LYON, D., JUNGE, A., WYDER, S., HUERTA-CEPAS, J., SIMONOVIC, M., DONCHEVA, N. T., MORRIS, J. H., BORK, P., JENSEN, L. J. & MERING, C. V. 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res,* 47**,** D607-d613.

TASSE, L., BERCOVICI, J., PIZZUT-SERIN, S., ROBE, P., TAP, J., KLOPP, C., CANTAREL, B. L., COUTINHO, P. M., HENRISSAT, B. & LECLERC, M. 2010. Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome research,* 20**,** 1605-1612.

TEUFEL, F., ALMAGRO ARMENTEROS, J. J., JOHANSEN, A. R., GÍSLASON, M. H., PIHL, S. I., TSIRIGOS, K. D., WINTHER, O., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*.

THE UNIPROT, C. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research,* 47**,** D506-D515.

THEERACHAT, M., EMOND, S., CAMBON, E., BORDES, F., MARTY, A., NICAUD, J. M., CHULALAKSANANUKUL, W., GUIEYSSE, D., REMAUD-SIMÉON, M. & MOREL, S. 2012. Engineering and production of laccase from Trametes versicolor in the yeast Yarrowia lipolytica. *Bioresour Technol,* 125**,** 267-74.

TIAN, W., CHEN, C., LEI, X., ZHAO, J. & LIANG, J. 2018. CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res,* 46**,** W363-w367.

TOBIAS, N. J. & BODE, H. B. 2019. Heterogeneity in Bacterial Specialized Metabolism. *J Mol Biol,* 431**,** 4589-4598.

TREUSCH, A. H., LEININGER, S., KLETZIN, A., SCHUSTER, S. C., KLENK, H.-P. & SCHLEPER, C. 2005. Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environmental Microbiology,* 7**,** 1985-1995.

UFARTÉ, L., LAVILLE, É., DUQUESNE, S. & POTOCKI-VERONESE, G. 2015. Metagenomics for the discovery of pollutant degrading enzymes. *Biotechnol Adv,* 33**,** 1845-54.

WANG, H., LI, Z., JIA, R., HOU, Y., YIN, J., BIAN, X., LI, A., MÜLLER, R., STEWART, A. F., FU, J. & ZHANG, Y. 2016. RecET direct cloning and Redαβ recombineering of biosynthetic gene clusters, large operons or single genes for heterologous expression. *Nat Protoc,* 11**,** 1175-90.

WANG, Y., CHEN, Y., ZHOU, Q., HUANG, S., NING, K., XU, J., KALIN, R. M., ROLFE, S. & HUANG, W. E. 2012. A Culture-Independent Approach to Unravel Uncultured Bacteria and Functional Genes in a Complex Microbial Community. *PLOS ONE,* 7**,** e47530.

WEBB, O. F., PHELPS, T. J., BIENKOWSKI, P. R., DIGRAZIA, P. M., WHITE, D. C. & SAYLER, G. S. 1992. Enzyme nomenclature.

WILSON, M. C., MORI, T., RÜCKERT, C., URIA, A. R., HELF, M. J., TAKADA, K., GERNERT, C., STEFFENS, U. A. E., HEYCKE, N., SCHMITT, S., RINKE, C., HELFRICH, E. J. N., BRACHMANN, A. O., GURGUI, C., WAKIMOTO, T., KRACHT, M., CRÜSEMANN, M., HENTSCHEL, U., ABE, I., MATSUNAGA, S., KALINOWSKI, J., TAKEYAMA, H. & PIEL, J. 2014. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature,* 506**,** 58-62.

WOYKE, T., DOUD, D. F. R. & SCHULZ, F. 2017. The trajectory of microbial single-cell sequencing. *Nat Methods,* 14**,** 1045-1054.

WU, Z., JOHNSTON, K. E., ARNOLD, F. H. & YANG, K. K. 2021. Protein sequence design with deep generative models. *Current Opinion in Chemical Biology,* 65**,** 18-27.

XIA, X., GURR, G. M., VASSEUR, L., ZHENG, D., ZHONG, H., QIN, B., LIN, J., WANG, Y., SONG, F., LI, Y., LIN, H. & YOU, M. 2017. Metagenomic Sequencing of Diamondback Moth Gut Microbiome Unveils Key Holobiont Adaptations for Herbivory. *Frontiers in Microbiology,* 8.

YAFFE, E. & RELMAN, D. A. 2020. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat Microbiol,* 5**,** 343-353.

YEOM, S.-J., KIM, M., KWON, K. K., FU, Y., RHA, E., PARK, S.-H., LEE, H., KIM, H., LEE, D.-H., KIM, D.-M. & LEE, S.-G. 2018. A synthetic microbial biosensor for high-throughput screening of lactam biocatalysts. *Nature Communications,* 9**,** 5053.

YOU, R., HUANG, X. & ZHU, S. 2018. DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation. *Methods,* 145**,** 82-90.

YOUNGS, N., PENFOLD-BROWN, D., BONNEAU, R. & SHASHA, D. 2014. Negative Example Selection for Protein Function Prediction: The NoGO Database. *PLOS Computational Biology,* 10**,** e1003644.

YU, G., LAM, T. T.-Y., ZHU, H. & GUAN, Y. 2018. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular biology and evolution,* 35**,** 3041-3043.

ZALLOT, R., OBERG, N. & GERLT, J. A. 2019. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry,* 58**,** 4169-4182.

ZAPRASIS, A., LIU, Y.-J., LIU, S.-J., DRAKE, H. L. & HORN, M. A. 2010. Abundance of novel and diverse tfdA-like genes, encoding putative phenoxyalkanoic acid herbicide-degrading dioxygenases, in soil. *Applied and Environmental Microbiology,* 76**,** 119-128.

ZHANG, J. J., TANG, X. & MOORE, B. S. 2019. Genetic platforms for heterologous expression of microbial natural products. *Natural product reports,* 36**,** 1313-1332.

ZHANG, M. M., WANG, Y., ANG, E. L. & ZHAO, H. 2016. Engineering microbial hosts for production of bacterial natural products. *Natural product reports,* 33**,** 963-987.

ZHANG, Y., BUCHHOLZ, F., MUYRERS, J. P. & STEWART, A. F. 1998. A new logic for DNA engineering using recombination in Escherichia coli. *Nat Genet,* 20**,** 123-8.

ZHOU, N., JIANG, Y., BERGQUIST, T. R., LEE, A. J., KACSOH, B. Z., CROCKER, A. W., LEWIS, K. A., GEORGHIOU, G., NGUYEN, H. N. & HAMID, M. N. 2019. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology,* 20**,** 1-23.

ZIEMERT, N., ALANJARY, M. & WEBER, T. 2016. The evolution of genome mining in microbes–a review. *Natural product reports,* 33**,** 988-1005.

10. Legends

Figure 1. Flowchart of experimental and computational approaches in metagenomic enzyme discovery.