Exploring functional annotation through genomic and metagenomic data mining

by

Briallen Lobb

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Biology

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Robert Beiko Professor, Faculty of Computer Science, Dalhousie University
Supervisor:	Dr. Andrew Doxey Associate Professor, Department of Biology, University of Waterloo
Internal Member:	Dr. Gabriel Moreno-Hagelsieb Professor, Department of Biology, Wilfrid Laurier University, and Adjunct Associate Professor, Department of Biology, University of Waterloo
Internal Member:	Dr. Trevor Charles Professor, Department of Biology, University of Waterloo
Internal-External Member:	Dr. Bin Ma Professor, Cheriton School of Computer Science, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Functional profiling of genomes and metagenomes, as well as data mining for novel proteins, all rely on computational methods for functional annotation of protein sequences. Standard methods assign protein function based on detected homology to reference sequences, but often leave behind a significant fraction of hypothetical sequences ("dark matter") that cannot be annotated. To maximize our ability to extract new biological insights from newly sequenced genomes, it is critical to understand the advantages and limitations of homology-based annotation, and explore alternative methods for inferring function. In this thesis, I performed a comprehensive exploration of computational protein annotation, with a focus on bacterial genomes and metagenomes. First, I applied homology-based methods to functionally annotate and analyze original datasets including newly sequenced *Streptomyces* strains, a wastewater metagenome, and microbial communities involved in vertebrate decomposition. These studies identified genes and functions of interest including cellulases, antibiotic resistance genes, and virulence factors. I then explored the limits of homology-based annotation by measuring annotation coverage, the fraction of annotated proteins in a proteome, across ∼27,000 organisms in the microbial tree of life. This study demonstrated a wide range in annotation coverage across bacteria, from 2-86%. In addition, it revealed multiple factors including taxonomy, genome size, and research bias, as heavy influences on the degree to which proteomes could be annotated. To gain biological insights into hypothetical proteins of unknown function, I analyzed 4,049 domains of unknown function (DUFs) from Pfam. Using phylogenomic, taxonomic and metagenomic information, I detected statistical associations between domains and biological traits. Association-based methods uncovered environment, lineage, and/or pathogen associations in just under half of all DUFs and highlighted new families such as DUF4765 as intriguing virulence factor candidates. Finally, I constructed a database of "ORFan" metagenomic sequences that cannot be annotated using standard approaches, and inferred functions for tens of thousands of these sequences using profile-profile comparison approaches. Motif analysis and genomic context validated these predictions, enabling the discovery of hundreds of novel candidate metalloproteases. Protein "dark matter", which includes a large pool of unannotated coding sequences, is an incredible resource to find new proteins and functions of interest, and included are suggestions on how to prioritize these sequences for future study. A combination of homology-based and alternative annotation methods will be most effective for broad functional profiling of genomes and metagenomes, and can push the boundaries for functional interpretation of sequence data.

Acknowledgements

First, I want to thank my supervisor, Dr. Andrew Doxey, who was always a well-spring of excitement and ideas. Without your trust in me, this would not have happened. Doxey lab members past and present, with extra thanks to Hina Bandukwala, Mike Mansfield, and Jen Aguair, have been essential as both sound-boards and emotion-sinks. Many thanks to the faculty members who I've met and who have helped me along the way: Dr. David Rose, for taking a chance on me, as well as Dr. Trevor Charles, Dr. Jiujun Cheng, Dr. Elizabeth Meiering, and Dr. Josh Neufeld for giving me the opportunity to experience so many facets of science.

Additional thanks to my collaborators from other labs, Dr. Gabriel Moreno-Hagelsieb, Dr. Anthony Adegoke, Dr. Olayinka Aiyegoro, Dr. Olubukola Babalola, Dr. Kesen Ma, and Dr. Paul Craig who have provided me with the data I structured my degree around and for their help in guiding our manuscripts. Your input was an essential part of this thesis.

I am grateful for receiving funding from NSERC, OGS, and the University of Waterloo. This has given me the freedom to be flexible in my research interests and spend more time exploring other bioinformatic disciplines. This work was also made possible through SHARCNET (https://www.sharcnet.ca) supercomputing resources.

Special thanks to my partner, Mike Van Dorp, for always being patient with my ranting and putting up with my moods when even I could not. Your calm voice of reason has kept me on track. Thank you to both of our parents for their kindness and generosity. And finally, I can't express how lucky I am to have parents who have been truly excited about every step that I've taken.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> [...] there seem to be two possible views on functional completeness: first, that we can reliably predict functions for the majority of proteins; or second, that there is a seemingly endless repertoire of specialized families and we cannot predict whether we are approaching the limits of protein function space.
>
> *Protein Function Space: Viewing the Limits or Limited by our View?*
> Jeroen Raes et al.[254]

Material in this chapter has been published as part of Lobb and Doxey (2016).[177] The published manuscript is available here:

> B. Lobb and A. C. Doxey. Novel function discovery through sequence and structural data mining. *Current Opinion in Structural Biology*, 38:53-61, 2016.[177]
> https://doi.org/10.1016/j.sbi.2016.05.017

Function is an expansive and complex term that is hard to define. With regards to biology, this can refer to a biochemical level (e.g. with residues interacting to facilitate reactions), a molecular network (e.g. metabolic pathways within the cell), and a higher-level cellular role (e.g. in a community or within a multi-cellular organism). The concept of function used within this thesis is broad and includes "molecular function" as well as "biological process", consistent with functional ontologies[19] and assessments of prediction methods.[253] This functional information comes from a patchwork of purification, biochemical assays, physiological experiments and phenotypic observations. One experiment alone cannot fully

describe a protein's role at all perceived levels. In order to build a picture of the protein's functional facets, multiple sources of functional information must be compared and combined. Translation of experimentally-derived functions into usable functional terms, *annotations*, is an on-going process, with 40,230 molecular function and biological process Gene Ontology (GO) terms currently applied to proteins across many databases[1]. Annotations also include other database ontologies/vocabularies, compiled notes (e.g. functional summaries from Interpro or Uniprot), researcher-bestowed protein names (e.g. autoagglutinating adhesin), identified domains, and protein family associations.

With the development of cheaper and faster sequencing technologies, sequence databases have been flooded with new submissions. Since Dec. 2019, there have been over 1,000,000 new entries in Genbank alone[2]. In the absence of thorough experimental characterization, a protein would be without any functional annotations without some kind of transfer of functional information from one sequence to another. As orthologous[3] proteins generally possess the same functions, finding a protein's homologs can provide a source for annotation transfer. Thus, classic methods, like BLAST,[8] were developed in order to efficiently find similar sequences with a high probability of homology. However, the sequence alignments that BLAST uses are sometimes not sensitive enough to find divergent, but still functionally similar, protein homologs. Newer methods, discussed in this thesis, incorporate protein family models, search iterations, and combinations of different reference databases to achieve greater levels of annotation success. When these homology-based methods do not find sequence matches, other annotation strategies including domain, motif, genomic neighborhood, association, co-occurrence, and strucural analyses can be explored.

Annotation is used to gather putative functions for a sequence, analyze the functional potential of genomes and metagenomes, and to guide future study. With a pool of uncharacterized proteins, finding novel proteins, that have unexplored roles or locations, is an added benefit of annotation. Data mining, or the extraction of useful, interesting information from a dataset, is the basis for many important discoveries (e.g. proteolytic flagellin[63]). With the wealth of sequence data available, in databases and through new sequencing ventures, finding interesting novel proteins is eminently achievable. Tailored approaches leverage known information from different sources and methods in order to strengthen predictions. In this chapter, I will introduce some of the many different strategies for finding proteins of interest, and discuss how this can be accomplished through different annotation techniques.

[1] http://geneontology.org/stats.html for the current release v2020-06-01

[2] 1,789,213 sequences were deposited into GenBank from Dec. 2019 - Jun. 2020 (https://www.ncbi.nlm.nih.gov/genbank/statistics/)

[3] Sequences that are related across a speciation event.

1.1 Homology-based functional annotation

1.1.1 Protein sequence and homology

Homologous sequences share a common ancestor and are thus more similar to each other than they are to other unrelated sequences. As homologs, especially orthologs[4], are generally considered to be functionally similar, this concept forms the basis for functional annotation.[160] Using sequence alignments enables one to infer homology between sequences, by comparing their sequence similarity. In an alignment, ideally the related amino acids or nucleotides are aligned, with gaps indicating insertions or deletions relative to their most recent common ancestor. The alignment is then scored, to get a sense of alignment quality. In the classic BLAST[8] (basic local alignment search tool) implementation, a statistical measure of alignment "significance" is also used, the *E-value* (expect value). The *E*-value does not only look at the alignment itself but also the context in which the match was found[5], taking into account the chance that it is a random match and not a homologous sequence. For example, an *E*-value of 1 means that one sequence match is expected by chance in a database of the same size, with a similar alignment score. If the *E*-value is low, it indicates that the two sequences have a good alignment quality and therefore, have evidence for homology. It is important to note that low alignment quality is not a guarantee that there is no evolutionary relationship between sequences, as the sequences may have diverged so far that significant sequence similarity is no longer detected. BLAST is not the only sequence-sequence search tool. There are many sequence alignment tools with different levels of sensitivity and speed.[27,66,236,263,291] These can all be used in order to find sequences with high similarity, transferring functional information from one sequence to another.

1.1.2 Protein and domain family profiles

Groups of related sequences with a common ancestor together form a protein family. Multiple sequence alignments (MSAs), with multiple family members aligned, enable an exploration of shared sequence traits between the proteins. After aligning family members, certain positions often show up as less variable. Conserved residues in these families are sometimes catalytic sites or ligands, like in the metalloprotease HExxH motif found in the

[4]Paralogs, sequences related across a gene duplication event, generally diverge to the point of shifting in function.[160,231]

[5]The *E*-value incorporates the length of query sequence and the database size as well as the bit-score of the sequence alignment.

Peptidase_M60 (PF13402) family. These residues can also be conserved due to structural importance (e.g. disulfide bridges or binding site pockets).[38,147,172] These conserved residues are used as a functional and structural *signature* for the family, allowing more divergent members to be discovered. PSSMs and HMMs[6] act as models or profiles of these protein families, using information about the amino acid distribution at each position to inform sequence searches. Example sequence-model searches are hmmscan,[64] that scan protein sequences against models of protein and/or domain families, and PSIBLAST[7] (position-specific iterative BLAST), which builds a model of top-scoring database hits across multiple iterative model-to-sequence searches.

Models can be built, not just from full-length proteins but also from domains. A domain is a folded structural unit within a protein and proteins are made up of one or more of these domain units. Domain profiles are used to find matches to building blocks of proteins (such as catalytic or binding regions), sometimes allowing functional information transfer even in the absence of a full protein match.[75] Thus, a domain model is a powerful tool for functional annotation, taking advantage of the evolutionary phenomenon of domain recombination as a means of generating new functional combinations.[22,163] Through sequence-profile methods, functional annotations are transferred based on either collective information about a protein or a domain family, with a more sensitive approach for protein classification.

1.1.3 Where are we now?

Modern sequencing technologies continue to accelerate the collection of new genes and genomes. This sequence information has become invaluable to protein researchers, fuelling advances in computational methods for structure and function prediction,[89,119,362] analysis of protein family evolution,[26,92,192,361] and protein design.[25,185] Sequence databases are improving with regards to annotations[5,34] and coverage of protein domain space.[67,286] Interpro announced in 2019 that its annotation coverage had further increased to 80.9% of the ∼125,000,000 sequences in UniProtKB.[214] In addition, structural data is growing through structural genomics initiatives,[97,140] further enabling large-scale homology modelling efforts.[168,244]

The accuracy of protein function prediction has improved over the years as a result of better methods, as well as increased experimentally-based annotations.[126,253] Many proteins predicted from genomes can now be at least partially annotated[166,214] through detected homology to existing proteins (e.g., via BLAST search) or through matches to domain databases such as CDD,[183] Pfam,[67] CATH,[286] and FIGFAMs.[210] CDD and Interpro,[115]

[6] HMMs also incorporate information about gap propensity.

in particular, combine domain and protein models from their own and other databases in order to have more comprehensive annotations. Focused, niche databases have also been created for specific functions (antibiotic resistance - Comprehensive Antibiotic Resistance Database;[125] CARD) and organisms (http://iant.toulouse.inra.fr/S.meliloti) which seek to collect exceptionally well-curated annotations. These predictions form the initial landscape of functional annotations in newly sequenced genomes, upon which further questions may be investigated.

One important and common question following functional annotation is how to pinpoint the most functionally novel and biologically interesting predictions. This task is challenging due to the scale at which function predictions are often made and also because of the complexities surrounding the definition of "function".[254] As a result, expert biological knowledge is needed to interpret predictions and identify those providing particularly novel or unexpected biological functionality.

1.1.4 Finding homologs in unexpected places

Homology search has been described as the single most powerful tool in bioinformatics and, for decades, has been the core strategy in protein annotation.[8] Beyond its utility in finding new members or relatives of existing families, homology search can also reveal profound functional novelty when a homolog is found in a novel/unexpected biological setting. This setting may be a new species or environment,[54,190,264,327] or an unexpected co-occurrence with other proteins/pathways.[40,324] The discovery of bacterial rhodopsins,[16,327] archaeal ammonia monooxygenases,[155,327] and, complete nitrification by *Nitrospira*,[40,324] are all examples of important biological phenomena predicted through sequence homology.

The discovery of complete nitrification[40,324] illustrates the power of detecting unexpected enzyme combinations. By identifying genes encoding ammonia monooxygenase and hydroxylamine dehydrogenase together in a single genome, two studies[40,324] were able to identify the microbial basis for the long-sought-after process of complete nitrification (oxidation of ammonia to nitrate, "comammox"). Undersampled phyla from the tree of life are a likely hotspot for functional novelty of this kind as their genomes have been less explored. Indeed, analyses of hundreds of new microbial "dark matter" genomes obtained by single-cell genome sequencing have revealed novel and unexpected metabolic features such as archaeal sigma factors previously considered exclusive to bacteria.[264] Ultimately, even if molecular function is completely conserved in newly detected homologs, finding homologs in unexpected biological settings can reveal novelty at the pathway to organismal to ecological level.[40,264,324,327]

1.1.5 Problems with homology-based annotation

Homology-based annotation techniques can be used to find protein novelty but there remain serious pitfalls with these methods. Sequence similarity is not a guarantee of a full overlap of function. As proteins diverge, their function at lower and higher levels can change (e.g. a shift in binding affinity and/or substrate leading to a change in a protein's cellular role). Large protein families have subdivisions within them that can be regulated in entirely different ways and have different functions.[277,290] While proteins evolve at different rates, Tian and Skolnick[311] found that a sequence identity of 40% was enough to transfer the first three levels[7] of an Enzyme Classification (E.C.) number between sequences but for a full enzyme classification per-family thresholds were needed for accuracy.[160,277] To combat the problem of overannotation[8], databases like Pfam and CARD[125] have implemented model-specific thresholds in order to provide guidelines for higher accuracy. Without careful attention to the way that certain residues, insertions and deletions can alter function, misclassification is possible. A recent look at the DmdA family of peptidases,[92] prone to paralogous divergence within its family phylogeny, found overannotation of the protein using automated methods. A more accurate model was constructed by incorporating environmental data to fill in underepresented taxa, identifying sequences directly annotated by experimental data, and refining the model with phylogenetic analysis.[92] Once a protein is given an incorrect functional association, that error can propagate throughout a database. For example, during the discovery of "comammox" organisms, ammonia monooxygenase subunit A sequences were found in the NCBI nr database misclassified as methane monooxygenases.[324] Annotation errors, either due to sequence divergence or incorrect database entry, are the reason that clear history from annotation to experimental data is so vital. This allows researchers to trace information back to its source and assess the validity of an annotation.

Another shortcoming is that homology-based annotation methods are not able to functionally annotate all query sequences. Experimental characterization of proteins is an expensive, time-consuming task and, as an example, there are still 23% (4155) of the domain families in Pfam v33.1 that are called domains of unknown function (DUFs). There are also still genomic and metagenomic CDSs without database coverage. Some of these may be pseudogenes or a result of inaccurate or fragmented assemblies. This problem is enhanced in metagenomes where many predicted coding sequences are incomplete. High community complexity combined with low coverage can seriously impact the annotation

[7] The levels of E.C. numbers are denoted by digits.

[8] Overannotation refers to assigning a "full" function to a new protein where it may only be loosely related in function to its sequence match (i.e. should be found within a protein superfamily like the enolases but not in the specific subgroup in which it was placed).

process as short contigs lead to short CDSs. Sequencing technology, final assembly quality, and the coding sequence prediction software can also affect the length and accuracy of the CDSs.[245,254] Shorter sequences are harder to annotate due to lower possible alignment scores and poor database coverage. Viral sequences are also extremely underrepresented in current databases, with viromes having some of the worst annotation coverage.[4,44,245] The majority of taxonomically classified genomes in NCBI Genome are from Proteobacteria and Firmicutes,[232] with a study in 2019 finding them overrepresented in 16S rRNA databases compared to the estimated taxonomic diversity in other phyla.[182] Lack of taxonomic representation and protein characterization for unculturable or hard-to-culture organisms creates limits for homology-based functional annotation. Due to the naive aspects of homology-based functional annotation, and its inability to find accurate functional annotations for underrepresented protein families and organsisms, alternate methods for extracting functional information are an increasingly popular option.

1.2 Alternative approaches for analyzing and inferring protein function

1.2.1 Detecting functional shifts in sequences

As newly identified homologs may have diverged in function with respect to their reference, finding functional shifts in sequences or families (for example through detection of site-specific changes in evolutionary rate or amino acid preference[297]) is another way to uncover function. Several studies have applied the evolutionary trace (ET) method[172] to identify conserved and likely functional sites that differ between protein subfamilies.[134,266,301] Applications of these methods to families of G protein-coupled receptors have uncovered specificity-determining residues (SDRs) that differentiate substrate affinity and specificity.[134,266,301] These studies also highlight the important role of changes to allosteric pathways in shaping the evolution of specificity.

Analyses of functional diversification have also been expanded to entire protein superfamilies.[26,79,111,200] An effective approach has been to map structural and functional properties onto large-scale sequence similarity networks of enzyme superfamilies, thus revealing broad-scale differentiation of substrate specificity and how it correlates with sequence and structural features.[26] Such approaches have revealed functional differentiation in ligases,[111] cytosolic glutathione transferases,[200] dipeptide epimerases,[184] and diverse trans-polyprenyl transferases.[332] In a recent study, Furnham et al.[79] examined changes in

enzymatic function within 379 protein domain superfamilies, revealing how both subtle and large-scale changes in enzymatic machinery can lead to functional changes in chemistry and substrate specificity.

Building on past approaches,[1] databases have attempted to subdivide known protein families into functionally distinct subfamilies. The FunFHMMer method has subdivided 6,119 CATH superfamilies into 67,598 subfamilies (FunFams) with increased functional coherency.[43,286] Similarly, the Selectome database has predicted positive selection across thousands of vertebrate protein phylogenies, facilitating large-scale exploration of adaptive evolution.[220]

While the above approaches tend to examine functional shifts over macroevolutionary time scales, others are better suited to detecting microevolutionary positive selection on single nucleotide polymorphism (SNP) and indel variants.[330] Genome-wide scans for positive selection for example have revealed a wide array of adaptive events in recent microbial[233] and human evolution.[99] Methods such as SIFT/Provean[151] have been used to estimate the functional impact of protein variants, and genome-wide screens using these methods have uncovered bacterial protein adaptations for increased pathogenicity and antibiotic resistance.[297,315]

1.2.2 Remote homology detection

While many proteins can at least find a match in popular annotation databases, anywhere from 2-81% of genomes and up to 86% of metagenomes are frequently left without any assigned functional information, lacking detectable homology to proteins of known function (Figure 1.1). These are the most challenging, yet biologically intriguing, targets for function prediction. These sequences include so-called ORFans,[76,284] DUFs,[123,221] and protein "dark matter".[240] Many apparent ORFan proteins have been predicted to be highly divergent homologs of known structural families.[90,123] For these cases, remote evolutionary relationships to known families can potentially be predicted using methods such as HHpred,[261] Protein Homology/Analogy Recognition Engine[138] (PHYRE2), and Iterative Threading Assembly Refinement[345] (I-TASSER).

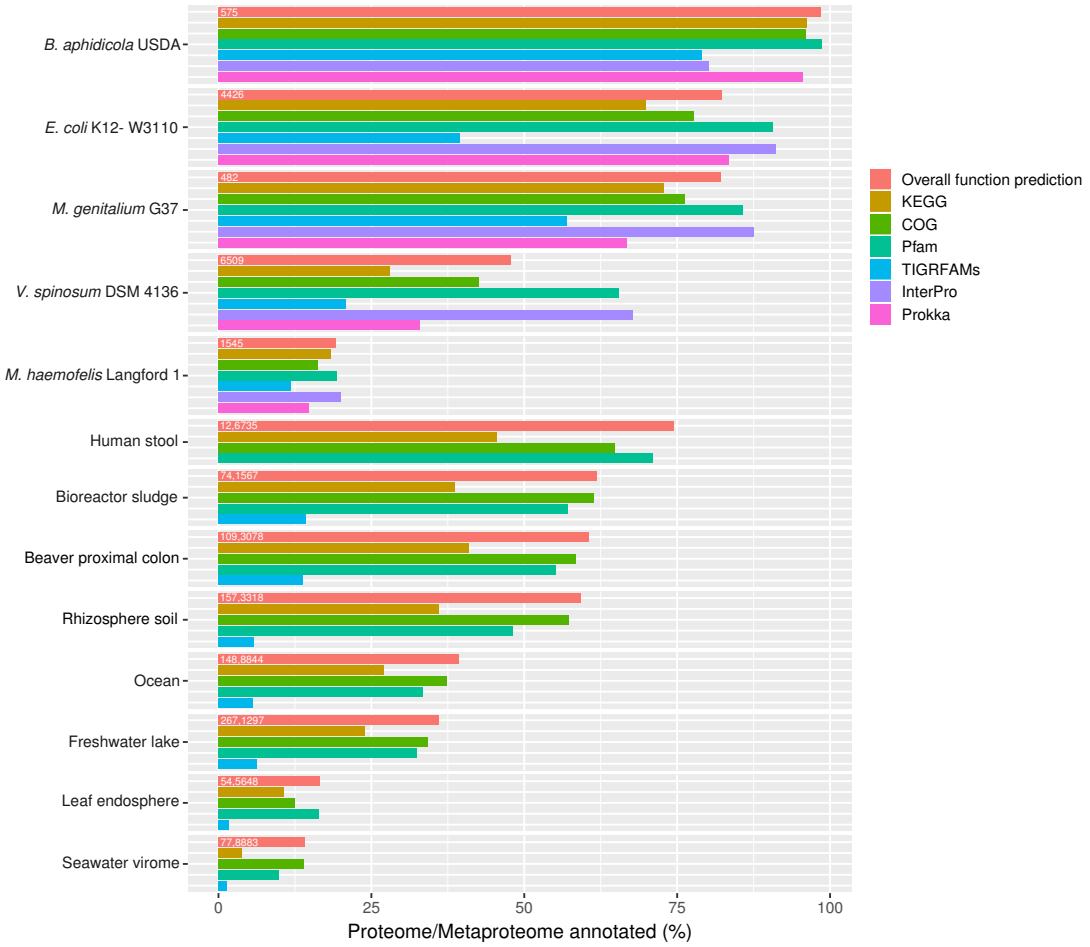

Figure 1.1: Annotation coverage of genomes and metagenomes from JGI's[201] Integrated Microbial Genomes and Metagenomes (IMG/M) database. GOLD (Genomes Online Database) analysis project IDs are: Ga0244168, Ga0334891, Ga0334942, Ga0376466, Ga0373948, Ga0373643, Ga0335017, Ga0325419, and Ga0326737. For the metagenomes, all but the human stool sample were sequenced with the Illumina Novaseq (Illumina HiSeq 2500 was used for the human sample) and annotated with the IMG Annotation Pipeline v.5.0.1-3 (v.4.16.4 used for the human sample). The number of annotated coding sequences were divided by the number of predicted coding sequences provided to get a fraction. *Figure based on Prakash and Taylor, 2012.*[245]

HHpred[261] uses profile HMM-HMM searches to sensitively compare the conservation profiles between families, ideally discovering shared functional or structural signatures that

suggest an evolutionary link or convergent evolution. This tool is used for finding templates in threading methods like PHYRE2 and I-TASSER[9], which subsequently model the query sequence onto the chosen templates and further refine the structures by incorporating secondary structure information and using ab initio modelling on unaligned regions.[138,346] Resulting models can then be compared to functionally annotated proteins to gain functional insights, like enzyme superfamily, ontology terms, and possible ligand binding sites, from their distantly-similar match.

Remote homology detection has been successful in elucidating the structures and possible functions of apparent protein dark matter.[123,221,240] Perdigão et al.[240] surveyed the "dark proteome" (segments of proteins that lack detectable similarity to known structures). Surprisingly, dark matter made up almost half of the eukaryotic proteome, and again dark proteins were found to be associated with certain functions such as secretion, disulfide-bonding and proteolytic cleavage.

Ultimately, protein dark matter is a particularly intriguing target for future characterization efforts. Recent studies suggest that the proportion of novel folds in newly discovered domain families may be as high as 36%.[265] Identifying which DUFs are most likely to provide new folds is thus an important goal. Developments in de novo structure prediction (e.g., covariation approaches[112,197]) have been suggested as a promising strategy to complement experimental approaches and accelerate the identification of new structures.

If structural data is available for a protein or can be modelled (either through threading or ab initio methods), attempts can be made to predict biochemical function *directly* from structure. Structure-based function prediction is therefore a potential solution to uncovering function for DUFs solved by structural genomics initiatives.[70,360] In addition, these methods may uncover new functionality in structures with existing annotations and predict new protein interactions.[355]

1.2.3 Motifs and domain architectures

In the absence of global sequence homology, motifs and domains can be used to identify a protein's functional or structural pieces. These techniques can also be used to try and support, expand, or further refine the functions determined from remote homology (or other methods). Linear motifs[10], examples of which include the PxxP (SH3 domain

[9] I-TASSER uses LOMETS[339] to find templates which incorporates predictions from nine different tools, including HHpred.

[10] Also called short linear motifs (SLiMs), these are stretches of 3-15 adjacent amino acids that are a mix of high and loosely conserved residues.[96]

binding motif) and PxY (WW domain binding motif), tend to be embedded in disordered regions and mediate protein–protein interactions.[50,226,314] "Binding motifs" like these (together with posttranslational modification sites) are widespread in proteomes, with repositories in the PROSITE[285] and Eukaryotic Linear Motif[150] (ELM) databases, and yet are largely understudied.[314] Because they are short and have a propensity to arise independently they can also capture convergent evolution of function in unrelated proteins, but are statistically difficult to predict without additional (e.g. structural) information. One example application of motif predictions in function discovery is the identification of host-like proteins in pathogenic organisms, or so-called "mimicry".[35,57,329] This feature of many pathogen proteins has been exploited by computational methods to predict novel virulence factors.[57,241] For example, Doxey and McConkey[57] predicted widespread mimicry of human extracellular matrix proteins across a diverse range of human pathogenic bacteria, based on detected similarities between motifs in collagen and leucine-rich repeat proteins. The predicted mimics represent new candidate virulence factors.

Multi-domain architectures, or combinations of domains across a protein, can reveal information about the protein itself, but also about which domains cooperate together. Domains have been duplicated and recombined extensively throughout protein evolution.[163] Experimental and computational approaches have shown that domain shuffling can significantly impact the organization of signalling networks,[239,364] and new domains may also alter protein function and enzymatic activity.[63,163] This was demonstrated in a recent study of the domains found in the flagellin hypervariable region of bacteria. A metallopeptidase insertion between the two flagellin domains indicates an enzymatic role for flagella, making the flagella the largest known proteolytic complex.[63] Identification of novel domain combinations may signify new functionality, however, predicting the functional consequences of domain combinations is a challenging and important goal.[253]

1.2.4 Genomic context and inferred functional associations

In contrast to methods that identify function directly within protein sequence and structure, function can sometimes be inferred using associated information. Functional associations may be inferred using a wide variety of techniques including detected protein enrichment in certain species or environments,[57] genotype-phenotype correlations,[62,173] networks analysis,[331] and analysis of neighbouring gene or domain functions (genomic or domain context).[114,288] For example, a comparison of genomes with and without a certain phenotype can reveal genes associated with the phenotype in question (genotype-phenotype correlations), whereas prokaryotic genome organization (i.e. genes with similar functions or

genes for similar pathways end up near each other[42,81,145,195,271,344])[11] enables high-level function predictions based on proximity (genomic context). In addition, databases such as STRING provide predicted interactions based on gene fusions, gene neighbourhoods, coexpression, and gene co-occurrence.[303] Since these methods use contextual information, they can be used to infer function of completely uncharacterized sequences. Indeed, these methods have been instrumental in historical examples of function discovery (e.g., the initial prediction of the CRISPR/Cas system[187,188]).

One area in which these methods have played an important role is metagenomic enzyme discovery.[320] The human gut microbiome has become a major target for finding novel Carbohydrate-Active enZYmes (CAZymes) due to its considerable diversity of uncharacterized glycan-degrading activities associated mostly with the phylum Bacteroidetes.[135] To discover novel enzymes with important roles in the human gut, studies have searched for proteins with increased relative abundance in gut metagenomes,[68,225] high sequence novelty[282] and genomic context suggestive of carbohydrate metabolism.[308] Taking advantage of the tendency for CAZymes to be genomically clustered in operons, Terrapon et al.[308] used both genomic and domain context to automate the prediction of polysaccharide utilization loci (PUL) in Bacteroidetes genomes (see Figure 1.2 for an example). Predicted PULs facilitate hypothesis generation and experimental discovery of new metabolic activities.[110,158] For example, Martens et al.[198] identified a PUL in *Bacteroides ovatus* that was transcriptionally upregulated by galactoxyloglucan, which led to the discovery of a novel xyloglucan metabolism locus found ubiquitously in human gut metagenomes.[158] The same study also illustrates how context-based predictions can sometimes be misleading. The study's characterization of the "Bacteroidetes-Associated Carbohydrate-binding Often N-terminal (BACON)" domain, a domain initially predicted to have carbohydrate-binding activity based on its recurring association with CAZyme families,[206] found no evidence of carbohydrate-binding.[158] Instead, the domain played a role in membrane anchoring and positioning of its partner catalytic domain.

[11] This phenomenon is exemplified by operons, a single promoter-controlled gene cluster.

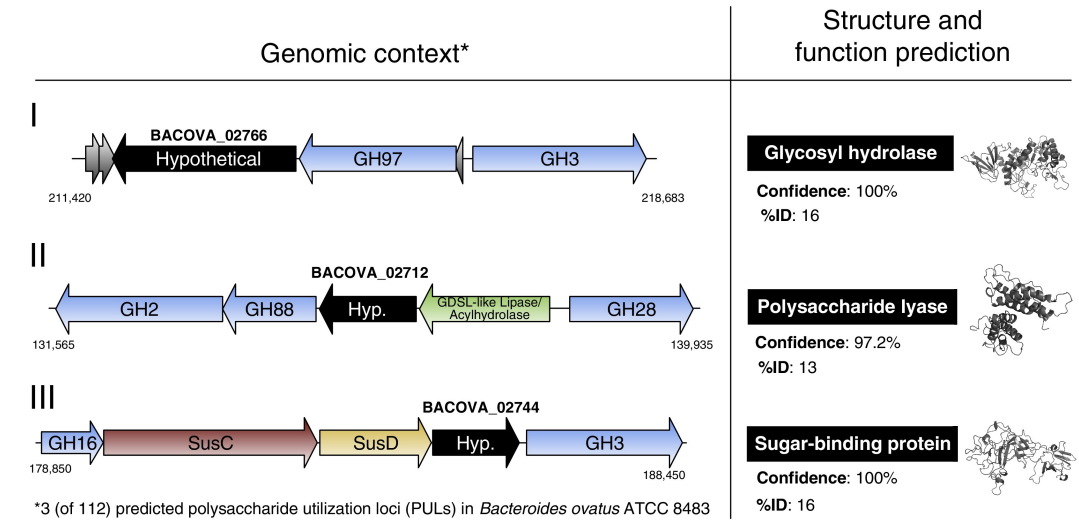

Figure 1.2: Genomic data mining for novel CAZyme activities by integration of genomic context with structural modelling. The genomes of human gut *Bacteroides* species are an excellent resource of novel carbohydrate-active enzymes and associated proteins. For targeted discovery of new carbohydrate metabolic functions, one approach is to first identify general genomic regions called polysaccharide utilization loci (PULs) that contain a high density of predicted carbohydrate-active enzymes (CAZymes).[308] Second, based on gene neighbourhood, possible CAZyme activity can be inferred for hypothetical genes found within these loci. To provide added evidence of CAZyme activity, predicted protein sequences can be structurally modelled and analyzed for similarity to CAZyme structures. The three structure predictions shown above were made using PHYRE.[139] Proteins satisfying all conditions are potentially new CAZymes with novel specificities or activities.

1.3 Applications of integrative approaches to function prediction

An important, long-standing theme in computational function prediction is the gain in predictive accuracy from data and methodological integration.[156,195,362] Greater confidence in function prediction can also be gained from multiple lines of evidence as demonstrated in Figure 1.2, which illustrates the use of genomic context and remote homology detection to predict novel CAZymes. Indeed, the best performing method in the latest Critical Assess-

ment of Functional Annotation (CAFA) challenge[362] combined five component classifiers including information about sequence properties (e.g. molecular weight, the isoelectric point, and a measure of instability), sequence alignment scores, GO term frequency, and domain and motif matches.[352] Increased coverage of protein structure space has also lead to an increase in data integration, as this enables a greater fraction of predicted protein sequences to be homology modelled and analyzed using structural bioinformatics.[243]

Studies have seen a considerable integration of data and methods including combinations of homology modelling, docking, sequence similarity networks, phylogenetics, genomic context, and metabolic analysis.[14,70,184,332,360,361] For example, the combination of homology modelling and ligand-docking has allowed sequences that lack available structural data to be virtually screened for novel activity. This approach successfully predicted pterin deaminase activity in a protein of unknown function.[70] Bastard et al.[14] combined homology modelling, docking, phylogenetic comparisons, genomic context, and metabolic analysis with high-throughput enzymatic screening, and uncovered 14 new enzymatic activities in the DUF849 family (now renamed as Pfam family "BKACE"). This impressive study reveals DUFs as an important source of new enzymatic functions and also highlights the tremendous functional diversity to be found within single enzyme families. Finally, Zhao et al.[360] combined structure-based approaches with genomic context information to predict the substrate specificity of several enzymes in a bacterial gene cluster. The approach not only predicted the function of an uncharacterized protein in the gene cluster, but also identified its role within a specific catabolic pathway by integrating information from the surrounding gene neighbourhood.

1.4 Thesis outline

As described thus far, there are a wide range of computational approaches available for assigning and analyzing protein function from sequence information. However, with the explosion of new sequence data from increased genome and metagenome sequencing, there are a number of important questions concerning functional annotation which form the basis of my thesis.

- How can new functions or biological insights be gained both by homology-based or alternative methods of functional annotation?
- How effective are standard methods of homology-based protein annotation at annotating entire genomes or metagenomes?

- And what trends (biological or otherwise) are associated with the sequences that homology-based methods fail to annotate?

In this thesis, I explore these questions using a combination of focused bioinformatic studies of original sequencing datasets, as well as large-scale analyses of existing protein databases. In Chapter 2, I begin by exploring the use of homology-based methods to analyze newly-generated genomic and metagenomic datasets, with the goal of detecting specific protein families/functions of interest including cellulases (2.1), antibiotic resistance proteins (2.2) and virulence factors (2.3, Table 1.1). I also perform a data-driven study (2.3) to detect global functional differences that occur in a time course of decomposing fish, which provides a rich resource of new genomic and protein sequence information.

These studies indicate that homology-based annotation is only able to capture a fraction of the total protein diversity in each dataset, which motivates my work in Chapter 3. In Chapter 3, I apply standard homology-based methods to functionally annotate over 27,000 bacterial genomes in the Genome Taxonomy Database, measure their "annotation completeness", and identify major factors influencing annotation completeness.

Then in Chapters 4 and 5, I explore additional methods for inferring the function of uncharacterized proteins. In Chapter 4, I analyze all ∼17,000 protein domain families in Pfam, including domain families of unknown function (DUFs), and develop strategies for detecting statistical associations between these families and other biological information. In Chapter 5, I turn my attention to uncharacterized proteins that are not even accounted for in current databases. By analyzing human gut, marine, and soil metagenomes, I construct a dataset of these "ORFan" proteins, that lack any detectable homology to existing reference databases. I then apply powerful remote homology-based approaches to infer their molecular functions, profiling the "dark" fraction that homology-based annotation leaves behind.

Table 1.1: Summary of thesis chapters.

Group analyzed	Datatype	Problem	Approach used
Chapter 2			
2.1 Novel *Streptomyces* strains	genomes	find pathways of interest and high-confidence cellulase predictions	*homology-based* metabolic pathway database and multi-method comparison
2.2 Research farm wastewater	metagenome	uncover antibiotic-resistance genes	*homology-based* focused antibiotic resistance database
2.3 Time-series of a rainbow darter necrobiome	metagenomes and metagenome-assembled genomes	profile the necrobiome and explore any potential pathogens	*homology-based* metabolic pathway database focused virulence factor database
Chapter 3			
Bacterial genomes across the tree of life	genomes	compare annotation completeness throughout bacteria	*homology-based* protein and domain family databases
Chapter 4			
Uncharacterized Pfam protein domain families	domain families	provide biological context to prioritize families for characterization	*alternate* phenotype associations with environmentally-classified metagenomes and with taxa and pathogen-classified genomes
Chapter 5			
ORFan sequences (lacking detectable homology to current databases)	metagenomes	profile protein dark matter	*alternate* remote homology, genomic context, and motif analysis

Chapter 2

Case studies of homology-based genome and metagenome analysis

First, I aimed to examine how we can use standard homology-based annotation methods to find novel proteins. There are numerous databases and tools for performing sequence - sequence or sequence - model annotation methods. Using focused databases for certain protein families or combining multiple databases and methods to lend more confidence to predictions are strategies explored here on newly sequenced genomes, metagenomes, and metagenome-assembled genomes (MAGs). Each of the following case studies is an example of either a targeted or exploratory search for novel proteins of interest, with the annotation strategies tailored for the individual circumstances.

In the first case study, collaborators isolated two *Streptomyces* strains from rhizosphere soil that were able to grow on starch, xylan and cellulose. As cellulases can be used in the production of environmentally friendly biofuels, a key goal of this study was to learn more about the isolated organisms and about the cellulases they may be able to produce. Phylogenetic analysis with other *Streptomyces* genomes and a comparison of the genomic sequence similarity between closely matched strains allowed placement of the newly sequenced organisms within *Streptomyces*. A thorough look at a metabolic database annotation of the organisms revealed inferences about their metabolic potential, and combining the annotations from four different methods/databases allowed a moderate number of high-confidence cellulase predictions.

A newly sequenced wastewater metagenome was analyzed in the second case-study for antibiotic resistance. This sample was of importance to a group of microbiologists working in association with a research farm in South Africa. Other farms have been

shown to be a source of antibiotic resistance, in part due to use of antibiotics for livestock. Of interest were antibiotic resistance genes present in the microbial community, as well as any potentially pathogenic genera present that could contain these genes. A niche antibiotic resistance database was used to target the resistome, as its well-curated models and stringent match threshold cut-offs result in more accurate analyses than less-focused databases. Genes associated with tetracycline and streptomycin resistance were the most frequent, with *Thauera*, a genera already associated with wastewater and sludge, dominating the community.

Little is known about the microbial decomposition of aquatic vertebrates from a functional and environmental context. In the final case study, a common North American fish (rainbow darter) was analyzed for temporal changes in its "necrobiome". By combining 16S rRNA gene and shotgun metagenomic sequence data from four time points, I studied the progression of decomposers from both taxonomic and functional perspectives. Metagenomic analysis of metabolic pathway annotations revealed significant changes throughout decomposition in degradation pathways for amino acids, carbohydrates/glycans, and other compounds. Binning of contigs confirmed a predominance of *Aeromonas* in the necrobiome, including novel strains related to the human and fish pathogen *Aeromonas veronii*. A virulence factor annotation database revealed that the *Aeromonas* bins encoded known hemolysin toxins (e.g., aerolysin) which were particularly abundant early in the process, potentially contributing to host cell lysis during decomposition.

2.1 Draft genome sequences of two novel cellulolytic *Streptomyces* strains isolated from South African rhizosphere soil

Material in this section has been published as part of Adegboye et al. (2018).[2] The published manuscript is available here:

> M. F. Adegboye, B. Lobb, O. O. Babalola, A. C. Doxey, and K. Ma. Draft genome sequences of two novel cellulolytic *Streptomyces* strains isolated from South African rhizosphere soil. *Genome Announcements*, 6(26):e00632-18. 2018.[2] https://doi.org/10.1128/genomeA.00632-18

2.1.1 Introduction

Streptomyces species are known for their diverse metabolic potential, wide range of antibiotic biosynthesis capabilities, and their ability to degrade unique compounds, such as lignocellulose, keratin, xylan, pectin, cellulose, lignin, chitin, and styrene.[136,167,259,272] They also produce various hydrolytic enzymes, such as amylase, lipase, esterase, gelatinase, xylanase, and cellulases.

The cellulase family is comprised of three different types of enzymes: endoglucanase or endo-1,4-β-D-glucanase [EC 3.2.1.4] which breaks down the internal β-1,4 glycosidic bonds, exoglucanase or cellobiohydrolases [EC 3.2.1.91] which release two (cellobiose)[86] or four (cellotetraose)[365] saccharide units from the ends, and β-glucosidase [EC 3.2.1 .21] which hydrolyzes the short oligosaccharides produced during cellulose degradation into glucose.[289] Cellulases can be used for the hydrolysis of lignocellulose to fermentable sugars which can be used as feedstock for the production of biofuels that have been proven to be environmentally friendly, help reduce dependence on fossil fuel, and serve as an alternative for declining petroleum reservoirs.[12] Novel cellulases with properties that will improve industrial processes like higher catalytic efficiency or tolerance to various temperature/pH levels are sought after.[58] As many industrial cellulases have been from *Trichoderma* spp. and *Aspergillus* spp.,[359] the varied *Streptomyces* genus represents an opportunity for enzyme discovery. The isolation of environmental *Streptomyces* species capable of lignocellulose degradation is therefore of considerable interest.

2.1.2 Methods

Sample preparation, sequencing and assembly

Initial isolation, experimental characterization, and sample preparation done by Dr. Mobolaji Adegboye, Dr. Olubukola Babalola, and Dr. Kesen Ma.

Streptomyces strains NWU339 and NWU49 were isolated from rhizosphere soil as described in Adegboye et al., 2013[215] and subsequently cultured using starch casein agar. Genomic DNA (50 ng) was extracted using the Wizard genomic DNA purification kit from the Promega Corporation. Sequencing libraries were prepared using the Nextera DNA sample preparation kit (Illumina). Sequencing was performed on an Illumina HiSeq platform, and genome assembly was performed using NGen v14 with Q25 trimming[1]. The raw reads were deposited in the Sequence Read Archive (SRA) under the accession number

[1]Molecular Research LP (USA) provided sequencing and assembly services.

SRP148117. The assemblies for *Streptomyces* sp. NWU339 and *Streptomyces viridosporus* NWU49 were deposited in GenBank under the accession numbers QFRK00000000 and QFXB00000000, respectively.

Annotation and phylogenetic analysis

Gene finding and genome annotation were then performed using Prokka v1.12 (databases downloaded 26 January 2018). Parsing Prokka's output to determine annotation coverage was performed as described in Methods 3.2). MetAnnotate[242] with its default set of taxonomic markers on the sequence similarity "fast mode" was used to confirm phylogenetic placement in March 2018. Closely related *Streptomyces* genomes from NCBI (based on BLASTN results from the 16S rRNA sequences) were used for the phylogenetic analysis. The tree was made using RAxML v8.2.12[295] with the LG likelihood model made from concatenated single-copy core protein sequences detected with Anvi'o[69] (Campbell et al. set[30]). Average nucleotide identity (ANI) was calculated with calcANI.pl v1 (available at `https://github.com/Computational-conSequences/SequenceTools`) using the FastANI v1.3[120] option. RAST (SEED) annotations were provided by Molecular Research LP (USA) using default settings. Any function identified only as "hypothetical protein" was removed to determine annotation coverage. For metabolic analysis, KEGG annotations were identified with GhostKOALA[133] using the "prokaryotes" setting in March 2018. CAZyme annotations were obtained via the dbCAN meta server[349] with default settings in March 2018. TIGRFAM annotations were determined with the TIGRFAM database v15.0[105] using a threshold of 1×10^{-3} with hmmscan from HMMER v3.1b1. Pfam annotations were derived from Pfam v27.0[75] and applied with HMMER v3.1b1 and Pfamscan (at `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/`). COG annotations were performed by Anvi'o v5.2 with the COG 2014 database[306] files sourced from `ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data/`.

2.1.3 Results

Two novel *Streptomyces* strains (NWU339 and NWU49) were isolated from the rhizosphere of maize in North West Province, South Africa, as described previously in Adegboye et al., 2013.[215] Both strains were capable of growing on polymeric carbohydrate substrates, such as starch, xylan, and cellulose. Sequencing and assembly of NWU339 produced 169 contigs, resulting in a draft genome of 9,425,309 bp, with a GC content of 70.8%. Whereas, the assembly of NWU49 produced 97 contigs, resulting in a draft genome of 8,905,076 bp, with

a GC content of 72.3%. The genomes of NWU339 and NWU49 encode 8,776 and 8,021 protein-coding sequences, 8 and 7 rRNA genes, and 88 and 100 tRNAs, respectively.

A 16S rRNA phylogenetic tree had been constructed previously by Adegboye et al., 2013.[215] Phylogenetic analysis of taxonomic marker genes using MetAnnotate[242] confirmed NWU339 to be a novel *Streptomyces* strain with 97.0% 16S rRNA identity to *Streptomyces poonensis* NRRL B-2319 (809 bp alignment with NR_043852.1). Whereas, NWU49 possessed 98.4% 16S rRNA identity to *Streptomyces viridosporus* NBRC15414[2] (1167 bp alignment with NR_112460.1). A phylogenetic tree using concatenated single-copy genes was constructed for better resolution of the organisms' placement within *Streptomyces* (Figure 2.1). This tree shows NWU49 falling within the well-supported *Streptomyces viridosporus* clade, and the overall topology matches the Genome Taxonomy Database[235] tree topology for similar organisms (including NWU49 and NWU339; visible with the AnnoTree[207] web interface http://annotree.uwaterloo.ca/app/#/?qtype=tax&qstring=67581). A comparison of NWU49 to the genome of *Streptomyces viridosporus* ATCC 14672[3] using an average nucleotide identity (ANI) calculator resulted in an ANI of 99.0% (Table 2.1), indicating that they are likely the same species (>95%[120]).

[2]Since this article was published, *Streptomyces ghanaensis* NBRC15414 and ATCC 14672 have been changed to *Streptomyces viridosporus* NBRC15414 and ATCC 14672 in NCBI.

[3]See footnote 2.

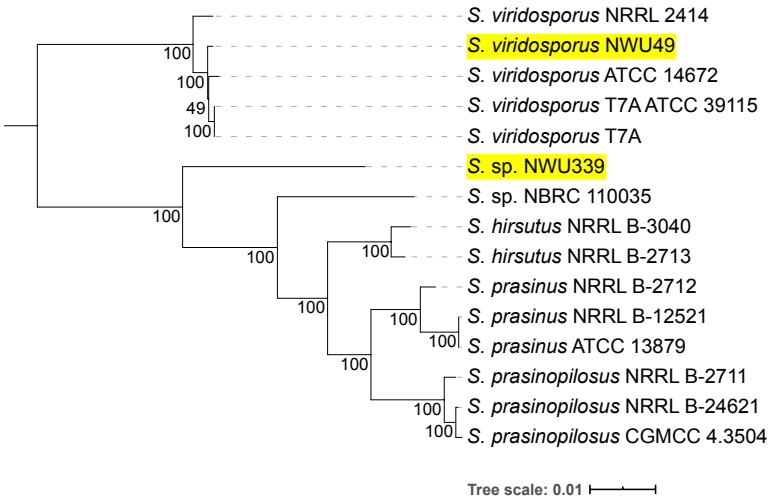

Figure 2.1: RAxML tree with the LG likelihood model made from concatenated single-copy core protein sequences detected with Anvi'o (Campbell et al. set[30]). The tree was outgrouped with *Streptomyces malaysiense*. *Streptomyces* genomes with similar 16S rRNA sequences were used for this tree and sourced from NCBI Genome. This tree was visualized with iTOL.[165]

In order to investigate the strains' metabolic potential, KEGG (a database with metabolic pathway information) annotations were analyzed. Matches to enzymes forming a metabolic pathway for benzoate degradation were detected in both organisms (including [EC 1.14.12.10] and [EC 1.3.1.25]). NWU339 also had matches to toluene and xylene degradation enzymes. NWU49 contained pathways for the degradation of sphingosine and trans-cinnamate and for the biosynthesis of polyamines and trehalose. Complete predicted pathways in NWU49 also included the biosynthesis of a nine-membered core molecule for enediyne, an anticancer metabolite.

While KEGG predicted cellulases in NWU339 and NWU49, Prokka was also used to provide more resolution for the cellulase family predictions. According to Prokka annotations, NWU339 contained 15 putative cellulase-related genes, including 5 predicted subtypes of endoglucanases, 3 subtypes of exoglucanases, and 4 subtypes of beta-glucosidases. NWU49 contained 18 putative cellulase-related genes, including 8 predicted subtypes of endoglucanases, 3 subtypes of exoglucanases, and 4 subtypes of beta-glucosidases. Comparing the Prokka, KEGG and additional RAST and dbCAN (with Carbohydrate-

Active enZyme or CAZy families) annotations revealed differences in which sequences were detected as cellulases. Only 11 of the initial Prokka-derived cellulases were consistently identified as cellulases with the other three methods (Table 1). Some of predicted cellulases were alternatively identified as other enzymes (such as 3-dehydroshikimate dehydratase, chitinase, or beta-mannosidase), had annotations for homologs that have not been confirmed to be cellulases, had annotations that are below the respective method's standard threshold, or had no annotation available.

Table 2.1: Average nucleotide identities for two novel *Streptomyces* strains. G1 is either *Streptomyces* sp. NWU339 or *Streptomyces viridosporus* NWU49 and G2 is the other compared *Streptomyces* genome. The ANI column is the average of the ANI from the two bidirectional ANI comparions in the G1-G2 and G2-G1 columns. Every member of the phylogenetic tree in Figure 2.1 is featured here.

	G1-G2 (%)	**G2-G1 (%)**	**ANI (%)**
NWU49			
S. viridosporus ATCC 14672	98.91	99.11	99.01
S. viridosporus T7A	99.15	99.24	99.19
S. viridosporus T7A ATCC 39115	99.22	99.22	99.17
S. viridosporus NRRL 2414	96.72	97.04	96.88
NWU339			
S. sp. NBRC 110035	88.86	89.15	89.00
S. hirsutus NRRL B-2713	89.44	89.74	89.59
S. hirsutus NRRL B-3040	89.23	89.67	89.45
S. prasinus ATCC 13879	89.13	89.69	89.41
S. prasinus NRRL B-12521	88.82	89.33	89.07
S. prasinus NRRL B-2712	88.89	89.43	89.16
S. prasinopilosus CGMCC 4.3504	89.46	89.46	89.57
S. prasinopilosus NRRL B-24621	88.99	88.99	89.25
S. prasinopilosus NRRL B-2711	89.40	89.40	89.07

To get a sense of how these two genomes fare overall during annotation, I applied some additional popular annotation methods: COG,[306] TIGRFAM,[105] and Pfam.[75] The annotation coverage varied dramatically depending on the method, ranging anywhere from 31 - 84%. Large sequence or domain databases that include uncharacterized proteins like COG, Pfam, and FIGfams (RAST) "annotated" more of the sequences. Even the method with the highest coverage, RAST, left an average of 18% of the predicted coding sequences unannotated across the two strains.

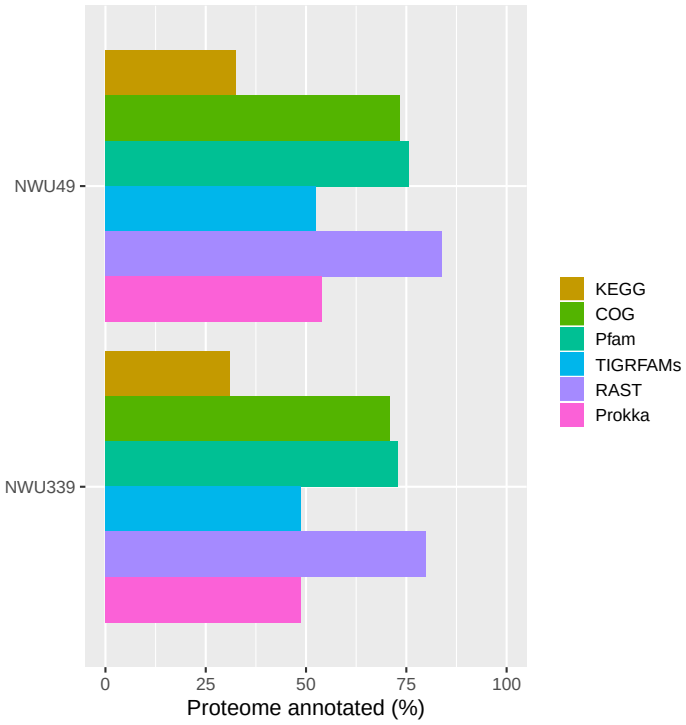

Figure 2.2: Annotation coverage of two novel *Streptomyces* strains: *Streptomyces* sp. NWU339 and *Streptomyces viridosporus* NWU49. The fraction of proteome annotated is determined by how many of the predicted protein coding sequences have any annotations using the respective methods.

2.1.4 Discussion

The newly sequenced *Streptomyces* strains NWU339 and NWU49 expand our knowledge of the *Streptomyces* genus and provide additional sources of these industrially and medically-relevant organisms. In order to place these genomes within currently known *Streptomyces* clades, 16S rRNA sequencing and other taxonomic marker genes were considered. Although the 16S rRNA identity of NWU49 to *Streptomyces poonensis* NRRL B-2319 is just under what is conservatively considered to be a species boundary ($\geq$98.7%[267,294]), based on placement within the *Streptomyces viridosporus* clade and >95% ANI with *Streptomyces viridosporus* ATCC 14672, NWU49 should be considered a new strain of *Streptomyces viridosporus*. NWU339, however, appears to not fall within any of the currently described species for *Streptomyces*, ending up just outside the *S. hirsutus*, *S. prasinus*, and *S. prasinopilosus*

clade.

KEGG was used to look at potential metabolic pathways to find interesting putative metabolic activity. A few *Steptomyces* have been shown to degrade aromatic compounds[13,15] including benzoate[234] and the presence of a benzoate degradation pathway in both strains could be of interest if experimentally corroborated. This could help reveal more about the metabolism of the environmentally-abundant aromatic compounds outside of the facultative anaerobes in which the processes have been most frequently studied.[84,337] Also of interest is the core enediyene biosynthesis pathway present in NWU49. Enediyene is used in cancer treatments[281] due to its high cytotoxicity. A *Streptomyces*-derived enediyene discovered in 2016[343] was the only "naturally" discovered one since 2005,[46] with actinobacteria singled-out in bioinformatics analyses[268,281] as having great potential as a source of new medically-relevant enediyene types.

In order to identify cellulases within these organsisms, the results of four different annotation methods were compared. A single, well-curated, trustworthy source of annotations can lead to high annotation accuracy but such sources do not exist for every organism and protein family. A comparison across different methods can provide confidence in the predictions and better annotation coverage,[98] although some manual assessment due to differences in naming is usually required. In this case, 11 coding sequences had congruent cellulase annotations. These sequences all have well-established cellulase homologs and are thus, most likely to have cellulose-degrading activity. Other predicted cellulases either had other enzymes predicted via different methods, less confident predictions (i.e. putative), or no annotations at all. These represent either false predictions, or diverse homologs of known cellulases that may have some cellulose-degrading ability or may have some other glycosidase activity. Sequences with CAZY glycoside hydrolase family matches and carbohydrate binding modules are probably divergent glycoside hydrolyases, while sequences without are riskier to devote further experimental resources to.

Even more annotation methods were run on these two genomes to get an overview of annotation coverage. Like other genomes seen in Figure 1.1, there was substantial variation in the number of predicted coding sequences each method was able to assign sequence or model matches to. This is in part because of how each database treats uncharacterized proteins or proteins with only partial information (e.g. a biological process they have been associated with). Some databases, like COG, include a large number of families with no or limited information apart from which taxa they are found in. Matches to families of uncharacterized proteins can inflate the annotation coverage. Another factor is that domain databases (like Pfam), due to domains being protein modules, often allow for more protein matches than when using databases of full-length proteins. The databases have different sizes as well as types (e.g. sequence versus model) which can affect how many

matches are found. Even with the most optimistic case, here at least 16% of the predicted coding sequences remain unannotated. But in spite of the unannotated sequences, these newly sequenced genomes are a source of proteins of interest, with putative benzoate and enediyene metabolic pathways, high confidence cellulases, and uncharacterized glycoside hydrolases to test.

2.2 Metagenomic sequencing of wastewater from a South African research farm

Material in this section has been published as part of Lobb et al. (2018).[176] The published manuscript is available here:

> B. Lobb, A. A. Adegoke, K. Ma, A. C. Doxey, and O. A. Aiyegoro. Metagenomic sequencing of wastewater from a South African research farm. *Microbiology Resource Announcements*, 7(16):e01323-18. 2018.[176] https://doi.org/10.1128/MRA.01323-18

2.2.1 Introduction

Antibiotics are used to promote growth and manage disease in livestock at the Agricultural Research Council–Animal Production in South Africa. However, the spread of antibiotic resistance is a pervasive concern. Waste from farm animals has been shown to spread antibiotic-resistant bacteria, sometimes due to selective pressure found in antibiotic-dosed livestock.[83,202,363] One of a farm's effluents, wastewater, is a documented reservoir of antibiotic resistance genes that could transfer to human pathogens.[229,341,358] Wastewater is also known to contain animal pathogens, some of which are opportunistic and can spread zoonoses.[10,222] Sequencing of the wastewater microbiome can help identify pathogenic species that might exist on the institute's farm and detect antibiotic resistance genes that may be active in these microbial communities. The goal of this study was to profile the antibiotic resistome at the Agricultural Research Council–Animal Production site, enabling them to make more informed decisions about antibiotic use moving forward.

2.2.2 Methods

Sample preparation, sequencing and assembly

Sample collection and preparation done by Dr. Anthony Adegoke and Dr. Olayinka Aiyegoro.

The metagenome was created from expended water taken from Agricultural Research Council–Animal Production (ARC-AP) in Irene, South Africa. A 1-liter composite sample was created by combining five 200-ml samples collected from different wastewater gutters in the pig facility. The composite sample was centrifuged at 3,500 rpm for 10 min at room temperature to separate the biomass and water. The water was filtered to trap microbes, and DNA was extracted from the pellet on the filter paper. The DNA extraction was done using the FastDNA Spin kit for water (MP Biomedicals, Solon, OH, USA) and the FastPrep apparatus, according to the instructions given by the manufacturer. The DNA was sequenced with the Illumina HiSeq platform and the Illumina HiSeq reagent v3[4]. The raw reads were deposited in the Sequence Read Archive (SRA) under the accession number SRP159184. The reads were trimmed with Sickle v1.33[129] and Trim Galore! v0.5.0[148] and then assembled using MEGAHIT v1.1.2,[169] resulting in 58,129 contigs longer than 1 kb. The assembly was then deposited at GenBank under the accession number QXGG00000000. Prodigal v2.6.3 with the -p meta option[116] was used next, facilitating the prediction of 612,922 coding sequences.

Taxonomic profiling

MetAnnotate[242] was used to create a taxonomic profile for the metagenome using the usearch option with its default set of taxonomic markers. An average coverage (mean per bp across the coding sequence) for each marker gene hit was calculated using Bowtie 2 v2.3.4.2,[157] SAMtools v1.9,[170] and BEDtools v2.27.1.[252] The relative frequency of each genus was determined for every marker gene based on the cumulative average coverage. Average relative frequency across each marker gene was then calculated.

Metagenome annotation

A BLASTP search using the BLAST v2.6.0+ package of the "protein homolog" model types in the Comprehensive Antibiotic Resistance Database[125] (CARD) (databases downloaded on 28 June 2018) using CARD's own per-model bit score cut-off was used to find

[4]Sequencing services provided by Agricultural Research Council–Biotechnology Platform Laboratory.

putative antibiotic resistance genes. Average coverage of each gene hit was calculated as described earlier. KEGG annotations were identified with GhostKOALA[133] using the "prokaryotes" setting on 9 January 2020. TIGRFAM annotations were determined with the TIGRFAM database v15.0[105] using a threshold of 1×10^{-3} with hmmscan from HMMER v3.1b1. Pfam annotations were derived from Pfam v27.0[75] and applied with HMMER v3.1b1 and Pfamscan (at ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/). COG annotations were performed by Anvi'o v5.2[69] with the COG 2014 database[306] files sourced from ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data/.

2.2.3 Results

A total of 28,540,348 read pairs with an average read length of 119 bp each were generated. The total assembly length was 311,492,658 bp, with an N50 value of 861 bp. A profile of the community based on taxonomic marker genes was constructed with MetAnnotate.[242] The average coverage of each gene was calculated as the mean coverage per base pair across the coding sequence. The most common genera present (based on the average coverage across all taxonomic markers) are *Thauera* (19%), *Oscillibacter* (7%), *Pseudomonas* (6%), and *Prevotella* (5%) (Figure 2.3a).

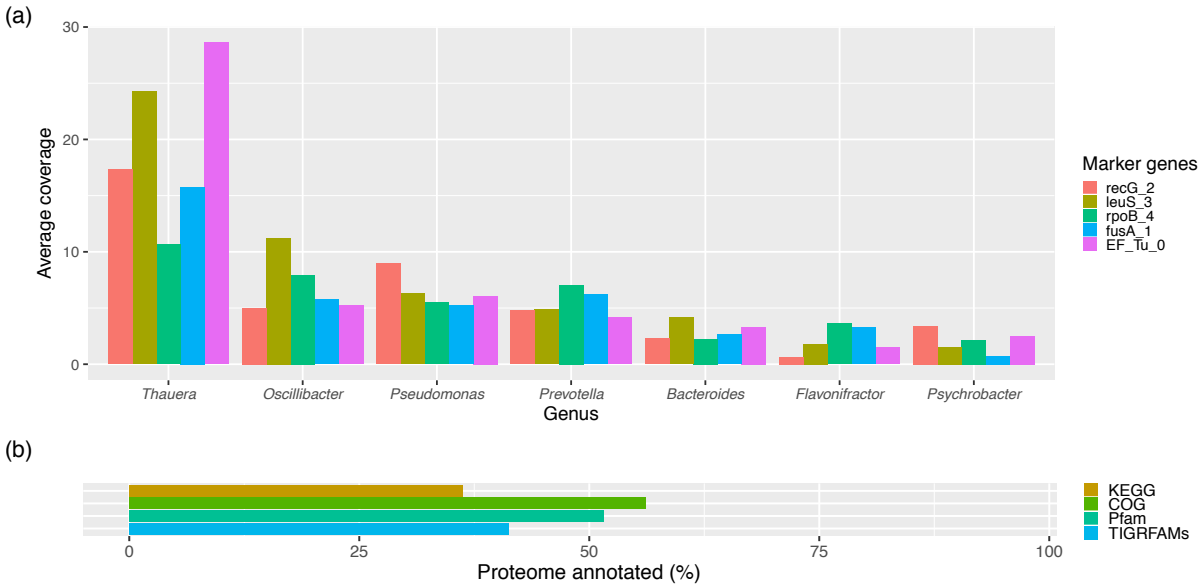

Figure 2.3: Taxonomic profile and annotation coverage of a farm wastewater metagenome from South Africa. (a) Coverage of marker genes at a genus-level for the farm wastewater metagenome. The average coverage is the mean of the coverage per base pair across the matched coding sequences. Most common genera are shown. (b) Annotation coverage of the farm wastewater metagenome.

A BLASTP search of the homolog models in the Comprehensive Antibiotic Resistance Database (CARD)[125] identified 31 different antibiotic resistance genes that passed CARD's strict score threshold. Ten of these genes are predicted to confer tetracycline resistance, and five genes are predicted to confer streptomycin resistance (Table 2.2). Annotation with the KEGG database[131] revealed complete pathways for tetracycline, streptomycin, aminoglycoside, cationic antimicrobial peptide (CAMP), vancomycin, and macrolide resistance, with near-complete pathways for beta-lactam, erythromycin, fluoroquinolone, and lincosamide resistance.

Table 2.2: Antibiotic resistance associated with three or more genes in the wastewater metagenome. The type of predicted antibiotic resistance came from CARD "parent terms" for each gene in their database.

Predicted antibiotic resistance	**Average coverage of associated genes**	**Count of associated genes**
tetracycline	33.49	10
minocycline	63.99	4
chlortetracycline	63.99	4
doxycycline	62.00	3
oxytetracycline	62.00	3
demeclocycline	62.00	3
streptomycin	32.77	5

Broader annotation methods were applied to get an overview of this metagenome's annotation coverage. KEGG, COG, Pfam, and TIGRFAMs contributed to a modest proportion of protein coding sequences annotated, from as low as 36% up to only 56% (Figure 2.3b). This leave almost half of all the predicted coding sequences without any biological information, apart from the environment they were found in.

2.2.4 Discussion

The sequencing of the South African Agricultural Research Council–Animal Production wastewater provides information on the microbial community that lives on the farm and whether there exists the potential for any human pathogens to be present. The most common genera found in the wastewater sample were *Thauera*, *Oscillibacter*, *Pseudomonas*, and *Prevotella*. *Thauera* has been found previously in agricultural wastewater, hot springs, a leachate treatment plant, and in sludge from ditches, water treatment plants, and oil-refineries.[204,278,279,304,347] This genus has been characterized as denitrifiers that have the ability to degrade aromatic compounds,[204,278] such as phenol,[293] that can end up in agricultural wastewater.[227] A high proportion of *Oscillibacter* also makes sense in farm wastewater as this genus has been found in cattle rumen, as well as being closely related to other bacteria found in sheep, cow, and goat gut samples.[161,186] Interestingly, several cases of bacteremia in Denmark were reported to be caused by *Oscillibacter ruminantium*, although all cases had risk factors for infection.[302] *Prevotella*, a common commensal in human gut associated with a plant-based diet,[95] is exceptionally abundant in cow rumen. One study in 2012[121] found that *Prevotella* made up an average of 52% of the rumen community. It is worth noting that some *Prevotella* are opportunistic pathogens in humans

and have been associated with chronic inflammation in periodontal disease, rheumatoid arthritis, and various gut disorders.[159] Potential pathogens aside, this metagenome has many links to the livestock-affected, high-nitrogen, aquatic environment that would compose the wastewater of an Animal Production research farm.

Antibiotic resistance is a well-studied topic as antibiotics are incredibly important for human health, as well as for livestock production. Rising levels of antibiotic resistance, especially in hospital settings, is a global health crisis.[45,125] There are whole databases focused on microbial antibiotic resistance. However, one of the broadest is the Comprehensive Antibiotic Resistance Database (CARD).[125] A global team of collaborators have curated this resistome database, generating a unifying Antibiotic Resistance Ontology (ARO) to guide the annotations. Protein models where antibiotic resistance can be accurately predicted via sequence similarity have strict alignment cut-offs and are separated from cases where antibiotic resistance is due to mutations. This database is a great resource for screening for antibiotic resistance in sequence data. Applying it to the research farm wastewater sample revealed a resistome geared primarily towards tetracycline resistance with other genes putatively providing streptomycin, minocycline, and chlortetracycline resistance, amoungst others. Annotation with KEGG confirmed matches to genes for tetracycline and streptomycin resistance. In a previous study by Noyes et al.,[229] sequences associated with tetracycline resistance were the most frequent in their 34 soil, manure and wastewater samples from various livestock operations across the U.S.A. and Canada. Aminoglycoside resistance (including streptomycin) was also fairly abundant in their detected resistomes.[229] Other resistome studies have found tetracycline to be common[276,310] in agricultural soil and honey bee gut communities, speculated to be due to years of oxytetracycline use in those environments.

Full-scale annotation was preformed on this metagenome sample to see how its annotation coverage compares to other annotated metagenomes. To briefly reiterate, due to differences in the way that databases create their annotation labels, database size, and database types (e.g. sequence versus model and full-length protein versus domain), there is a range of annotation coverage. However, even with differences in annotation coverage, only around half of the CDSs overall have annotated functional information. As seen from Figure 1.1, metagenomes often have a substantial proportion of their predicted CDSs that end up with no annotations. This metagenome follows that trend with the highest annotation coverage only reaching 56%. The problems that lead to low annotation coverage in metagenomes (fragmented sequences and organisms that are potentially quite distant from well-studied species) could affect the detection of antibiotic resistance, especially if there are understudied systems present. But this work, to the standards of our current knowledge on antibiotic resistance annotations, contributes to the growing data on human-influenced resistomes.

2.3 Functional profiling of a fish necrobiome reveals a decomposer succession involving toxigenic bacterial pathogens

Material in this section has been published as part of Lobb et al. (2020).[178] The published manuscript is available here:

> B. Lobb, R. Hodgson, M. D. Lynch, M. J. Mansfield, J. Cheng, T. C. Charles, J. D. Neufeld, P. M. Craig, and A. C. Doxey. Time series resolution of the fish necrobiome reveals a decomposer succession involving toxigenic bacterial pathogens. *mSystems*, 5(2):e00145-20. 2020.[178] https://doi.org/10.1128/mSystems.00145-20

2.3.1 Introduction

The decomposition of animal tissues is a fundamental ecological process that impacts nutrient cycling and species composition in terrestrial and aquatic ecosystems. Vertebrate tissue decomposition creates a unique ecological niche supporting a wide variety of specialized decomposer species, including insects, predators, and microorganisms. These species form an interconnected community whose combined activities lead to the decomposition of an organism from its initial death to the complete degradation of its exterior and internal contents.

The microbial communities involved in decomposition, including bacteria derived from the surrounding environment (e.g., water, soil) and the host (e.g., digestive tract and lungs), are collectively referred to as the "necrobiome" (from nekrós, the Greek word for dead body),[36] or alternatively, the "thanatomicrobiome" (from Thanatos, the Greek god of death).[124] Studies of necrobiome structure and function in several model systems (e.g., human, cow, pig, and mouse) have revealed strong microbial succession with distinct taxonomic and functional shifts linked to the phases of tissue decomposition.[28,100,117,208,209,237] After cellular autolysis breaks down tissue following death, anaerobic bacteria such as *Clostridium* spp. increase in relative abundance and metabolize available carbohydrates and proteins from the body, producing organic acids and gas.[33] Functional shifts occur; these shifts include increases in catabolic pathways, carbohydrate and energy metabolism, nitrogen cycling, and processes related to bacterial invasion. Foul-smelling compounds associated with the process of putrefaction are also produced as by-products of fermentation

and amino acid decomposition, including putrescine, cadaverine, and indole. Because putatively pathogenic bacteria proliferate within vertebrate necrobiomes, such as *Clostridium botulinum*,[31] it has been proposed that bacterial toxins secreted by these bacteria may play roles in decomposition by interfering with host cellular functions.[191]

Although much knowledge of necrobiome community structure and function has come from studies of terrestrial mammals, less is known about the structure, function, and dynamics of decomposition in aquatic ecosystems. Previous studies of fish carcass decomposition demonstrate that as in terrestrial systems, both macroinvertebrates and microorganisms play important roles as aquatic decomposers.[211,246] But what metabolic activities/functions are present in aquatic necrobiome communities and how do they change over time? Comparing necrobiomes between two different locations in the Grand River (southwestern Ontario, Canada), upstream and downstream of a wastewater treatment plant, allows analysis of community members and their functional potential both spatially and temporally. In this study, I used a variety of annotation tools and methods to functional and taxonomically profile the necrobiomes including broad metabolic database KEGG and specific database VFDB (Virulence Factor Database). Here, studying necrobiome-associated functions provides a unique way to better understand links to aquatic health, fish physiology, and ecosystem dynamics.

2.3.2 Methods

Fish collection

Sample collection and preparation done by Rhiannon Hodgson and Dr. Paul Craig.

On 24 October 2016, female rainbow darters (*Etheostoma caeruleum*) were collected from the Grand River (Figure 2.4), both upstream (Westmontrose [WMR]; 43°35'08"N; 80°28'53"W) and downstream (Economic Insurance Trail [EIT]; 43°28'24"N; 80°28'22"W) of the Waterloo wastewater treatment plant (WWTP) (43°29'16"N; 80°30'25"W). Forty-two fish (21 from each site) were collected using a backpack electrofisher (Smith Root, LR-20) and euthanized quickly with a sharp blow to the head. Then each fish was placed in an autoclaved 250-ml mason jar microcosm that contained a mixture of water and river substrate (see Lobb et al., 2020[178] for river water quality metadata and Figure 1a for an example mason jar setup). The lids were closed, but not sealed, in order to ensure oxic conditions that would accompany natural in-river decay events. The jars were then left to decay in a fume hood at room temperature. Three samples containing both fish and water/sediment from the same site were left to decompose for 1 day (24 h), 4 days, 8 days,

and 10 days for both the WMR and EIT sites, totaling 24 fish. For additional treatments to assess differences in water quality and aquatic microorganisms, three samples containing fish and water/sediment from different sites (i.e., WMR fish in EIT conditions and EIT fish in WMR conditions) were allowed to decay for 4, 8, and 10 days, totaling 18 fish. At each time point, decay was documented (Figure 1b), and fish were removed from the replicate jars, then rinsed with sterile water, and ground with liquid nitrogen using a clean mortar and pestle. The powdered tissue was stored at –80°C prior to genomic DNA extraction.

Experimental procedures and the use of animals in this study were approved by the University of Waterloo Animal Care Committee and within Canadian Council on Animal Care (CCAC) guidelines (AUPP 40318).

DNA extraction

DNA extraction done by Metagenom Bio Life Science Inc.

Unless noted, all chemicals and reagents were purchased from Sigma-Aldrich (Mississauga, Ontario, Canada). For DNA extraction, 100 mg of ground tissue was added to 1.2 ml of TE buffer (10 mM Tris-HCl, 1 mM EDTA [pH 8.0]), 100 μl of 10% sodium dodecyl sulfate (SDS), 20 μl of proteinase K, 8 μl of RNase A, and 200 μl of 5 M NaCl. This mixture was vortexed quickly and incubated at 55°C for 30 min. Then 160 μl of CTAB extraction solution (2% cetrimonium bromide, 100 mM Tris, 20 mM EDTA, 1.4 M NaCl [pH 8.0]) was added, and the samples were further incubated at 65°C for 1.5 h. Following this lysis incubation, 700 μl of the lysate was extracted with an equal volume of phenol and centrifuged at 10,000 × g for 5 min. The aqueous phase was retained and twice extracted with equal volumes of phenol-chloroform-isoamyl alcohol (25:24:1), followed each time with centrifugation at 10,000 × g for 5 min. One volume of isopropanol was used to precipitate aqueous phase DNA in a new ultracentrifuge tube, followed by centrifugation at 13,000 × g for 10 min at room temperature. The resulting pellet was washed twice with 70% ethanol, dried, and then dissolved in 50 μl of DNase- and RNase-free H2O (Sigma) at 50°C for 15 min. The quantity and quality of DNA were determined with a SpectraDrop (Molecular Devices) and stored at –20°C prior to sequencing.

16S rRNA gene and metagenomic sequencing

Sequencing services provided by Metagenom Bio Life Science Inc.

Extracted DNA was amplified in triplicate using Pro341F and Pro805R universal prokaryotic primers.[305] Triplicate amplicons were pooled, gel quantified, and sequenced to

a depth of at least 30,000 paired-end reads per sample using the MiSeq reagent kit v3 (2 × 300 cycles; Illumina).

For metagenomic sequencing, genomic DNA (1 ng) was fragmented and individually barcoded using the Nextera XT DNA Library Prep kit (Illumina) following the supplier's guidelines. Small fragments of library DNA were removed by adding 0.6 volumes of AMPure XP beads (Beckman Coulter). After washing twice with 80% ethanol and air drying for 10 min, DNA was eluted from the beads with 10 mM Tris-HCl (pH 8.5). Purified library DNA was quantified with the Qubit dsDNA (double-stranded DNA) HS (high-sensitivity) assay kit, diluted to 4 nM with the Tris-HCl buffer and then pooled in an equal volume. Library DNA was denatured with equal volumes of 0.2 N NaOH, diluted to 7 pM with hybridization buffer HT1, and sequenced with MiSeq reagent kit v2 (2 × 250 cycles; Illumina).

All 16S rRNA gene and metagenomic sequencing data for this project were deposited into the NCBI Short Read Archive (SRA) under BioProject accession no. PRJNA604775.

16S rRNA gene analysis

QIIME processing of 16S rRNA sequence data done by Dr. Michael Lynch.

Demultiplexed sequences were processed using DADA2 v1.4,[29] managed through QIIME2 v.2017.10.[21] Briefly, forward and reverse reads were truncated with decreasing quality metrics while maintaining sequence overlap (∼250 bases). Primers were removed, and paired reads were assembled after error modeling and correction, creating amplicon sequence variants (ASVs). Chimeric ASVs were removed by reconstruction against more abundant parent ASVs. The resulting ASV table was constructed for downstream analysis (see Lobb et al., 2020[178]).

Taxonomy was assigned to representative sequence variants using a naive Bayesian classifier implemented in QIIME2 with scikit-learn (v.0.19.0), trained against SILVA release 128,[251] clustered at 99% identity, and trimmed to the amplified region. Assignments were accepted above a 0.7 confidence threshold.

For ordination, a proportion matrix of ASVs were used across each sample with a sparsity cutoff (i.e., ASV detected in at least 3 of 42 samples). The metaMDS() and envfit() scripts from vegan package v2.4-2 in R were used to calculate ordination coordinates and data vectors. A stress or Shepard diagram was generated with stressplot() from the vegan package to determine the nonmetric fit.

Metagenomic data analysis

Raw reads were processed with TrimGalore v0.5.0,[148] coassembled with metaSPAdes (SPAdes v3.12.0),[230] and eukaryotic contigs were identified with Centrifuge v1.0.4[141] using their NCBI nr preindexed database (last updated 3 March 2018) and subsequently removed. Reads were mapped with Bowtie 2 v2.3.4.3[157] using default settings and binned using CONCOCT[6] with Anvi'o v5.2 (minimum 1 kb contig cutoff).[69] Mean coverage data for the metagenomic functional analyses and for the methanogen analysis were extracted from Anvi'o[69] using all contigs (no contig length cutoff).

For metagenomic and bin functional analysis, KEGG annotations were identified with GhostKOALA.[133] The average coverage for each gene (per base pair), normalized by dividing by the average sample coverage (per base pair), was summed to give a total coverage value for each KEGG pathway. The decostand() function from the vegan package v2.4-2 in R was used to determine the fractional value of each pathway with respect to the total summed coverage across all KEGG pathways detected in the sample. A Kruskall-Wallis test was done in R to identify KEGG pathways with significantly different distributions by day of decomposition. The decostand() function was also used to proportionally normalize each pathway value across every sample for plotting. For the bin functional analysis, the frequency of each KEGG orthology (KO) annotation in each MAG bin was counted. These counts were summed for each KEGG pathway, and fractional values were calculated across all KEGG pathways detected in the bins as before.

The VFanalyzer software from the Virulence Factor Database (VFDB)[174] identified virulence factors in the predicted coding sequences of Bin_4 using *Aeromonas veronii* B565 as a representative genome. The domain architecture from the *Aeromonas* toxin gene set from the VFDB was also used to identify *Aeromonas* toxin genes in the coassembly. Putative toxins longer than 150 amino acids were assessed with BLASTP for *Aeromonas* taxonomy and gene annotation. The *Aeromonas* phylogenetic tree was made using RAxML v8.2.12[295] with the LG likelihood model made from concatenated single-copy core protein sequences detected with Anvi'o[69] (Campbell et al. set[30]).

Additional annotation methods were used as described in Sections 2.1.2 and 2.2.2.

2.3.3 Results and Discussion

Time series community profiling of fish necrobiomes

To examine the structure and temporal succession of aquatic vertebrate necrobiomes, a 16S rRNA-based study of decomposing fish was performed at different time points and locations.

Female rainbow darters (*Etheostoma caeruleum*) were collected from the Grand River in Waterloo, Ontario, Canada, both upstream and downstream of the Waterloo wastewater treatment plant (WWTP) (Figure 2.4). Individual fish were subjected to decomposition with river water and sediment at room temperature for 1, 4, 8, and 10 days in sterile containers that acted as microcosms of a natural decomposition environment. Sample 16S rRNA gene profiles for fish decomposition microbiomes ("necrobiomes") for these four time points and two water/sediment sources revealed reproducible microbial communities among independent replicates and also between environments (i.e., fish and water source; Figures 2.5 and 2). This microbial succession was apparent at the order level of taxonomy (Figure 2.5) and at the level of amplicon sequence variants (ASVs) (Figure 2), although variation in ASV composition was evident among fish samples and environments (Figure 2).

Further discussions on the 16S rRNA profile and differences in taxa throughout the time course and up and downstream of the WWTP are available in the original article.[178]

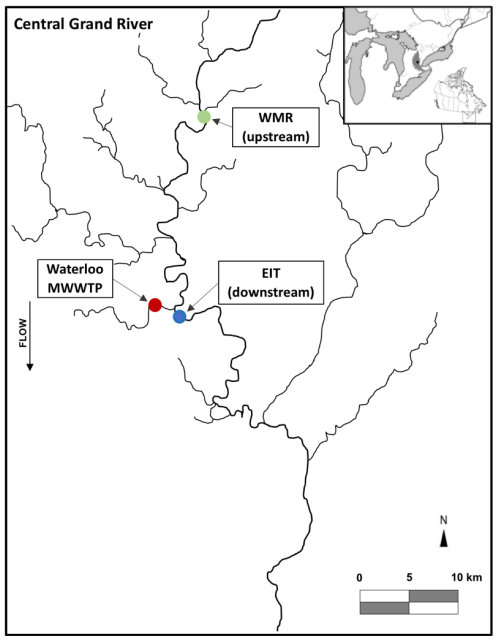

Figure 2.4: Map showing sampling locations of Grand River fish for metagenomic analysis. The municipal wastewater treatment plant (WWTP) for the city of Waterloo, Canada, and the two sampling locations, upstream at West Montrose (WMR) and downstream at the Economic Insurance Trail (EIT), are displayed. *Figure created by Dr. Paul Craig.*

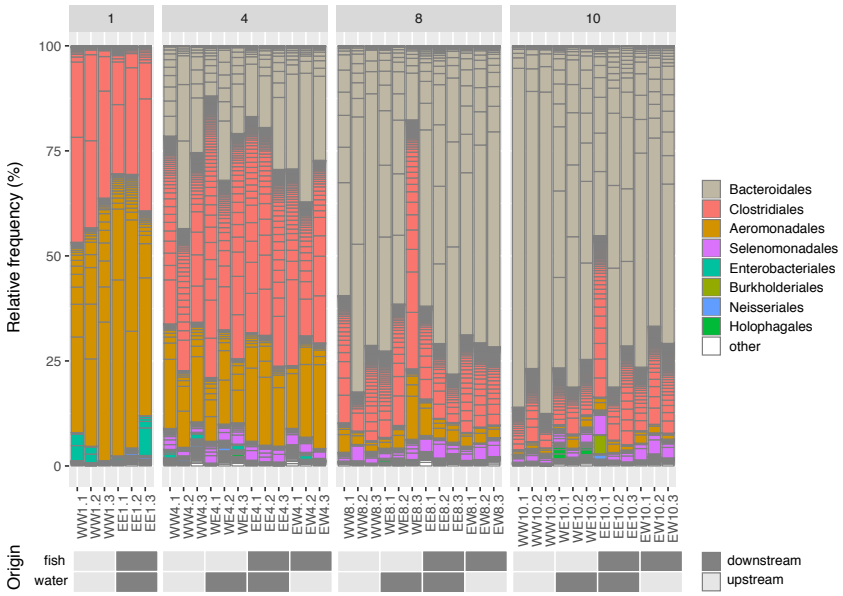

Figure 2.5: Relative frequency of ASVs within each sample colored by taxonomic order. Samples are sorted by decomposition time (1 day, 4 days, 8 days, and 10 days). The fish and water/sediment origin of the samples are displayed at the bottom of the figure, with upstream referring to the WMR site and downstream referring to the EIT site. Low-relative-abundance taxonomic orders are grouped into "other."

Metagenomic binning and analysis of decomposition pathways

To explore the genomes and genome-encoded metabolic/functional potential of the necrobiomes, metagenomic sequencing was performed on one replicate for each condition (14 total). Subsequent assembly and binning resulted in four MAGs with >85% completion and <5% redundancy. I examined the taxonomic composition of the MAGs using MetAnnotate.[242] These MAGs included two genomes affiliated with *Alistipes* (Rikenellaceae), a genome annotated as *Aeromonas veronii*, and a Selenomonadaceae-associated genome (Table 2.3). The bins are consistent with ASVs identified by the 16S rRNA gene sequencing, corresponding to *Acetobacteroides* (Rikenellaceae), *Aeromonas*, and various members of Selenomonadales (Figures 2.5 and 2). Other ASVs identified by 16S rRNA gene sequencing were also recovered in the lower-quality MAGs (Table 2.3). One bin was affiliated with the genus *Pseudomonas*, and another bin was affiliated with the family Rikenellaceae.

Table 2.3: Bins obtained from metagenomic sequencing of fish necrobiomes. Taxonomic affiliation is predicted by MetAnnotate.[242]

	Completion (%)	**Redundancy (%)**	**GC (%)**	**Total length (Mb)**	**Gene count**	**Contig count**	**Taxonomic affiliation**
Bin_4	98.6	0.7	60.7	3.85	3855	784	Bacteria; Proteobacteria; Gammaproteobacteria; Aeromonadales; Aeromonadaceae; *Aeromonas*; *Aeromonas veronii*
Bin_9	97.1	1.4	47.5	2.25	2216	402	Bacteria; Firmicutes; Negativicutes; Selenomonadales; Selenomonadaceae
Bin_3	87.1	2.2	47.0	2.64	2467	801	Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae; *Alistipes*
Bin_10	92.8	2.2	44.0	3.26	2882	368	Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae; *Alistipes*
Bin_7	38.8	7.9	61.4	0.78	1187	628	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; *Pseudomonas*
Bin_2	25.2	1.4	48.2	1.71	1872	960	Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae

The relative abundance of Bin_4 (*Aeromonas veronii*) decreased throughout decomposition from an average relative abundance of 3.7 (day 1) to an average relative abundance of 0.14 (day 10; Figure 2.6a), consistent with the 16S rRNA data. Because *Aeromonas* has been associated with fish gut microbiomes,[91,127,216,299,319] it is possible that Bin_4 and other *Aeromonas* taxa were initially derived from the fish guts and were important only for early stage decomposition. In contrast, Bin_3 (Rikenellaceae family) may represent a late-stage decomposer because its relative abundance increased in metagenomes from days 8 to 10 of decomposition (average relative abundance of 3.9 on day 8 to an average relative abundance 5.1 on day 10; Figure 2.6a). In the downstream fish-upstream sediment/water set, both

Rikenellaceae-affiliated bins (Bin_3 and Bin_10) were similar in relative abundance, implying site-specific influences on the relative abundance of different Rikenellaceae-affiliated taxa, consistent with 16S rRNA gene data for *Acetobacteroides* ASVs (Figure 2). Phylogenetic analysis of the two Rikenellaceae-associated bins revealed that Bin_3 was more closely related to *Acetobacteroides hydrogenigenes* RL-C and Bin_10 was more closely related to *Alistipes* sp. strain ZOR0009 (Figure 2.6b). Bin_9 (*Propionispira*) was present at low (0.0 to 0.54 average on days 1 to 10; Figure 2.6a) relative abundance, close to the sample's mean coverage across the entire course of decomposition, consistent with the abundance patterns seen for Selenomonadales based on 16S rRNA gene data (Figure 2.5).

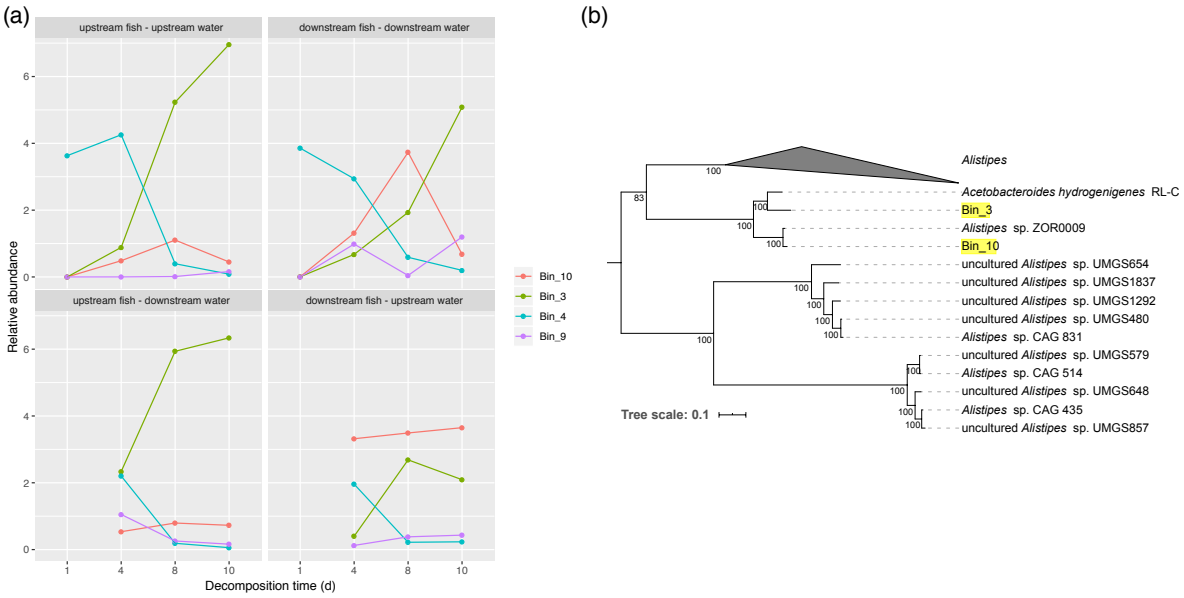

Figure 2.6: Metagenomic bin relative abundance and phylogenetic analysis of Bin_3 and Bin_10. (a) Relative abundance of four high-quality binned genomes across each necrobiome sample. Relative abundance was computed as mean bin coverage/mean sample coverage. Mean coverage was calculated per base pair using Anvi'o. (b) RAxML tree using the LG likelihood model made from concatenated single-copy core protein sequences detected with Anvi'o[69] (Campbell et al. set[30]). The tree outgrouped with *Lentimicrobium saccharophilum*. *Acetobacteroides hydrogenigenes*, representatives of *Alistipes* strains, and all uncharacterized *Alistipes* isolates were used for this tree and sourced from NCBI Genome. This tree was visualized with iTOL.[165]

Using a KEGG analysis of assembled contigs and binned metagenomes, metabolic

pathway potentials associated with the decomposition samples were examined. The resulting functional profiles had a highly similar grouping in ordination space compared to the 16S rRNA gene community profiles (Figure 3), whereby samples grouped primarily based on decomposition time point (Figure 2.7). Analysis of specific KEGG pathways revealed patterns consistent with a functional succession (Figure 2.8), mirroring the taxonomic succession described earlier. Pollutant degradation pathways for polyaromatic hydrocarbons such as naphthalene, styrene, and nitrotoluene showed increased relative abundances on day 1 (13% on average) compared to subsequent time points (6.2% on average). The initial fish bacterial community may have been enriched for microorganisms that could degrade river water contaminants, which can originate from both anthropogenic and natural sources and bioaccumulate in fish.[39,107,203] Naphthalene degradation in polluted sediment-water systems can be accomplished through several bacterial pathways, and bioremediation of this toxic molecule by native organisms is currently being studied.[162,313,321] Various biofilm formation pathways were also proportionally abundant (13%) within day 1 metagenomes (Figure 2.8), possibly reflecting skin and gut community functions originating prior to decomposition. Degrading river water contaminants and skin and gut biofilm formation may be functions that are more important for the bacterial communities living with their fish host and dealing with possibly contaminated river water than for the necrobiome that formed in the closed system after the fish's death.

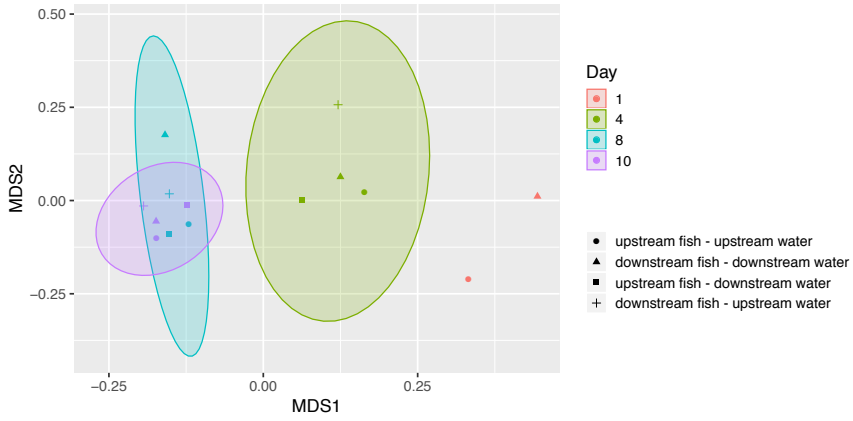

Figure 2.7: NMDS ordination of metagenomic functional profiles with Bray-Curtis distances calculated based on KEGG pathway frequencies. A strong agreement between the ordination space and the distance matrix was observed ($R^2 = 0.996$), and the stress value is 0.063.

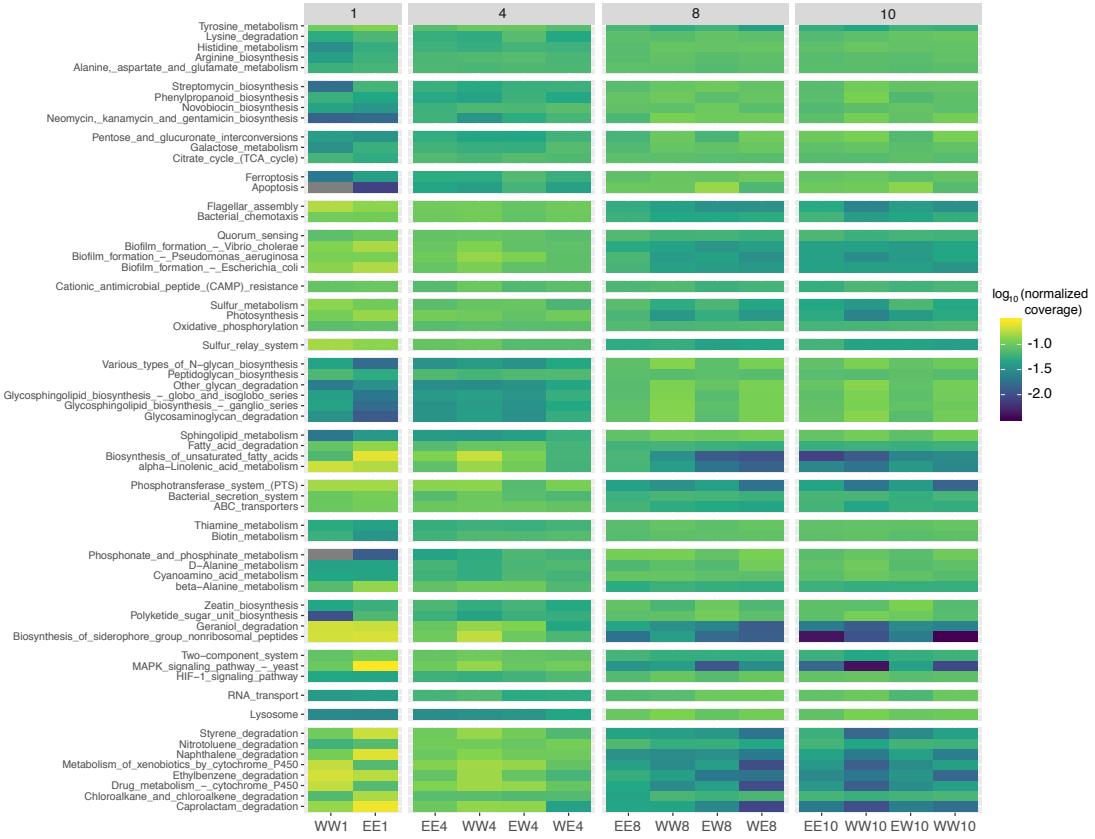

Figure 2.8: Selected KEGG pathways displaying significant differential relative abundance across the course of decomposition. Pathways were selected that had an unadjusted P value of <0.03 after a Kruskall-Wallis test comparing decomposition time (1, 4, 8 , and 10 days). Shown is the $\log_{10}$ value of the fractional coverage of the pathway with respect to the total coverage across all the pathways in the sample. Total pathway coverage is also proportionally normalized across every sample. Note that some pathways are based on a few representative genes. For example, coverage of the photosynthesis pathway is mainly derived from genes encoding sodium ion pumps.

Glycan metabolism generally increased in coverage from early stages (2.4% on day 1) to later stages of decomposition (10%). Glycan degradation pathways (e.g., glycosaminoglycans) increased in coverage by days 8 and 10, which may be involved in decomposition of fish skin and intestinal mucins. Late-stage increases in streptomycin, phenylpropanoid, novobiocin, neomycin, kanamycin, and gentamicin biosynthesis pathways (2.4-fold change from day 1 to 10) were also detected, implying that the remaining microorganisms by day

10 possess increased potential for antibiotic synthesis.

These metagenome-wide functional patterns closely matched the functional potentials of individual *Aeromonas* (early stage) and Rikenellaceae (late stage) bins, when taking into consideration their shifts in relative abundance through the time course (Figure 2.9). Genes belonging to pollutant degradation pathways were present in the *Aeromonas* bin yet mostly absent from other MAGs with lower relative abundance from days 1 and 4 metagenomes. Likewise, biofilm formation pathway genes had a 6.2-fold-higher frequency in the *Aeromonas* bin compared to the *Acetobacteroides*/*Alistipes* bins. In contrast, antibiotic biosynthesis pathway genes had a 2.5-fold-higher frequency in the Rikenellaceae-associated bins, in addition to multiple key glycan degradation genes. Thus, the detected shifts in functional profiles were in part due to the hand-off microbial community dominance from *Aeromonas* to Rikenellaceae. It is important to note that these apparent late-stage functional shifts could also be important for earlier phases when Rikenellaceae initially began to increase in relative abundance.

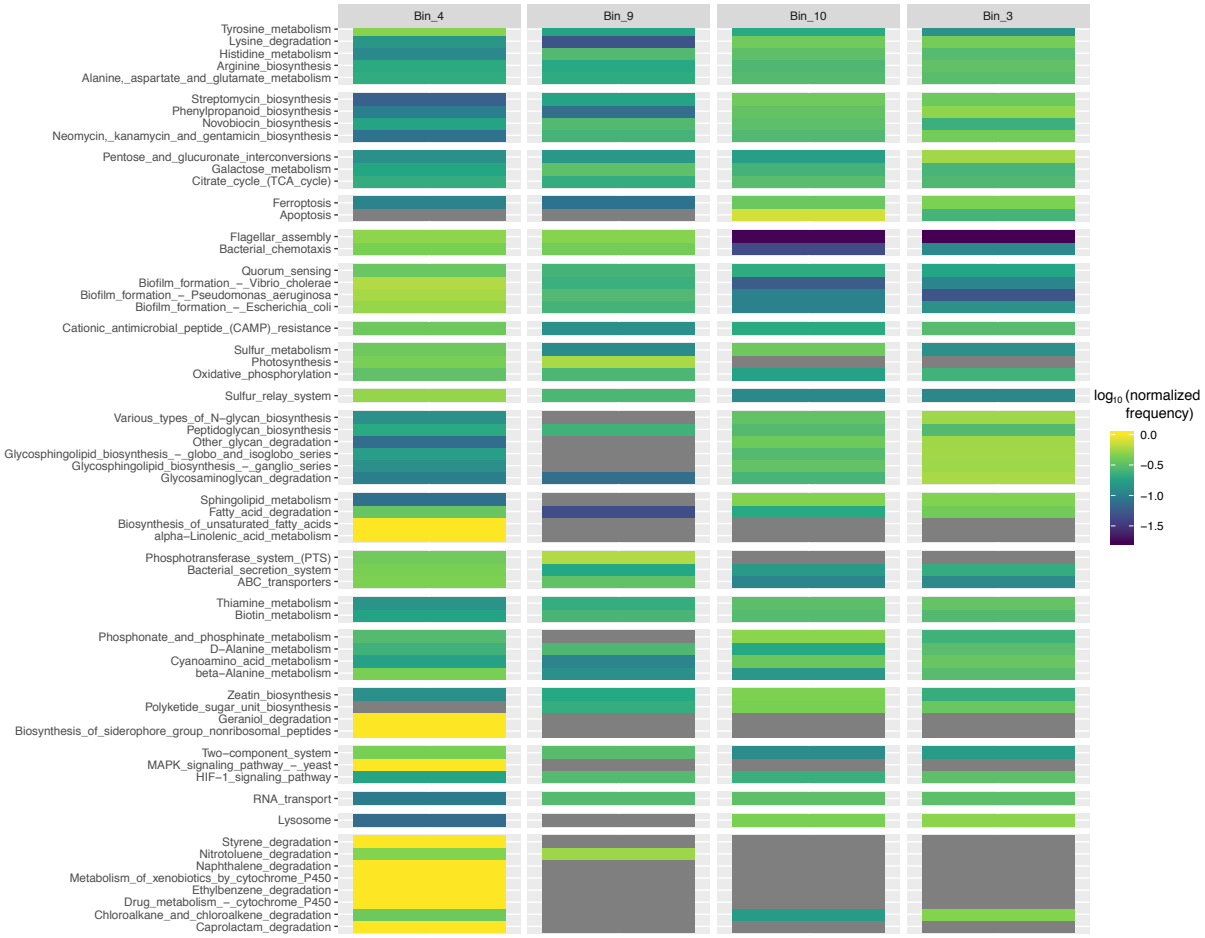

Figure 2.9: Count of KEGG annotations mapping to the corresponding KEGG pathways in Figure 2.8 across each MAG. Shown is the $\log_{10}$ value of the fractional frequency of the pathway with respect to the total across all the pathways in the sample. The total pathway coverage is also proportionally normalized across every sample.

The data suggests strong *Acetobacteroides* dominance in late-stage rainbow darter necrobiomes (Figure 2.5 and 2). Because related species have been implicated in anaerobic sugar fermentation,[298] I investigated the two MAGs affiliated with these bacteria for glycolytic enzymes. Both Bin_3 and Bin_10 possess a complete glycolysis pathway as well as l-lactate dehydrogenase for anaerobic fermentation (Figure 2.10). Bin_3 genes also encode pyruvate dehydrogenase, aldehyde dehydrogenase, and enzymes for conversion of d-fructose, d-fructose-1-phosphate (d-fructose-1P), and d-mannose-6P to glycolysis precursors. Based

on a previous analysis of decomposition pathways (51), Bin_3 and Bin_10 genes also encode components of potential pathways for production of indole [EC 4.1.99.1], putrescine [EC 3.5.3.11], and spermidine [EC 2.5.1.6 and 2.5.1.16], in addition to histidine degradation [EC 4.3.1.3, 4.2.1.49, and 3.5.3.8[5]].

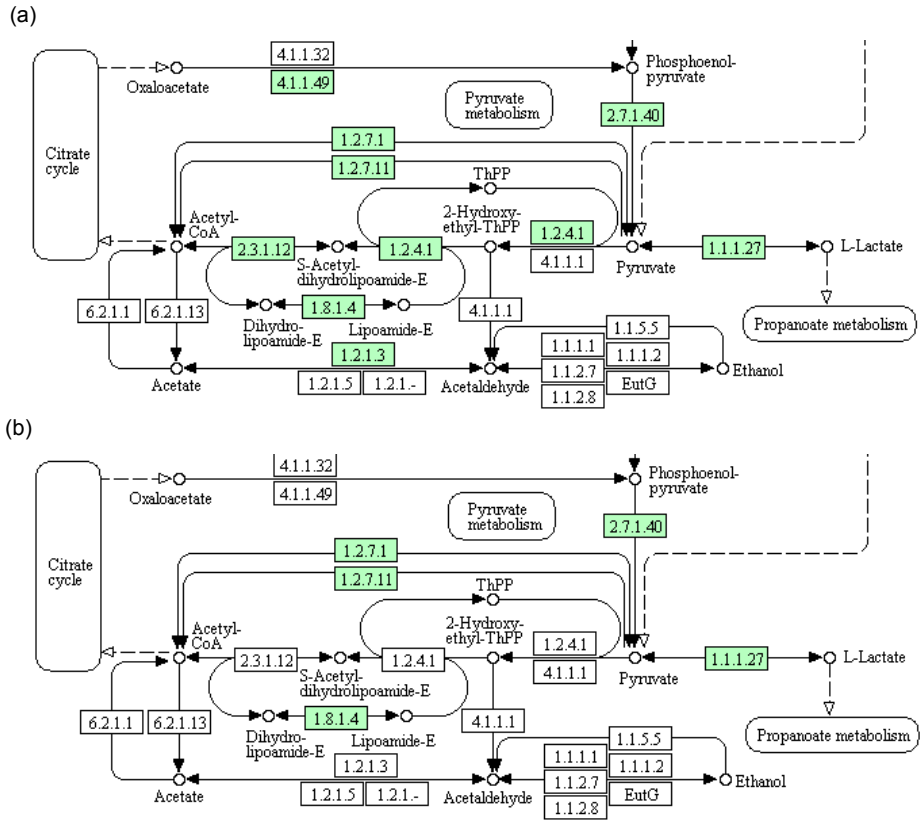

Figure 2.10: End of the KEGG glycolysis/gluconeogenesis pathway for Rikenellaceae (a) Bin_3 and (b) Bin_10. Green indicates the presence of a match to that enzyme. Images were generated using KEGG.[131]

[5]This enzyme is only detected in Bin_10.

A toxigenic strain of *Aeromonas veronii* is a dominant member of the necrobiome

Because Bin_4 affiliated with *A. veronii*, a well-established pathogen of fish and humans,[52,93,122,128,146,199,256,257] and a common inhabitant of the fish gut microbiome,[91,127,216,299] I explored its phylogenetic position, functional profile, and virulence repertoire. A maximum likelihood phylogeny of *A. veronii* and other related *Aeromonas* genomes from the NCBI was constructed based on a concatenated alignment of conserved ribosomal marker genes (Figure 2.11a). Within this phylogeny, Bin_4 grouped with a clade of *A. veronii* genomes but as a basal lineage outgrouping all *A. veronii* species except AMC34.

VFanalyzer from the Virulence Factor Database (VFDB)[174] was used to detect virulence factors within Bin_4 and compare it to a reference *Aeromonas* strain, *A. veronii* B565. VFDB focuses on experimentally-confirmed virulence factors and is a comprehensive database of virulence factors from medically-relevant pathogens. The VFanalyzer pipeline uses a hierarchical series of homology searches of the VFDB in order to find close matches with stringent cut-offs before moving to more permissive methods for divergent virulence factors.[174] VFanalyzer revealed that Bin_4 contained virulence-related genes for adherence, iron uptake, and secretion systems. Indeed, a total of 54 genes that were associated with secretion systems were identified, compared to only 15 in *A. veronii* B565. In addition, I identified 13 genes associated with endotoxin production. Like *A. veronii* B565, Bin_4 genes encoded hemolysin III, hemolysin HlyA, and a thermostable hemolysin gene (Figure 2.11b). I also recovered a relatively small incomplete bin (Bin_11, 0.64 Mb, 717 CDSs, 321 contigs) that correlated with Bin_4 in relative abundance. This small bin affiliated with *Aeromonas veronii* and also included a gene encoding aerolysin toxin production. Based on metagenomic mapped read coverage, the relative abundance of genes encoding *Aeromonas* toxins increased on day 4 of decomposition (Figure 2.11c), indicating an enrichment in *Aeromonas* strains carrying hemolytic proteins. A possible explanation for this is that lytic toxins, including those from *Aeromonas*, may function in host cell lysis during decomposition and therefore peak in relative abundance during earlier stages of decomposition. Bin_4 also possessed genomic potential for decomposition-related pathways, including histidine degradation (contains [EC 4.3.1.3, 4.2.1.49, 3.5.2.7, and 3.5.3.8]) and the production of putrescine [EC 4.1.1.19, 3.5.3.12, and 3.5.1.53], indole [EC 4.1.99.1], and cadaverine [EC 4.1.1.18].

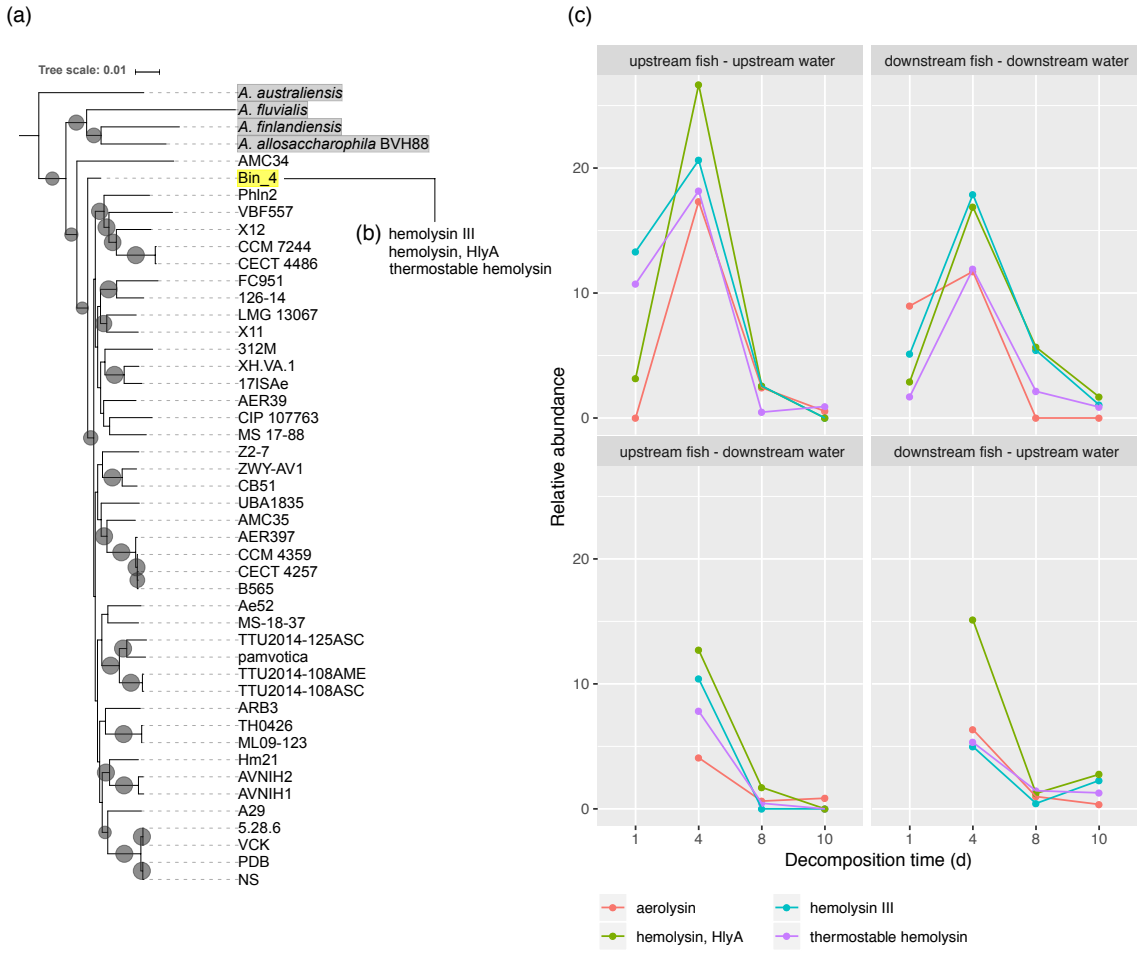

Figure 2.11: A toxigenic *Aeromonas veronii*-like strain is a dominant species in early decomposition. (a) RAxML tree using the GTR+GAMMA model made from concatenated single-copy core gene nucleotide sequences detected with Anvi'o (Campbell et al. set[30]). The tree was outgrouped on *Aeromonas hydrophila*. Gray circles are scaled to bootstrap support of ≥85, with the largest size representing 100. *Aeromonas* species outside *Aeromonas veronii* are highlighted in gray. Representative *Aeromonas veronii* strains from the NCBI Genome Tree report were chosen to display here (not highlighted), and only their strain name is shown. This tree was visualized with iTOL.[165] (b) Bin_4's predicted toxin repertoire from VFDB. (c) Relative abundance (mean gene coverage/mean sample coverage) of *Aeromonas hemolysin* toxin genes. Decomposition time is shown in days.

Overview of annotation coverage in metagenome and metagenome-assembled genomes

A selection of popular annotation methods were used on both the necrobiome (pooled and co-assembled) and the bins that were assembled from the necrobiome. All methods achieved a higher annotation coverage on the bins, as opposed to the metagenome itself. The necrobiome contains shorter fragments than the bins (as the binning procedure discards any sequence fragments smaller than 1 kb). These shorter fragments can lead to fragmented predicted coding sequences that are challenging to annotate. The metagenome could also contain organisms that have low annotation coverage, due to taxonomic distance from well-studied species or having a higher proportion of divergent/uncharacterized protein families. Thus, the metagenome ends up with around half of its predicted coding sequences annotated (the highest coverage reached being 58%). The MAGs had better annotation coverage (42 - 88%) than the pooled metagenome, albeit with a larger range between the methods with the lowest coverage versus the methods with the highest coverage. This is most obvious for Bin_3 and Bin_10. Both of these bins have lower annotation levels for all methods compared to Bin_4 and Bin_9. However, Prokka and KEGG have a larger drop at 16% less annotation coverage, on average, versus 11% for the other methods. This perhaps indicates that the KEGG database and Prokka pipeline are not as well set up for the more obscure *Acetobacteroides* genus and divergent *Alistipes* (Bin_3 and Bin_10). This clade probably contains protein families not as well covered in KEGG, Uniprot, and HAMAP (used in Prokka) and divergent from currently characterized proteins, possibly relating to the reed swamp and zebrafish gut-like environments *Acetobacteroides hydrogenigenes* RL-C and *Alistipes* sp. strain ZOR0009 (GOLD ID: Gp0042493) were found in.[298] Although this does affect all the annotation methods, just not to the same extent. As Rikenellaceae are extremely dominant in the community on days 8 and 10 (increasing to a relative abundance of as much as 87% in the decomposer community by the final day of sampling, Figure 2), this probably lowers the annotation coverage (but especially with the KEGG database) in the pooled necrobiome, being some of the harder-to-annotate organisms mentioned above.

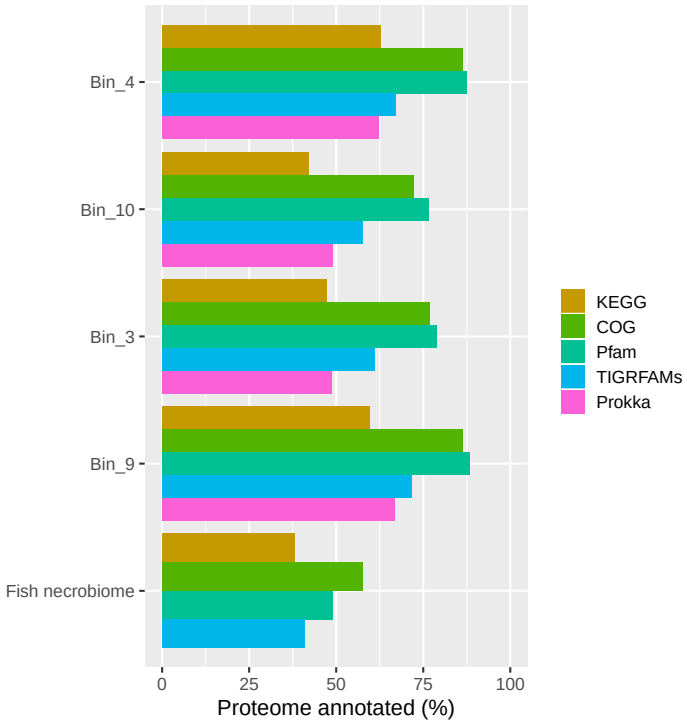

Figure 2.12: Annotation coverage in fish necrobiome and the four highest-quality MAGs. Prokka was not run on the metagenome as it is designed for genomes and the gene finding that is a part of its pipeline is not as effective on metagenomes.

2.3.4 Conclusion

Both 16S and metagenomic analysis revealed a strong succession in which initial time points were dominated by Clostridiaceae and *Aeromonas*, with Rikenellaceae species appearing by day 4 and becoming major community members by day 10. Analysis of functional profiles inferred from the metagenomic data revealed common decomposition pathways, as well as temporal shifts in function that mirrored taxonomic succession. Notably, pollutant degradation pathways and biofilm formation pathways were enriched in the early stages of decomposition and associated with Clostridiaceae and *Aeromonas*, and glycan metabolism and antibiotic synthesis increased in later stages and associated with Rikenellaceae. I also identified a toxigenic *Aeromonas* strain that was a dominant member of the necrobiome community. The presence of numerous hemolytic toxin genes in this organism suggests a potential role for toxins in the decomposition of host tissues as proposed previously.[191]

Lastly, investigating the overall annotation coverage of the metagenome-assembled genomes revealed that the two Rikenellaceae bins had a lower proportion of CDSs annotated compared to the other bins. Their annotation coverage possibly reflects their understudied place within the Rikenellaceae family. Further work investigating the prevalence and function of toxigenic and non-toxigenic bacterial species in decomposer communities will be important to explore their broader ecological roles and niches within natural ecosystems.

2.4 Summary

Homology-based annotation is the foundation of current genome and metagenome functional analyses. Through sequence-sequence and sequence-model methods, I annotated two newly sequenced *Streptomyces* strains, revealed the antibiotic resistome of a farm wastewater sample, and explored the functional profile of a rainbow darter necrobiome. There are many ways to annotate a dataset depending on the intended target and how high confidence the annotations should be. Targeted annotation databases are excellent resources if the genome/bin is part of a well-studied sub-group like *Aeromonas* or if the proteins you are targeting have their own focused, heavily-curated databases like CARD or VFDB. Multi-database comparisons can increase the number of coding sequences with functional information and can provide validation for function transfer, as done here in the search for cellulases. The case studies lead to the discovery of proteins of interest, from cellulase predictions in novel *Streptomyces*, to toxin genes found in decomposing fish. These case studies reinforce the concept that novelty can come from new contexts even if the protein family is already at least partially characterized.

The other take-away from these studies is that a significant proportion (12 - 44%) of newly sequenced genomes/metagenomes are not annotated with many current methods and databases. Intrinsic differences between annotation methods cause variability in the annotation coverage but all methods tested left a substantial number of predicted coding sequences unannotated. Even worse, some taxa were more in the "dark" than others. This raises important questions: what is the range of annotation completeness in other microbial taxa and what are the factors that can affect annotation coverage?

Chapter 3

Annotation completeness of bacterial genomes

Material in this chapter has been published as part of Lobb et al. (2020).[180] The published manuscript is available here:

> B. Lobb, B. J.-M. Tremblay, G. Moreno-Hagelsieb, and A. C. Doxey. An assessment of genome annotation coverage across the bacterial tree of life. *Microbial Genomics*, 6(3):e000341, 2020.[180] https://doi.org/10.1099/mgen.0.000341

Although gene-finding in bacterial genomes is relatively straightforward, the automated assignment of gene function is still challenging, resulting in a vast quantity of hypothetical sequences of unknown function. As seen in Figure 1.1 and the Chapter 2 case studies, a significant proportion of newly sequenced genomes/metagenomes are unannotatable with current methods and databases. But how prevalent are hypothetical sequences across bacteria, what proportion of genes in different bacterial genomes remain unannotated, and what factors affect annotation completeness? To address these questions, the genome annotation completeness of over 27,000 bacterial genomes from the Genome Taxonomy Database was surveyed, with a focus on annotation method, taxonomy, genome size, 'research bias' and publication date. Annotation coverage using protein homology-based searches varied significantly. However, taxonomy was a major factor influencing annotation completeness, with distinct trends observed across the microbial tree (e.g. the lowest level of completeness was found in the Patescibacteria lineage). Most lineages showed a significant

association between genome size and annotation incompleteness, likely reflecting a greater degree of uncharacterized sequences in 'accessory' proteomes than in 'core' proteomes. Finally, research bias, as measured by publication volume, was also an important factor influencing genome annotation completeness, with early model organisms showing high completeness levels relative to other genomes in their own taxonomic lineages. This work highlights the disparity in annotation coverage across the bacterial tree of life and emphasizes a need for more experimental characterization of accessory proteomes as well as understudied lineages.

3.1 Introduction

Genome annotation relies primarily on the detection of homology between newly identified genes/proteins and previously annotated sequences. Although complicated by varying definitions of "function" and "annotation", homology-based annotation transfer has been systematically explored, revealing success rates of upwards of 60–70% accuracy based on assessment of GO term prediction.[181,253] Studies of early model organisms, such as *Escherichia coli*, *Bacillus subtilis* and *Caulobacter crescentus*, are a major source of experimentally derived functional annotations. Therefore, it is important to note that such limited sources can be expected to result in biases in genome annotation, with a greater success rate in species that are phylogenetically closer to these and other commonly studies species.[98]

Both sequence-to-sequence and profile-based methods are implemented in common annotation pipelines such as Prokka,[280] the Joint Genome Institute Microbial Annotation Pipeline[201] and NCBI's Prokaryotic Genome Annotation Pipeline.[104] Annotation pipelines may also integrate a variety of methods and databases, and/or allow users to customize options towards specific reference databases or taxonomic lineages. Commonly used reference databases include UniProt/SwissProt, as well as the NCBI's reference sequence (RefSeq) database, and its non-redundant protein database. Other reference databases of protein and/or domain families include TIGRFAMs,[105] FIGfams,[210] COG[306] and Pfam.[75]

Even with sequence databases growing at an exponential rate and with ongoing expansion of annotation information in reference databases, well-studied organisms still have significant proportions of their CDSs functionally unannotated.[118,179,240,340] When predicted protein sequences cannot be functionally annotated, they are typically classified as "hypothetical" proteins, or sometimes as "conserved hypothetical" proteins if they are commonly detected in the genomes of numerous organisms.[80,82] These hypothetical sequences consist of proteins of unknown function as well as potential pseudogenes and even spurious gene predictions.[48,179]

An important question in genome-wide functional annotation is to what degree a genome (or more specifically, a proteome) can be assigned function.[273,284] Interestingly, across different bacterial species/genomes there is considerable variation in the completeness of genome annotations reported in the literature and in databases.[11,98] For example, according to the Joint Genome Institute database, well-studied model organisms such as *E. coli* K12-W3110 and *Bacillus subtilis* strain 168 have ∼86 and 81% of their proteome functionally annotated, respectively.[201] However, the proteome of *Verrucomicrobium spinosum* DSM 4136 is only 48% annotated. Ever more extreme than this is the feline parasite *Mycoplasma haemofelis*, which has functional annotations for only 19% of its proteome.[17,201] With such a wide range of annotation coverage found among bacteria, this study aimed to investigate the extent of annotation coverage across the bacterial tree of life, as well as to identify factors related to this important property of genomes.

3.2 Methods

Genome data sources

Bacterial genomes from AnnoTree[207] and their Pfam[75] and KEGG[132] annotations (gtdb_r86_bac_genomic_files.tar.gz, gtdb_r86_bac_pfam_tophits.tar.gz, and gtdb_r86_bac_ko_tophits.tar.gz, respectively) were accessed from https://data.ace.uq.edu.au/public/misc_downloads/annotree/r86/. Metadata for the downloaded genomes were retrieved from the Genome Taxonomy Database (GTDB)[235] at https://data.ace.uq.edu.au/public/gtdb/data/releases/release86/86.1/bac120_metadata_r86.1.tsv.

Gene annotation

As described elsewhere, Pfam[75] annotations were derived from Pfam v27.0[75] and applied with HMMER v3.1b1 and Pfamscan (at ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/). KEGG[132] annotations were computed based on DIAMOND v0.9.22[27] matches against the UniRef100 dataset, members of which were pre-annotated with KEGG orthology (KO) annotations. The percentage of unannotated CDSs from the Pfam and KEGG approaches for each genome was calculated by comparing the number of CDSs in the metadata file with the number of CDSs with Pfam or KEGG matches in the Pfam and KO "tophits" files from AnnoTree.[207]

Genome annotation was also performed using Prokka v1.13.7[280] with its default databases and with the rRNA and tRNA search options turned off. Mycoplasmatales (GTDB

taxonomic nomenclature that includes Entomoplasmatales and Mycoplasmatales from the NCBI taxonomic nomenclature) was analysed with translation table 4, while GTDB orders Absconditabacterales and BD1-5 (which include candidate division SR1 and '*Candidatus* Gracilibacteria' from NCBI taxonomic nomenclature) were analysed with translation table 25. The unannotated class of CDSs were identified as those containing "hypothetical protein" product names that also lacked Prokka database annotations. To analyse NCBI-derived protein annotations, protein .gpff files associated with 113,424 genome IDs in the GTDB metadata file were downloaded from NCBI's ftp server (`ftp://ftp.ncbi.nlm.nih.gov/genomes/all/`). Any protein annotation in the "product" line of the file containing the words "hypothetical", "uncharacteri(s/z)ed protein" or "unknown" were counted towards the "unannotated" fraction for that genome. The number of protein CDSs were also counted from the .gpff files for determining the percentage of unannotated CDSs. A data table containing the genome accession numbers and associated frequencies of annotated, unannotated and total gene counts produced by all three annotation pipelines is available online (`https://github.com/doxeylab/genomeAnnotationCoverage`).

Statistical analyses

Statistical analyses were performed using R v3.2.3. For all statistical tests, the logarithm of genome size was used, which resulted in distributions closer to normality. The aov() function within the R base library was used to perform analysis of variance (ANOVA) tests and ANOVA [aov(),type='III')] from the car v3.0–3 library was used to calculate analysis of covariance (ANCOVA) tests. Each ANCOVA identified a significant effect of the covariate GTDB taxonomic order on the annotation coverage, as well as a significant interference of the covariate with the effect of the independent variable. Linear regression was performed using the ggplot2 module stat_smooth(method='lm').

The PubMed June 6 2019 database was downloaded using Entrez Direct. 'Research bias' represented by PubMed mentions was determined using Entrez Direct to search PubMed for all abstracts or titles that contained a genus name (NCBI taxonomic nomenclature).

Protein lengths were derived from the predicted proteins generated by Prokka.[280]

3.3 Results

Annotation analysis

In order to explore patterns of genome annotation across bacteria, 27,372 bacterial genomes included as part of the AnnoTree database[207] were analysed. AnnoTree uses a phylogenetic tree originally derived from the GTDB[235] and allows users to visualize pre-computed functional annotations across the bacterial tree of life. Three popular approaches for functional annotation that utilize different tools and databases were used, in addition to externally computed NCBI annotations, which are describe later. (i) Prokka[280] (v1.13.7): predicted proteins were annotated by BLAST+ searches against databases of curated proteins, and by hmmscan[74] searches against the HAMAP HMMs library.[238] (ii) KEGG:[132] predicted proteins were annotated with KO numbers based on DIAMOND[27] searches against the KEGG database. (iii) Pfam:[75] predicted proteins were annotated by hmmscan searches against the Pfam-A HMM library.

Following annotation with these pipelines, for every genome, predicted CDSs were then subdivided into two categories: (i) *annotated proteins* – sequences matched to either functionally characterized or unnamed families; and (ii) *unannotated proteins* – sequences without any matches. CDSs matching protein families without an annotated molecular function were still included in the first group, since these domains may still possess limited information that can be transferred to a new sequence.

Based on Prokka results, the mean proteome annotation coverage was 52±9% (48% unannotated) (Figure 3.1a). This is expectedly lower than that reported for model organisms and higher than that reported for the low-end cases described earlier. It is worth noting that the default Prokka parameters for functional annotation are fairly strict, as only reference proteins with experimental evidence are considered for functional assignments,[280] and that annotation coverage can potentially be increased by adding custom databases of curated annotations. The KEGG-based annotation method produced similar results with 55±10% mean annotation coverage (Figure 3.1a). The third approach based on Pfam domain-based annotation produced a mean of 79±7.1% annotation coverage (Figure 3.1a), which is higher than that of the other methods. To compare the results against externally derived functional annotations, 113,424 previously annotated proteomes within the NCBI database were also examined. These proteomes had a mean annotation coverage of 79.8±10% (see section 3.2, Methods).

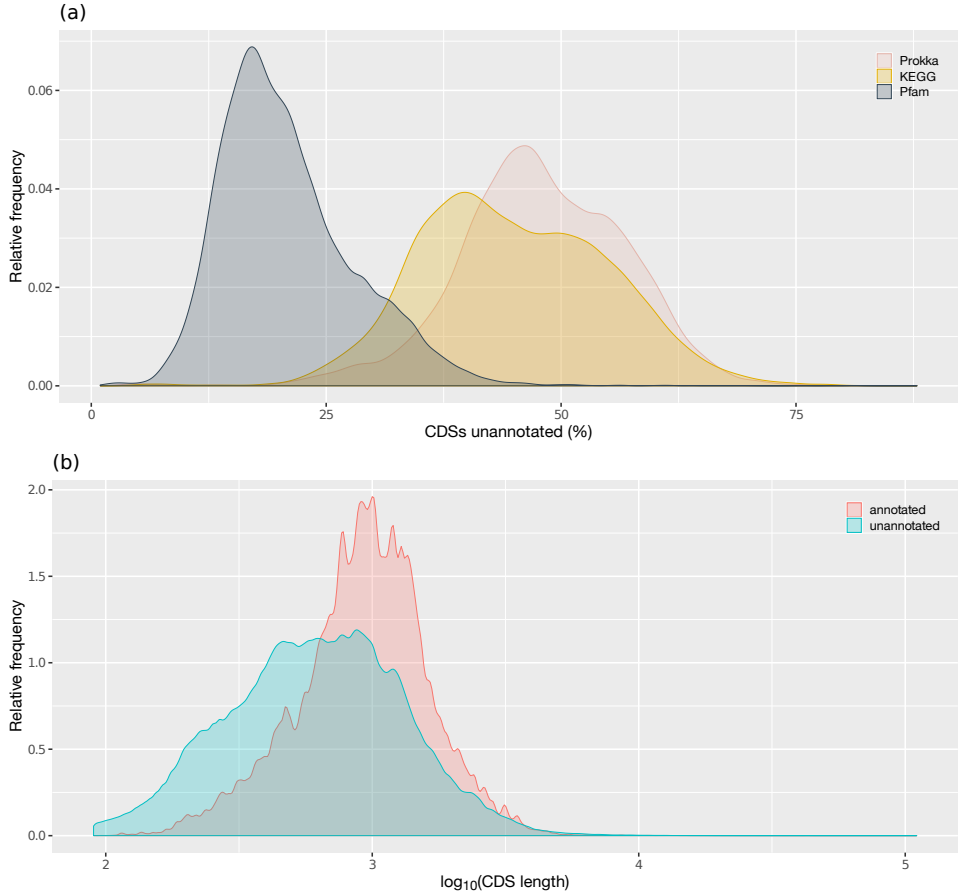

Figure 3.1: Distributions of genome annotation incompleteness across GTDB bacteria and length of annotated versus unannotated CDSs. (a) Relative frequency distribution of annotation coverage based on annotation with Prokka, KEGG and Pfam. (b) Relative frequency distribution of the length (bp) of CDSs in genomes present in AnnoTree. Annotation status was determined with this study's binary Prokka classification. The lowest length for both annotated and unannotated sequences is 90 bp, due to the length threshold in Prodigal.[116]

Another trend that was observed was that unannotated protein sequences tended to be shorter in length than annotated protein sequences (Figure 3.1b). Shorter proteins can be more difficult to annotate due to poor database coverage, lower match scores and an increased chance of being pseudogenes (one signature of pseudogenization is the accumulation of premature stop codons, which leads to shorter CDSs).[175] While it is challenging to uncover pseudogenes at such a large scale,[116,164] there was an observable

difference in the length distribution of the unannotated sequences, consistent with an increased proportion of pseudogenes. Despite this, a large proportion of the distribution was indistinguishable from that of annotated sequences (Figure 3.1b).

With all annotation pipelines analysed, extreme variation in annotation incompleteness across bacterial genomes was observed (Figure 3.1a). For example, based on protein homology searching using Prokka, annotation incompleteness ranged from 2.3% ('*Candidatus* Baumannia cicadellinicola') to 85.5% (*Mycoplasma haemofelis* Ohio2). Similar values were obtained using KEGG-based annotation, with incompleteness ranging from 3.1% ('*Candidatus* Evansia muelleri') to 87.9% (*Algoriphagus boritolerans*). Next, to further explore factors influencing this variation, the relationship between annotation coverage and various features, such as taxonomy, genome size and research bias, were examined.

Taxonomy

To study the potential taxonomic bias in genome annotations, annotation completeness was mapped onto the bacterial phylogeny, and was partitioned according to the taxonomic scheme defined by the GTDB (Figure 3.2). Differences in annotation coverage were visually apparent across the tree, and a strong degree of clade-specific patterns could be observed. This taxonomic annotation bias was supported by quantitative measurements at different taxonomic levels (Figure 3.3). Even at the phylum level, there were observable differences in genome annotation coverage between taxa (Figure 3.3a; ANOVA P value $<2\times10^{-16}$), with greater resolution revealed at every subsequent taxonomic level (Figure 3.3b). This taxonomic effect was consistent between Prokka (Figure 3.3a, b), KEGG (Figure 4a; ANOVA P value $<2\times10^{-16}$) and Pfam (Figure 4b; ANOVA P value $<2\times10^{-16}$) proteome annotations. Patescibacteria, a phylum recently formed from the highly underrepresented candidate phyla radiation associated with smaller genomes,[109,264] had the highest mean of unannotated CDSs across all three annotation systems. Spirochaetota, a smaller phylum, and Bacteroidota, found across many environments, also had higher unannotated proportions (54.8% mean and 55.7% mean, respectively). Proteobacteria and Firmicutes, the phyla of the majority of bacterial model organisms, had better annotation completeness across all three annotation systems with mean unannotated proportions of 42.6 and 42.3%, respectively. Thus, the taxonomic bias on genome annotation completeness may be in part due to what can be described as research bias or model organism bias (a larger scientific community effort towards functional characterization), which is explored further in a later section.

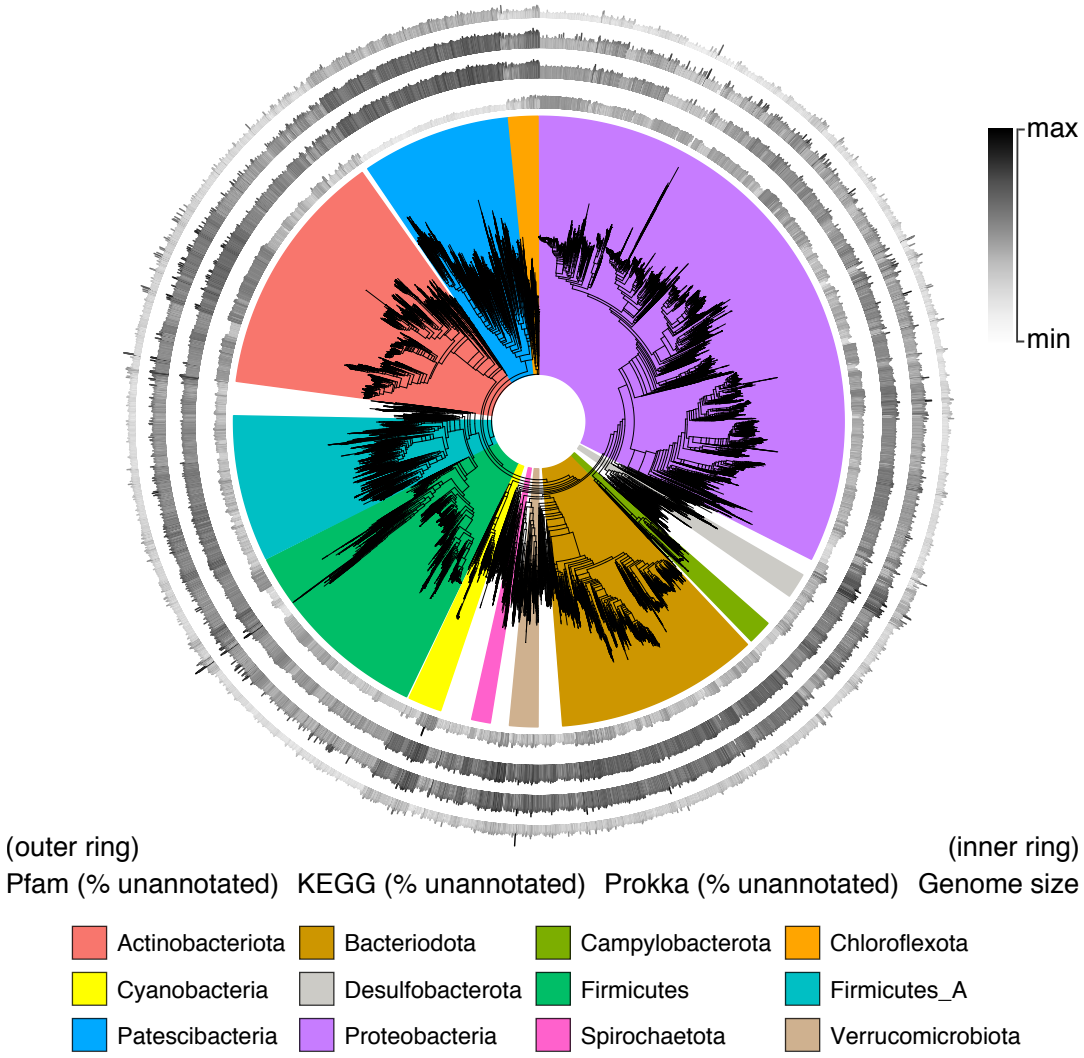

Figure 3.2: Genome annotation incompleteness across the bacterial tree of life. Annotation incompleteness has been mapped to the outer edges of the tree of life obtained from AnnoTree,[207] which was originally derived from the GTDB.[235] The height of each bar (and colour) depicts traits (annotation incompleteness and genome size), which have been normalized separately for each metric. For annotation incompleteness, the gradient goes from 0% (minimum) to 100% (maximum). Four metrics are shown, including annotation incompleteness as determined using Pfam (outer ring), followed by that determined using KEGG, that determined using Prokka and genome size (inner ring). *This figure was designed in collaboration with Benjamin Tremblay.*

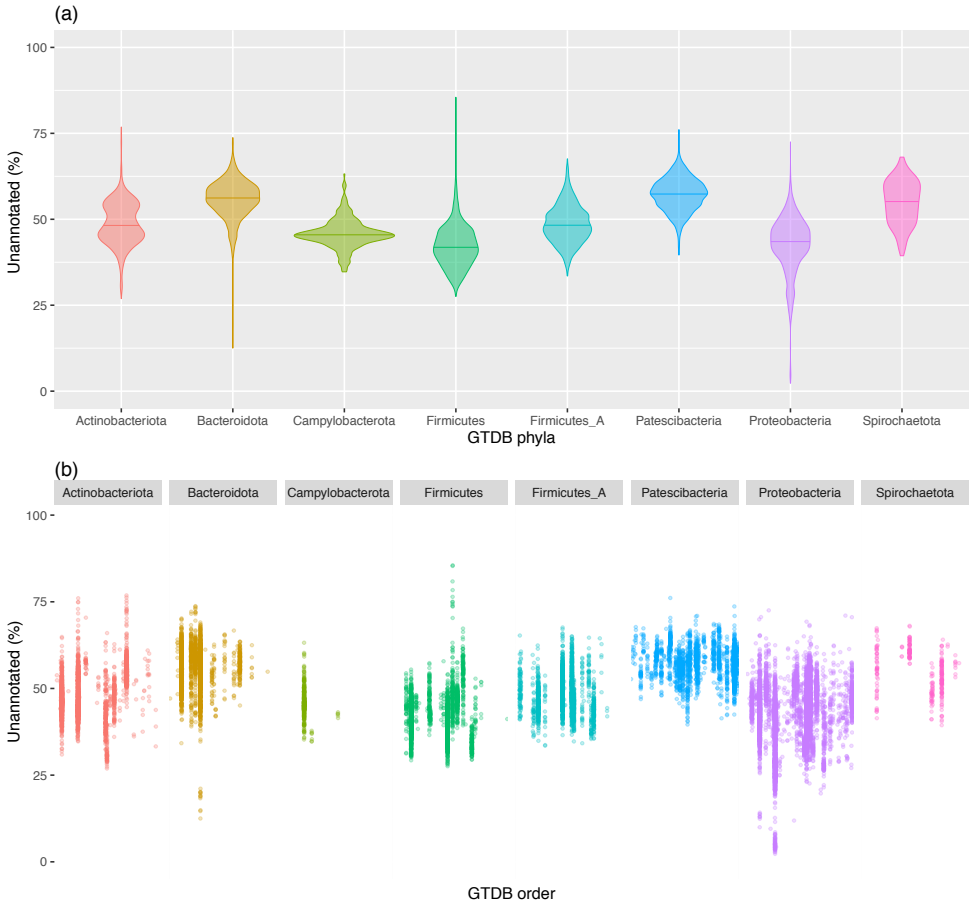

Figure 3.3: Distributions of genome annotation coverage subdivided by taxonomic group. Genomes were annotated using Prokka with default parameters (see section 3.2, Methods). Only the most common phyla from the GTDB[235] are shown. (a) Taxonomic separation by phyla. (b) Taxonomic separation by order. Labelled orders are using GTDB taxonomic nomenclature.

Genome size

Genome size, a trait related to taxonomy (as evident in Figure 3.2), also appeared to affect the annotation coverage of genomes. Even without accounting for the confounding impact of taxonomy, a relationship between genome size and genome annotation completeness was visible (Figure 3.4a). A closer look at this phenomenon within individual phyla revealed

an even clearer picture of this trend, where larger genomes were associated with a larger proportion of unannotated proteins [Figure 3.4b, 5a (KEGG) and 5b (Pfam)].

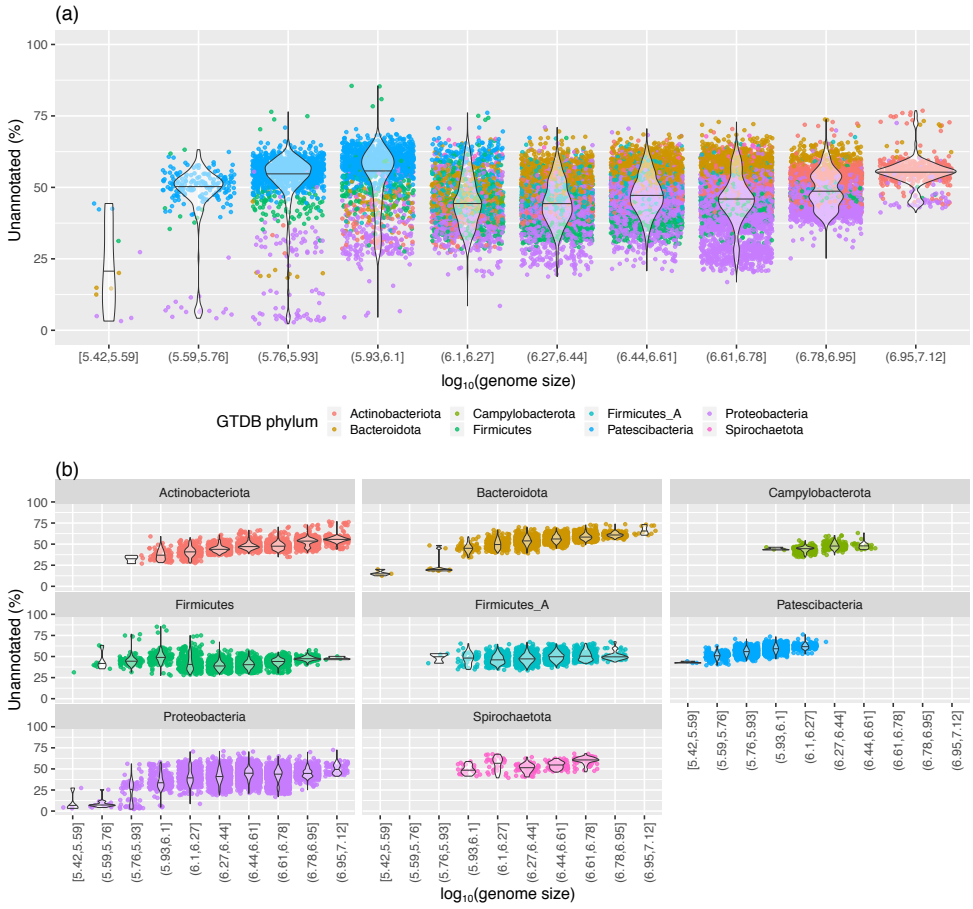

Figure 3.4: Relationship between genome size (bp) and Prokka genome annotation coverage. The $\log_{10}$(genome size in bp) is binned into 10 distinct bins to better display the trend. Square and open brackets indicate intervals that include and do not include the adjacent number, respectively. (a) Only the most common GTDB phyla are shown. (b) The most common GTDB phyla are displayed separately.

An interesting case demonstrating this relationship is the phylum Firmicutes. Although at a phylum level, the effect of genome size on annotation completeness was not entirely clear (Figure 3.4), when subdivided into lower taxonomic levels (Figure 3.5), the trend was readily apparent. That is, different taxonomic groups within the Firmicutes possessed

distinct distributions of genome completeness and each was also influenced by genome size. For example, Mycoplasmatales, RF39 and RFN20 (GTDB taxonomic nomenclature[235]) possess relatively small genomes, but had a high fraction of unannotated CDSs. Yet, within these taxonomic groups, genome size positively correlated with the level of annotation incompleteness. Thus, these cases illustrate how annotation incompleteness is driven by multiple factors.

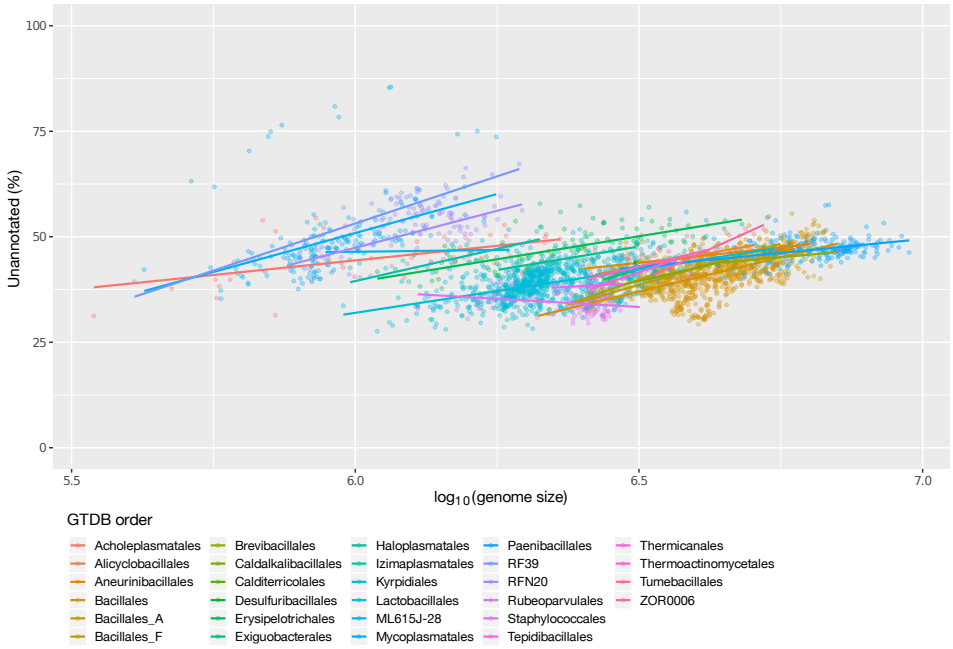

Figure 3.5: Prokka genome annotation coverage of Firmicutes (GTDB taxonomy) against genome size. Trend lines are displayed for each taxonomic order.

Consistent with these observations, an ANCOVA test controlling for the GTDB taxonomic order revealed a significant relationship between genome size and annotation incompleteness for Prokka, KEGG and Pfam annotations (P value=3.6×10^{-5}, 2.5×10^{-3} and 1.1×10^{-4}, respectively). The protein annotations in the NCBI database also showed a significant difference between taxonomic phyla (ANOVA P value $<2.2\times10^{-16}$; Figure 3.6a) and a relationship with genome size (ANCOVA, while controlling for GTDB taxonomic orders, P value=2.3×10^{-10}; Figure 3.6b). Since the largest factor influencing genome size variation in bacteria is the gain and loss of "accessory" genes,[20,316] it can be reasoned that this trend may reflect an increased difficulty in functional annotation of accessory genes

versus "core" genes (see Discussion). Since genome size is also related to other factors such as GC content, the correlation between GC content and annotation completeness was also examined. However, this relationship was not as clear (Figure 3.7) and was non-significant when controlling for taxonomy (ANCOVA P values of 0.6, 0.85 and 0.33 for Prokka, KEGG and Pfam, respectively).

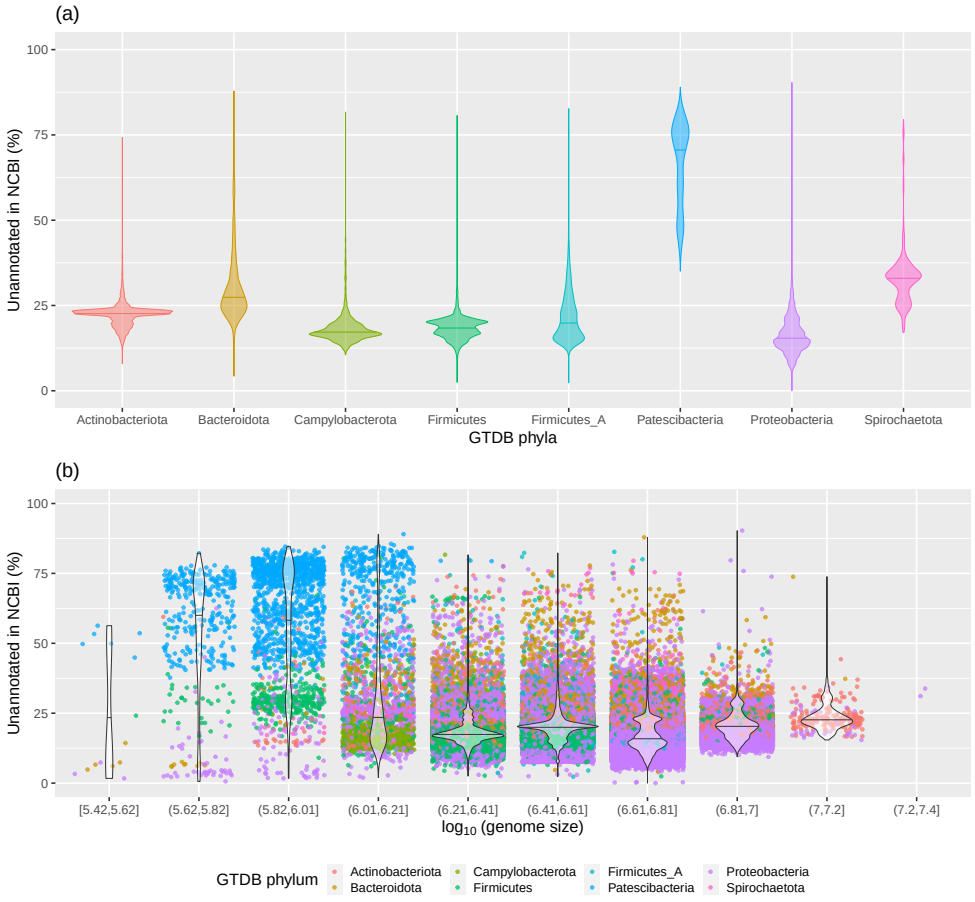

Figure 3.6: Genome annotations from NCBI (a) Taxonomic separation of genome annotation coverage by phyla using the taxonomic nomenclature from the Genome Taxonomy Database (GTDB). Only the most common GTDB phyla are shown. (b) Relationship between genome size and genome annotation coverage in NCBI. $\log_{10}$(genome size) is binned into 10 distinct bins to better display the trend. The most common GTDB phyla are displayed.

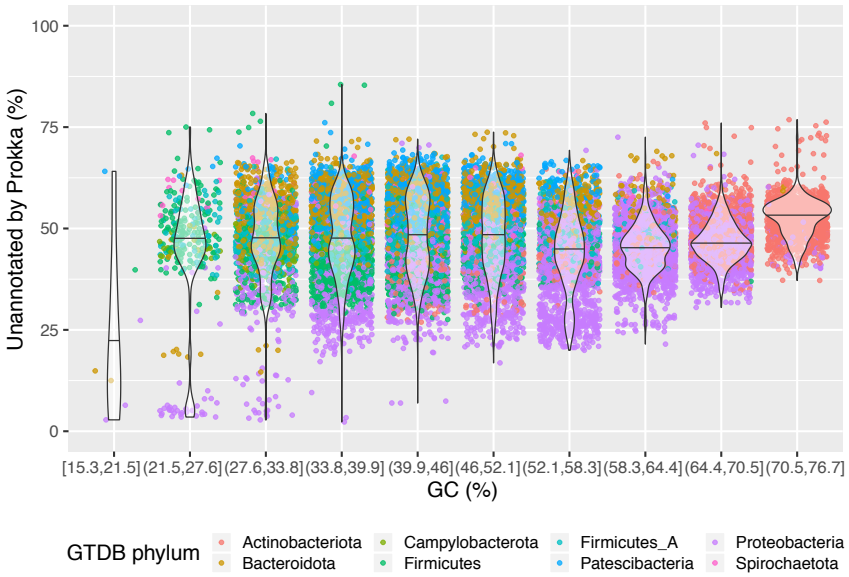

Figure 3.7: Relationship between GC percentage and genome annotation coverage by Prokka. Percent GC is binned into 10 distinct bins to better display the trend. The most common GTDB phyla are displayed.

Research bias

To explore the effects of research bias on annotation coverage, the number of times each genus was mentioned in abstracts or titles within the PubMed database was counted, and their genome publication dates were also recorded. Here, NCBI taxonomic nomenclature was adopted as those naming conventions are more common in literature. Genera with over 75,000 mentions (such as *Escherichia*, *Staphylococcus* and *Pseudomonas*) generally had a greater annotation coverage compared to genera that occurred less frequently in publications [Figs 3.8a (Prokka), 3.8b (KEGG), 3.8c (Pfam)]. Similarly, genomes released before 2003 tended to have a greater proportion of annotated CDSs [Figures 3.9a (Prokka), 3.9b (KEGG), 3.9c (Pfam)]. However, these effects were only apparent in the extreme cases (i.e. model organisms associated with extreme publication volume). Moreover, the majority of genera in this uppermost bracket were Proteobacteria and Firmicutes, consistent with the earlier analysis of taxonomic influence on genome annotation coverage.

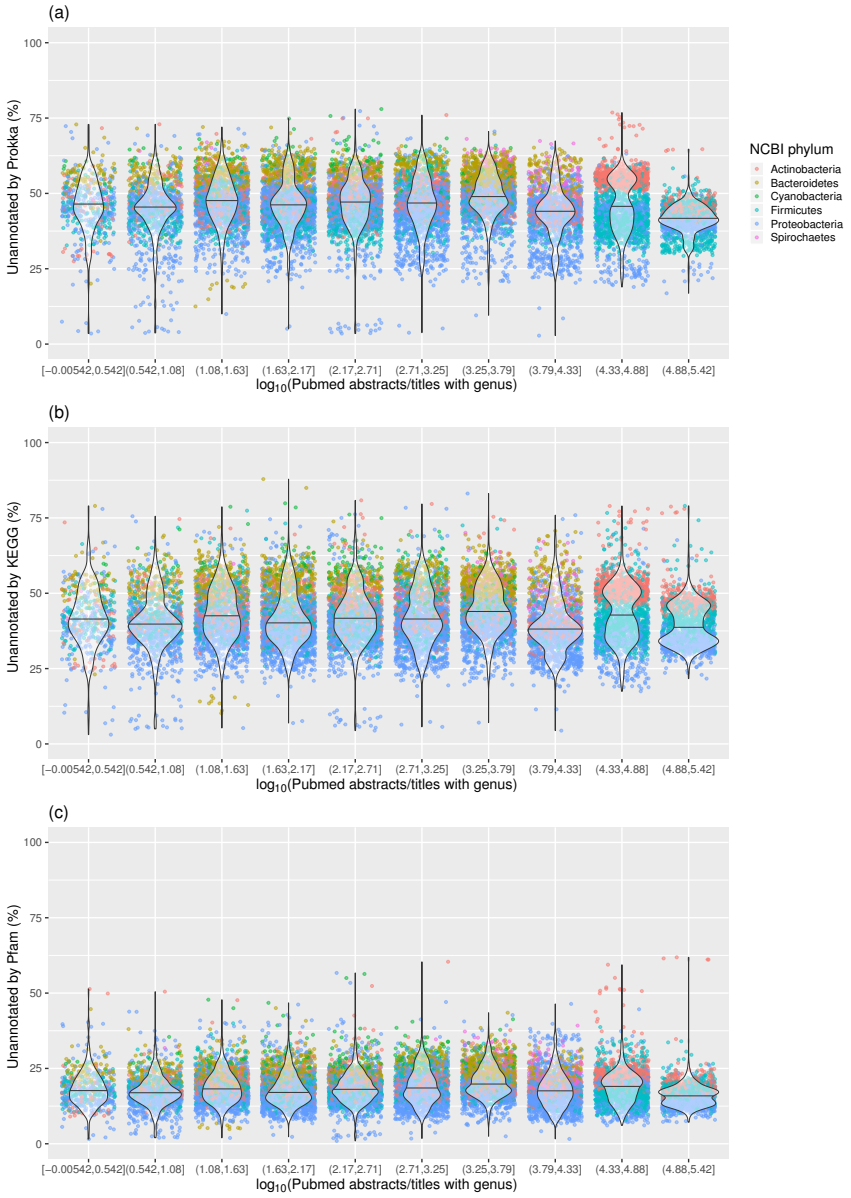

Figure 3.8: $\log_{10}$(Pubmed mentions of genera in titles or abstracts) by genome annotation coverage. The Pubmed mentions are binned into 10 distinct bins to better display the lack of any significant trend. Only the most common phyla according to NCBI taxonomic nomenclature are shown. (a) Prokka genome annotation; (b) KEGG genome annotation; (c) Pfam genome annotation.

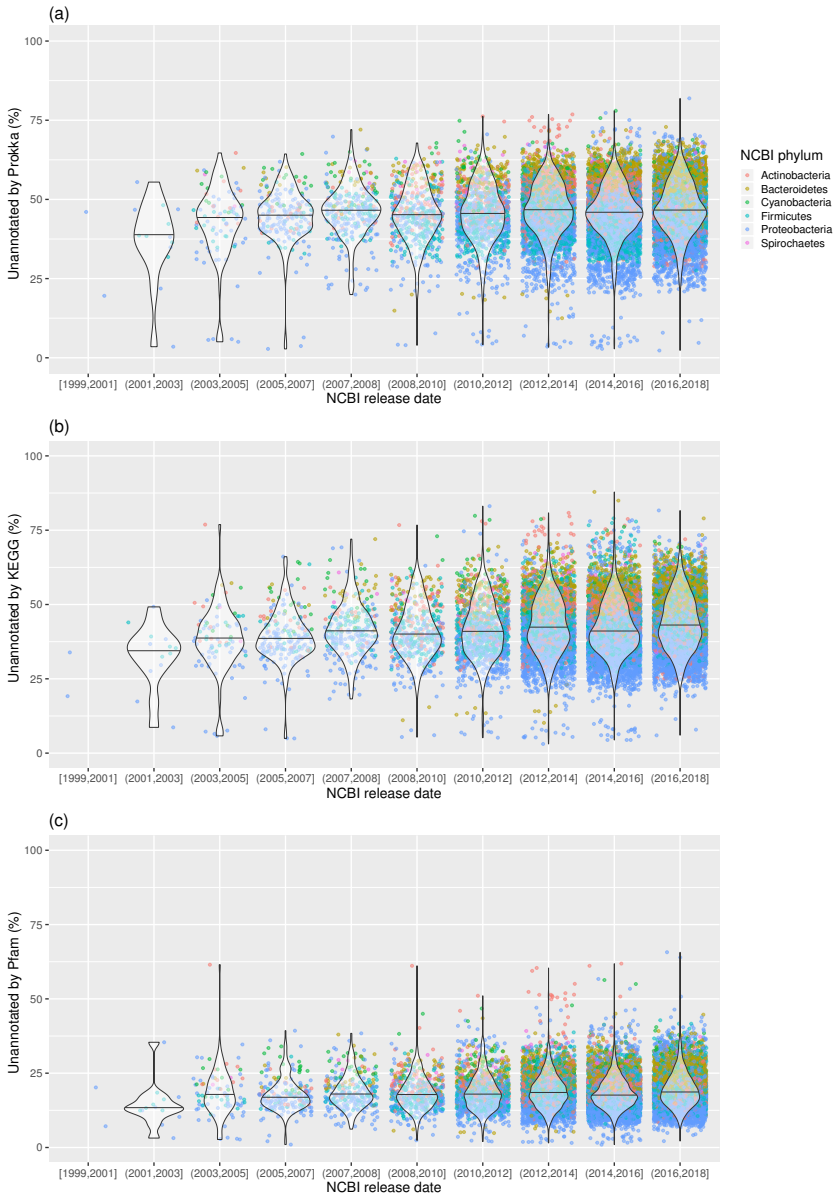

Figure 3.9: NCBI genome release date by genome annotation coverage. The NCBI release date has been binned into 10 distinct bins to better display the lack of any significant trend. Only the most common phyla according to NCBI taxonomic nomenclature are shown. (a) Prokka genome annotation; (b) KEGG genome annotation. (c) Pfam genome annotation.

To explore this phenomenon further, the distributions of genome annotation completeness was examined while subdividing by taxonomy, mapping only the most heavily studied taxa onto their respective lineages. This clarified the effect of research bias since model organisms (e.g. *E. coli*, *Bacillus subtilis*, *Mycobacterium tuberculosis*) stood out as being among the best annotated genomes in their respective taxonomic groups (Figure 3.10). There were, however, some exceptions to this phenomenon; within the Proteobacteria, a noticeable group of organisms had annotation completeness well exceeding that of *E. coli*. These organisms included endosymbionts with highly reduced genomes, such as *Buchnera aphidicola*, an endosymbiont of aphids, '*Candidatus* Blochmannia' (an ant symbiont), *Wigglesworthia* (a symbiont of tsetse flies) and others. This may be due to multiple factors, including an increased proportion of core or 'essential' functions associated with "minimal genomes" and, thus, easier-to-annotate processes in reduced genomes of parasitic organisms,[143,144,223] as well as the close evolutionary relationship of these genomes to the heavily studied model organism *E. coli*.[85,219]

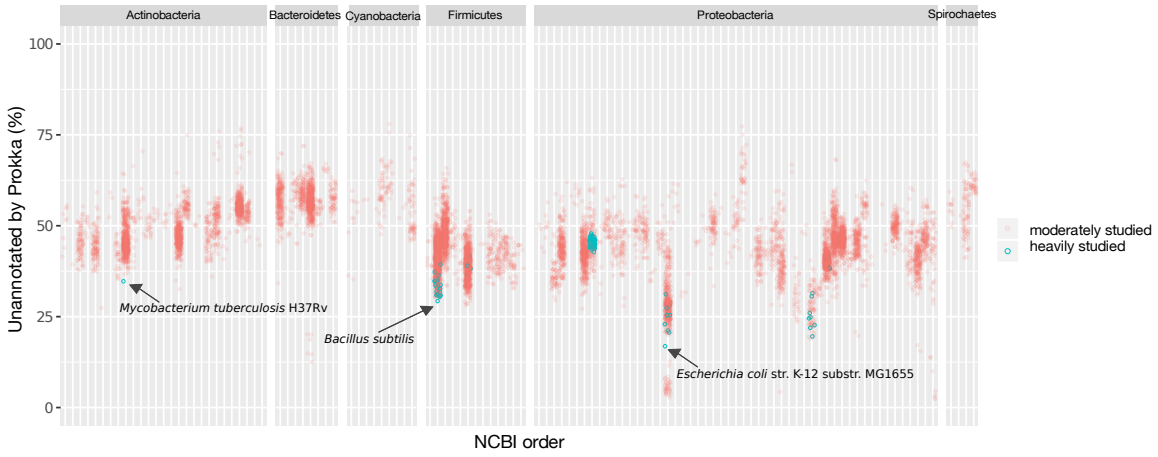

Figure 3.10: Influence of research bias on genome incompleteness. The top six most abundant phyla are shown and each is further subdivided by taxonomic order. Orders appear as distinct vertical columns. Heavily studied genomes, as measured by PubMed abstract counts per species (>15,000), show a marked reduction in unannotated sequences (annotated with Prokka) compared to other moderately studied genomes (500–1000) in their taxonomic group. Other heavily studied species include *Listeria monocytogenes*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Helicobacter pylori*, *Klebsiella pneumoniae*, *Haemophilus influenzae* and *Pseudomonas aeruginosa*. It must be noted that the terms "heavily" and "moderately" studied organisms are relative, are associated only with the frequency of published papers, and do not account for the true impact of publications and other work that contribute toward functional annotation.

3.4 Discussion

As genomes shape our understanding of organism function, not only individually but also as a community, it is important to assess our ability to annotate genomes across the tree of life and understand the factors that influence this important property. Here, GTDB[235] and AnnoTree[207] were used in combination with various annotation pipelines to perform a comprehensive assessment of genome annotation coverage across the bacterial phylogeny. This analysis revealed extreme variation in genome annotation coverage across and within taxonomic groups. Numerous factors appear to influence levels of annotation completeness across bacterial genomes, including annotation method, taxonomy, genome size and research

bias.

Overall, the mean annotation completeness of bacterial genomes varied from ∼52% for methods requiring high-stringency matches to reference proteins, to 79% for more sensitive domain-based annotation methods. While domain-based annotation methods produced the highest proportion of annotated CDSs, these estimates of annotation coverage may be not be realistic, since the mere presence of a domain in a predicted protein sequence is not necessarily sufficient to assign function, and consideration of domain architecture is more informative. Also, although three annotation pipelines were performed separately, a combination of methods would have likely resulted in greater annotation coverage, as observed in previous studies.[98] However, the goal of this study was not to optimize annotation coverage across bacteria, but rather to assess it using standard, commonly used pipelines.

Taxonomy was an important factor influencing genome annotation completeness. Some of this taxonomic bias may stem from research bias, whereby genomes that are more closely related to those of model organisms possess a greater chance of being successfully annotated based on detectable homology. Indeed, phyla containing many model organisms were found to have, on average, more annotated CDSs than their understudied counterparts. In addition, within broader taxonomic groups, specific model organisms (e.g. *E. coli*) stood out as outliers in terms of annotation coverage. This pattern was also demonstrated for other highly studied species as determined based on publication volume (occurrences of species names in PubMed abstracts and titles).

This analysis also uncovered an interesting, significant anticorrelation between genome size and annotation coverage, which was consistently detected across a range of taxonomic groups. Larger genomes showed lower annotation coverage, which suggests a relative lack of annotations and functional characterization concerning accessory proteomes. Indeed one study in 2007[254] even found a "weakly significant" positive correlation between the average genome size predicted in metagenomes and the number of novel protein families discovered. One interpretation of this finding is that core proteomes contain more essential and widely studied processes, resulting in increased genome annotation coverage. In contrast, the accessory gene content within a pangenome of a species may include a more diverse repertoire of genes,[255] including those derived from prophages[20] and integrated elements, which are known to be particularly challenging for annotation.[37] The dynamic accessory genome of a species may also possess increased pseudogene content, resulting in shorter (truncated) and potentially divergent ORFs that are harder to assign function through homology searches. The observed difference in the length distribution of annotated versus unannotated CDSs is consistent with this idea.

The reduced genomes of symbionts and parasites are extreme examples of how factors related to genome size may affect annotation completeness. In this analysis, reduced genomes were found at both ends of the spectrum of annotation completeness. Within the Firmicutes, for example, some parasitic genomes in the Mycoplasmatales were poorly annotated. This may be a result of increased pseudogene content, which is thought to accumulate in the reduced genomes of some organisms due to genetic drift.[20,152,153,218] However, the reduced genomes of endosymbiotic Proteobacteria such as *Buchnera aphidicola* were extremely well annotated, consistent with previous analyses,[254,323] which may be due to efficient purging of genes and pseudogenes over a longer evolutionary timescale with retention of core processes. These core or essential functions are in turn easier to annotate bioinformatically [for previous papers on the minimal genome concept see references by Mushegian (1999) and Koonin (2000)[143,223]]. Their increased annotation completeness may also in part benefit from their close relationship with a model organism (*E. coli*).

Finally, this analysis highlighted certain lineages (e.g. the Patescibacteria within the candidate phyla radiation group) as possessing a higher level of hypothetical gene content. This may reflect the presence of highly divergent gene families that escape the detection limits of standard homology-based annotation, or this may be indicative of new protein functions, metabolic activities and biological traits. To assign function to these sequences, the use of powerful/sensitive methods for protein function prediction may be useful; these include remote-homology detection and structure prediction approaches.[19,179] Methods for function prediction will also benefit from continual expansion of Gene Ontology[1] and other controlled vocabularies.[41,312] In addition to sequence-to-function methods, a complementary "function-to-sequence" type of approach may also be useful, where a required parts list of functions is used to guide the search for potential gene functions.[292] Finally, our ability to assign function computationally to these and other bacterial genomes is inherently tied to the quantity and quality of experimentally derived functional information contained within references databases. Continued experimental characterization of understudied organisms and hypothetical/novel gene families will be critical to widen the net of annotation coverage and lead to more accurate genome analyses and functional insights derived from genomic and metagenomic studies.

[1] 704 new "biological process" and "molecular function" GO terms were added between Jun. 2019 and Aug. 2020. Retrieved from http://geneontology.org/stats.html

Chapter 4

Inferring biological associations for conserved domain families

As just investigated in Chapters 2 and 3, a significant fraction of genome sequence data is currently unannotatable with homology-based methods. For this fraction of proteins, is there a way to use alternative non-homology based annotation methods to uncover functional information and mine for novel proteins of interest? Many of these hypothetical genes and proteins of unknown function have been compared and amassed into protein or domain families, just as characterized proteins have been organized into families based on sequence, structure and/or functional similarity. In the absence of experimental data for these conserved but uncharacterized protein families, alternate methods for teasing out functional information can be attempted. Detecting associations between these families and other biological traits (such as particular environments, taxonomic lineages, and phenotypes) has the potential to provide functional insights and give a broad range of data for researchers interested in targeting domains for experimental characterization. Here, a comprehensive analysis of 17,929 domain families within the Pfam database is provided, including over 4,000 domains of unknown function (DUFs), all scored based on various biological and statistical attributes. Statistically significant associations for a substantial fraction of DUFs and other protein families of unknown function were uncovered, providing a guide for future experimental characterization.

4.1 Introduction

Domains are modular units of proteins that adopt specific three-dimensional structures and functions. Related domains can be grouped by sequence homology into domain families, which have a common evolutionary ancestry, and adopt similar structures and functions.[309] Domain families have been bioinformatically classified by databases such as CATH,[287] the NCBI Conserved Domain Database,[194] Interpro,[115] and Pfam.[73]

The Pfam database v32.0 contains a total of 17,929 domain families. These can be further classified into "clans", sometimes referred to as domain superfamilies. In v32.0, 22% (4049) of all domain families in Pfam are defined as "domains of unknown function" or DUFs. DUFs can be recognized bioinformatically as sequence families in genomes but have not been assigned function. Many DUFs are essential in bacteria[94] and thus are an important target for functional characterization.[221] Beyond DUFs, additional collections of uncharacterized protein families have been constructed, including ORFan proteins derived from metagenomes[179] and the recently generated Function Unknown Families (FunkFams) dataset.[340] DUFs and other collections of uncharacterized protein families are a fascinating target for bioinformatic analysis, since many potentially encode novel biochemical activities and biological functions.[177] At the same time, they are extremely challenging cases for function prediction because by nature they tend to lack detectable homology to other families, and so cannot be assigned function by standard tools such as BLAST.

As an alternative to homology-based functional annotation methods, functional insights into DUFs may be obtained by detecting *statistical associations* between DUFs and various biological traits. By analyzing the distribution of protein families across the genomes of different species, it is possible to uncover several types of associations. First, a protein/gene family may show an association with a particular *taxonomic lineage* (in which case it may be called a "signature" gene[101]), which may help place that family under a certain biological context.[142] With improved taxon sampling of genomes across the tree of life, lineage-specificity of gene/protein families can potentially be measured at a greater resolution than ever before.[207] Second, the *abundance* of a protein family across different environments can help provide functional context.[68,72,154,179,317,357] Detecting protein families that associate with certain environments has become increasingly possible through the availability of metagenomic datasets from a growing diversity of biomes and associated environmental metadata. For example, Ellrott et al.[68] used an automated computational procedure to identify protein families specific to the human gut microbiome, and discovered 835 sequence families *de novo* in metagenomic data. Subsequent experimental characterization of some of these protein families have revealed functions that are important for microbial physiology in the human gut environment. Third, presence/absence of a protein family may show

a statistical association with a certain *phenotype*. For example, numerous studies have compared protein family abundance between pathogenic and non-pathogenic genomes to detect those that may play roles in virulence.[57,78,354] Recent studies have also identified phenotypic associations for bacterial genes, including genes of unknown function, *en masse* through genome-wide screens using transposon sequencing.[247]

In this work, several association-based methods have been applied to analyze the full set of 17,979 domain families in Pfam, with a focus on DUFs to gain insights into their biology. For each domain family, its distribution across available genomes and metagenomes was examined to measure a variety of biologically relevant characteristics including: abundance, taxonomic breadth and specificity, environmental association, and pathogen association (Figure 4.1). By performing multiple association-based analyses, I was able to uncover statistically significant biological associations for a large number of protein domain families. An online database (virfams.uwaterloo.ca) is provided to allow researchers to explore these pre-computed statistical analyses of Pfam domain families, providing biological and statistical information to guide future experimental studies.

4.2 Methods

Abundance and taxonomic breadth

The NCBI sequence database domain alignments were sourced from `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.full.ncbi` (Pfam v.32.0;[75] retrieved Feb.9, 2019). The proteins that were aligned to Pfam domains and the total number of hits were taken from this file. An environmental average of the normalized adjusted family size for each domain (see the Environmental association section 4.2 of Methods) present in at least 5% of the selected samples used to determine environment-association was calculated. The taxon ID and taxonomy of proteomes with Pfam domain matches were retrieved from `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/pfamA_ncbi.txt.gz` and `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/taxonomy.txt.gz`, respectively (Pfam v.32.0; retrieved Oct.16, 2018). The percentage of species each domain is present in, and the corresponding percentage for the Genus, Family, Order, Class, Phylum, Kingdom and Superkingdom taxonomic levels are included on `virfams.uwaterloo.ca`. Spearman rank correlation between the different abundance measures (percentage of species, environmental average, and protein hits in NCBI) was calculated with corr in R v3.3.3.

Environmental association

Metagenomic assemblies and raw reads were taken from public repositories (NCBI Sequence Read Archive: SRA045646 and SRA050230; NCBI Assembly: GCA_900245835.1, GCA_000208365.1, GCA_900245825.1, GCA_000496495.1, GCA_002059125.1, GCA_900216645.1, GCA_002059105.1, GCA_002059145.1, GCA_002059065.1, GCA_002059085.1, GCA_002058945.1, GCA_002058925.1, GCA_002059005.1, GCA_002059045.1, GCA_002059025.1, GCA_002058985.1, GCA_002058965.1, GCA_002058885.1, GCA_002058905.1, GCA_002058845.1, GCA_002058865.1, GCA_002058825.1, GCA_900216805.1, GCA_900216795.1, GCA_900215965.1, GCA_900215875.1, GCA_900245845.1, GCA_900216675.1, GCA_900216935.1, GCA_900291615.1, GCA_900291665.1, GCA_900216775.1, and GCA_900216765.1; CAMERA: CAM_PROJ_GOS; EBI: PRJEB6337; MGRAST: 4504797.3 and 4504798.3; http://www.bork.embl.de/~arumugam/Qin_et_al_2010/). No samples smaller than 1,000,000 bp were used. The raw reads from the human gut studies (Qin et al., 2012, 2014) were processed and assembled with the following procedure. Any read that aligned to the human genome (GCA_000306695.2) with Bowtie 2 (v2.2.9)[157] default settings was removed (along with its pair). Quality trimming was performed by sickle v1.33. The reads were assembled with Megahit v1.0.6-3-gfb1e59b[169] with default settings. The raw reads from the Global Ocean Sampling study[269] were not assembled as the reads, which were sequenced with a modified form of Sanger sequencing, were already quite long. FragGeneScan v1.30[262] was used to detect CDSs in the samples. To remove any putatively spurious CDSs, any sequence with greater than 40% repetitive sequence, detected by segmasker from the BLAST package v2.2.28+, was removed. Annotation with PfamScan (version updated on Feb. 28, 2017) using HMMER3 v.3.1b2[65] against the Pfam database v32.0 (retrieved Oct. 16, 2018) with a threshold of 1×10^{-3} was performed on the remaining sequences. The annotated region of each metagenomic sequence (aligned with a Pfam domain) was clustered with CD-HIT v4.6.8[171] to 99% similarity for each sample within each set of domain matches. This removed redundant domain matches to give a measure of adjusted family size of the domain families for each sample. To normalize to sample size, the adjusted family member count was divided by the number of base pairs in the assembly and multiplied by 1,000,000. A ratio of samples across each human gut study analyzed was chosen to maximize regional diversity while making the sample size in each environment more comparable. All 14 healthy samples from the Spanish cohort[248] were used, and then 34, 16, and 16 healthy samples from the Danish cohort,[248] the Chinese cohort originating from Peking University Shenzhen Hospital, Shenzhen Second People's Hospital and Medical Research Center of Guangdong General Hospital,[249] and the Chinese cohort originating from the First Affiliated Hospital of Zhejiang University[250] were randomly selected, respectively. However, in per-domain figures all human gut samples have been added back in for visual comparison. Domains not present in greater than 95% of the

selected samples were excluded. Domains where at least one environment (soil, marine or human gut) showed significant differences based on the normalized adjusted family size were determined with the Kruskal-Wallis test. P values were adjusted with p.adjust using the Benjamini-Hochberg model. The logarithm of the normalized adjusted family size (base 10) and the subsequent scaling across the domain hits (scale) was done in R v3.3.3 for the heatmap. Enrichment of DUFs in the environment-associated domain sets compared to the background frequency of DUFs in Pfam was tested using the binomial test (pbinom in R). To determine GO term enrichment within the environment-associated domain sets, a Pfam to GO term map was retrieved from http://geneontology.org/external2go/pfam2go (last updated February 12, 2019). The frequency of GO terms in domains associated with one of the three environments (soil, marine and human gut) and the frequency of GO terms corresponding to other Pfam domains present in at least 5% of the selected samples were compared with the hypergeometric test (phyper in R), with P values again adjusted with the Benjamini-Hochberg model.

Lineage association

The sensitivity and precision of the Pfam domain distribution across the NCBI taxonomy system was calculated from the Pfam files pfamA_ncbi.txt and taxonomy.txt (see the Abundance and taxonomic breadth section 4.2 in Methods). The total number of proteomes within any one taxonomic group is based on the taxon IDs in the pfamA_ncbi.txt file. These scores were calculated for the most common taxon (presence/absence counts of a domain hit per proteome) in each domain family at the Superkingdom, Kingdom, Phylum, Class, Order, Family and Genus taxonomic levels. The best taxonomic level to describe a domain's lineage specificity was chosen based on the F1 score 4.1. In the case of a tie between taxonomic levels, the higher level in the taxonomic hierarchy (e.g. Superkingdom) was given preference. If the majority of proteomes that the domain was present in did not have any classification at a certain taxonomic level, this taxonomic level would not be considered for "best taxonomic level." The enrichment of DUFs in extreme lineage-specific cases was determined in the same way as with the environmental-associated domain set.

Pathogen association

354 proteomes in Pfam were designated as bacterial pathogens based on PATRIC (https://www.patricbrc.org)[335] bacterial pathogens with metadata relating them to disease and a manually curated set of pathogens from Dhillon et al.[49] Enriched pathogenic domains were

detected with the hypergeometric test (phyper in R) based on the number of pathogenic proteomes in Pfam the domain is present in compared to the non-pathogenic bacterial proteomes in Pfam the domain is present in. P values were FDR corrected with p.adjust using the Benjamini-Hochberg model. The enrichment of DUFs in pathogen-associated domains was calculated in the same way as with the environment-associated domain set. The frequency of the pathogenesis GO term in domains identified as pathogen-associated and other Pfam domains present in bacterial proteomes were compared with the hypergeometric test (phyper in R). Eukaryotic-like domains in bacterial pathogens were identified as being most common in eukaryotic proteomes as well as pathogen-associated (P value < 0.05) or with hits in bacterial pathogens but without hits in non-pathogen proteomes. I expanded past the pathogen-associated domain set in this case, to capture domains present in a low number of proteomes (which meant they weren't statistically significant) that seemed like promising "mimicry" candidates.

Additional filters

All data was taken from Pfam v.32.0 (files retrieved on Oct.16, 2018). A list of Pfam families with PDB structures was taken from `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/pdb_pfamA_reg.txt`. Domain architectures were sourced from `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/architecture.txt`. Predicted transmembrane and disordered regions in sequences with Pfam domain alignments were retrieved from `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/other_reg.txt`. Overlap of predicted transmembrane or disordered regions with an annotated domain was evaluated by comparing to `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.regions.uniprot.tsv`. The standard deviation for domain family percentage disorder was calculated using std from the NumPy package v1.16.1. Domains that were prioritized for structural feasibility had no representatives in the PDB, an average across the domain family members of less than 10% of the domain sequence predicted to be disordered, less than 10% of their members with a predicted transmembrane region (anywhere along the protein), and less than 10% of their members with transmembrane-domain overlap.

4.3 Results and Discussion

Abundance and taxonomic breadth of domain families

All 17,929 protein domain families in the Pfam 32.0 release were analyzed, including 3961 DUFs and 88 UPFs[1] (from now on collectively referred to as DUFs) (Figure 4.1). To gain insights into the abundance of DUFs and other Pfam families, three different datasets were surveyed. I examined: N_{NCBI}, the number of protein family members in the NCBI sequence database; $N_{species}$, the percentage of species containing the domain family in the Pfam proteome collection; and N_{meta}, the number of non-redundant matches in a diverse dataset of metagenomes (Figure 4.2). DUFs were abundant in all three datasets, present in a total of 13,201,304 sequences (see Table 4.1 for the most abundant DUF and other Pfam families). The abundance distributions for DUFs overlapped with that of other domains in Pfam, but DUF families tended to be smaller in size (Figure 4.2). Despite a strong expected sampling bias towards Proteobacteria, Actinobacteria, and Firmicutes in the NCBI database, all three metrics correlated with one another ($r = 0.86$ for N_{NCBI} vs $N_{species}$, $r = 0.69$ for N_{NCBI} vs N_{meta}, and $r = 0.71$ for N_{meta} vs $N_{species}$).

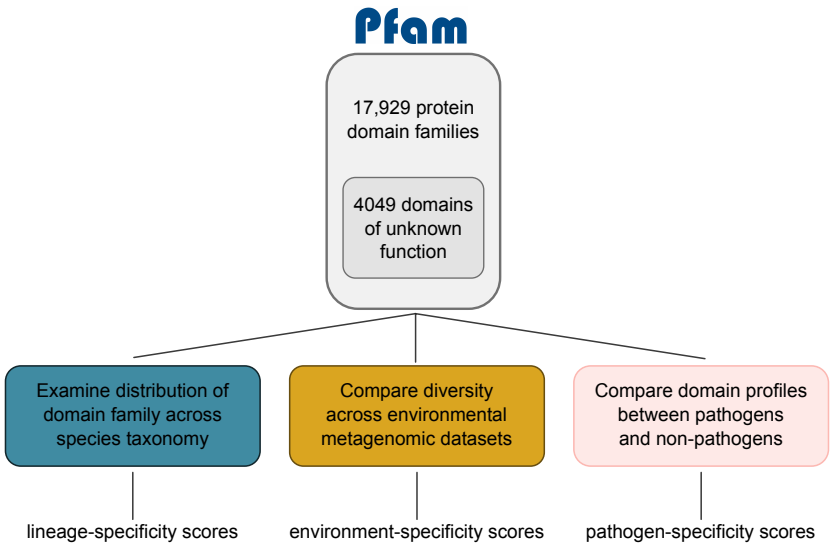

Figure 4.1: Overview of computational framework for DUF categorization and functional prioritization.

[1]UPFs stand for uncharacterized protein families and were created separately by Swiss-Prot.

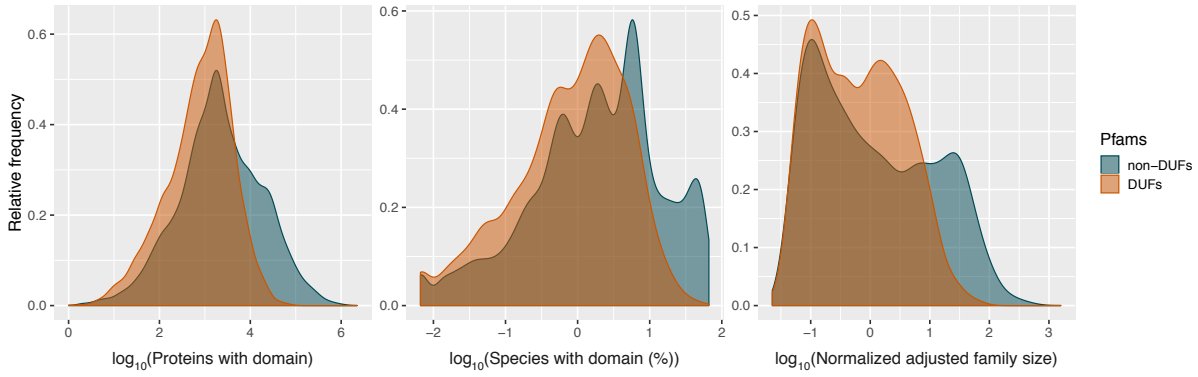

Figure 4.2: Domain abundance distributions. Frequency histograms show the number of NCBI proteins containing the domain, the % species in Pfam proteomes with the domain, and the average of the normalized adjusted family size of non-redundant hits across a diverse set of metagenomic samples. All Pfam domain families, DUFs and the other Pfam families that are not DUFs, are represented here.

Table 4.1: Top five most abundant domains and DUFs. Families were ranked using three different measures of abundance: number of proteins with the domain in the NCBI, percent of species with the domain in Pfam proteomes, and the average of the normalized adjusted family size across all environmental samples.

	Proteins in NCBI	**Presence in species (%)**	**Environmental average**	**Average abundance rank**
Pfam domains (non-DUFs)				
ABC_tran	2,236,463	62.74	1579.20	2.00
HATPase_C	1,136,217	62.69	618.05	4.00
Helicase_C	641,004	66.24	290.12	16.33
HTH_3	449,128	62.80	313.15	22.67
Glycos_transf_2	458,343	61.35	426.08	25.00
DUFs				
DUF21	63,537	52.22	56.53	507.67
UPF0004	58,608	50.42	67.68	528.33
UPF0051	40,910	40.03	65.50	795.00
UPF0020	72,562	36.65	22.57	1030.67
UPF0054	25,360	52.15	33.08	1058.67

Although DUFs could be identified in only 3-6% of total open-reading frames, they make up 22% of all protein domain families in Pfam and therefore constitute a sizeable fraction

of the domain diversity in proteomes. In an analysis of 15,803 proteomes, DUFs were found to make up to 16%, 9%, and 100% of unique domain families found in prokaryotes, eukaryotes, and viruses, respectively (Figure 4.3). The abundance of DUFs and other proteins of unknown function underscores the need to detect associations between domain families and various biological attributes.

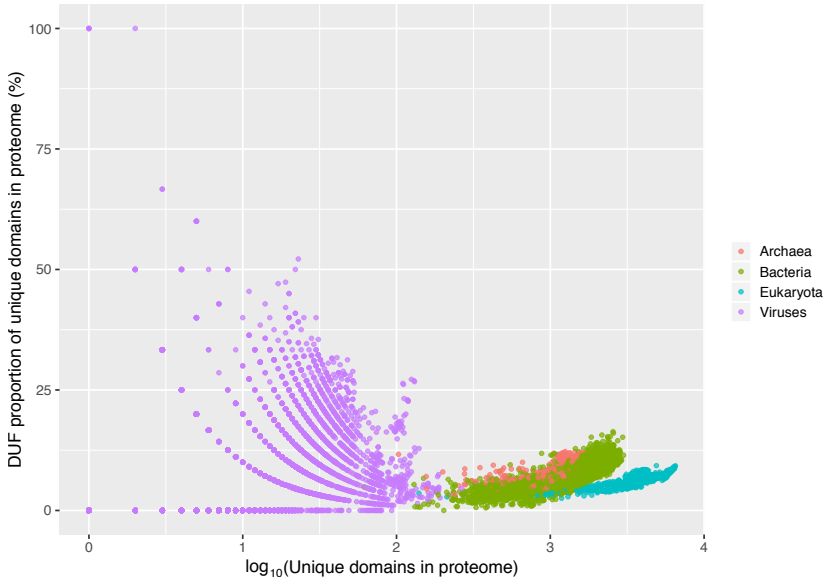

Figure 4.3: The percentage of unique DUFs in each Pfam proteome compared to the total number of unique Pfam domains in each Pfam proteome.

Environmental association

To evaluate the environment-association of Pfam and DUF families, their metagenomic abundance across three major environments were compared: human gut, terrestrial (soil), and marine ecosystems. A total of 392 metagenomic assemblies from global soil, marine, and human gut samples were collected from public repositories and databases and annotated with Pfam models. For each protein domain, its *adjusted family size*, the number of unique (99% redundancy threshold) domain occurrences in each metagenome assembly, was computed. In order to avoid sample size bias, the number of human gut samples was reduced while maximizing their regional diversity (see 4.2, Methods). The Kruskal-Wallis test was used to determine which domains have significantly differing adjusted family size in

at least one of the three environments. A stringent threshold of $P_{adj} < 1{\times}10^{-15}$ was found to capture domains with extreme differences in adjusted family sizes between environments. This identified a set of 4357 domains with strong environment-specificity, including 1050 in soil, 1246 in marine systems and 2061 in human gut (see heatmap in Figure 4.4a). Interestingly, soil-associated families showed a greater tendency to occur in marine (top right) and human gut (bottom left) environments (Figure 4.4a). Within the set of 4357 environment-associated families, there were 1056 DUFs. DUFs were slightly enriched in soil-associated families (1.13-fold, $P = 0.016$) and more strongly enriched in human gut associated families (1.20 fold, $P = 1.37{\times}10^{-5}$). However, DUFs were underrepresented in marine-associated families (0.82 fold, $P = 3.0{\times}10^{-4}$). Example DUFs with extreme environmental specificity are shown in Figure 4.4b and top-scoring Pfam and DUF families are listed in Table 4.2 and 4.3.

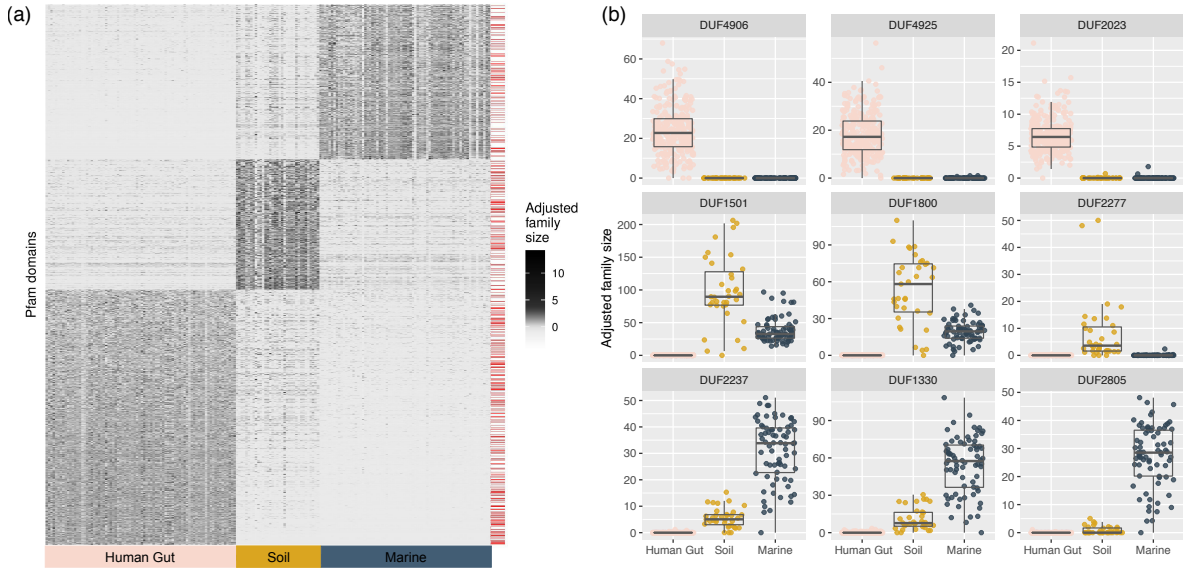

Figure 4.4: Detected Pfam families with strong environmental associations. (a) Abundance heatmap of Pfam families with significant environmental-specificity scores ($P_{adj} < 1{\times}10^{-15}$). The adjusted family size was calculated as the logarithm of the normalized adjusted family size (base 10), scaled across the domain values. The red lines on the right-side of the plot denote DUF rows. (b) Selected DUF families with strong environment-specificity scores. Plotted are the per-sample distributions of normalized adjusted family size in three environments: human gut, marine, and soil.

Table 4.2: Top five environment-associated domains from soil, marine, and human gut metagenomes. Shown in the table are the normalized average adjusted family sizes of each domain family in each environment. The list is ranked by the P value and then by the normalized average adjusted family size of the environment the domain family is associated with.

	P_{adj}	Soil	Marine	Human Gut
Soil-associated				
Ycel	3.27×10^{-33}	50.96	16.56	0.60
Virul_fac_BrkB	8.95×10^{-33}	113.75	13.41	33.69
zf-HC2	9.67×10^{-33}	72.42	1.29	17.35
DUF1501	1.10×10^{-32}	97.85	38.51	0.00
GerE	1.22×10^{-32}	242.00	21.89	109.79
Marine-associated				
T4_neck-protein	2.13×10^{-33}	0.07	31.79	0.03
UvsY	2.13×10^{-33}	0.07	22.92	0.01
Gp5_OB	2.13×10^{-33}	0.04	17.31	0.00
Phage-Gp8	2.14×10^{-33}	0.03	43.29	0.01
DUF2237	2.14×10^{-33}	5.61	30.92	0.01
Human gut-associated				
DUF4906	2.13×10^{-33}	0.00	0.00	23.80
DUF4925	2.13×10^{-33}	0.00	0.04	17.62
LPD16	2.13×10^{-33}	0.05	0.00	14.15
Cys_rich_VLP	2.13×10^{-33}	0.01	0.00	10.28
Lipocalin_8	2.13×10^{-33}	0.06	0.02	7.87

Table 4.3: Top five environment-associated DUFs from the soil, marine, and human gut. Shown in the table are the average adjusted family sizes of each domain family in each environment. The list is ranked by the P value and then by the normalized average adjusted family size of the environment the domain family is associated with.

	P_{adj}	Soil	Marine	Human Gut
Soil-associated				
DUF1501	1.10×10^{-32}	97.85	38.51	0.00
DUF1800	1.35×10^{-32}	53.35	19.41	0.00
DUF2277	1.52×10^{-32}	7.99	0.04	0.00
DUF2382	1.94×10^{-32}	24.83	0.00	0.01
DUF488	2.42×10^{-32}	29.67	0.96	9.72
Marine-associated				
DUF2237	2.14×10^{-33}	5.61	30.92	0.01
DUF1330	4.54×10^{-33}	10.95	54.19	0.08
DUF2805	7.59×10^{-33}	0.87	27.34	0.00
DUF4815	7.90×10^{-33}	0.55	95.72	0.09
DUF2061	1.08×10^{-32}	0.73	15.74	0.01
Human gut-associated				
DUF4906	2.13×10^{-33}	0.00	0.00	23.80
DUF4925	2.13×10^{-33}	0.00	0.04	17.62
DUF2023	2.14×10^{-33}	0.02	0.03	6.55
DUF4317	2.91×10^{-33}	0.28	0.05	25.03
DUF5119	2.91×10^{-33}	0.07	0.03	16.50

Next, the top function enrichments for environment-specific domain families were explored. Marine-specific protein families were enriched in GO terms related to photosynthesis, consistent with the higher proportion of cyanobacteria in marine environments rather than in soil and gut environments (Table 4.4). The top function enrichments for soil-specific protein families included transposase activity and heme binding, which was associated with nine heme binding domains, including catalase and numerous cytochrome enzymes. The top enriched GO terms for human-gut associated protein families included the phosphoenolpyruvate-dependent sugar phosphotransferase system, O-glycosyl hydrolase activity, and carbohydrate metabolic process (Table 4.4). Also among the top human-gut specific domain families are domains with known roles in host adhesion/colonization and gut microbial metabolism (Table 4.4). For example, DUF4906 (PF16249; ranked 1) appears to be a homolog of the fimbrial proteins Mfa2 (PF08842) and P_gingi_FimA (PF06321), known to be involved in cell adhesion. Fimbrillin_C (PF15495; ranked 11) is also associated with P_gingi_FimA. These domain families appear to be members of a broader superfamily of fimbrial proteins[342] in the human gut microbiome, and may be responsible for cell adhesion

to the human gut epithelium. The identification of the carbohydrate-binding module CBM32 (PF18344; ranked in top 10) also makes sense from the perspective of microbial carbohydrate metabolism in the human gut. Finally, the identification of Maff2 (PF12750) within the top 10 domains also agrees with previous literature since this protein family is associated with tetracycline resistance cassettes that are extremely abundant in the human gut microbiome.[137]

Table 4.4: GO term enrichment in environment-associated domain sets. Fold change is within domains with a Pfam annotation in the environmental samples.

GO term	**Environment-associated domains with GO term**	**Non-environment-associated domains with GO term**	**Fold change**	P_{adj}
Soil-associated				
transposase activity	9	2	50.4	9.85×10^{-6}
transposition, DNA-mediated	11	3	41.1	7.95×10^{-7}
heme binding	9	15	6.72	4.38×10^{-2}
oxidation-reduction process	43	184	2.62	1.36×10^{-4}
Marine-associated				
cytochrome complex assembly	6	0	Inf	3.08×10^{-4}
photosystem II reaction center	6	1	55.7	1.26×10^{-3}
photosynthesis, light reaction	5	1	46.4	1.02×10^{-2}
flavin adenine dinucleotide binding	9	3	27.8	6.55×10^{-5}
photosystem I	8	3	24.7	3.19×10^{-4}
oxidoreductase activity, acting on the CH-CH group of donors	5	2	23.2	2.63×10^{-2}
tricarboxylic acid cycle	5	2	23.2	2.63×10^{-2}
nickel cation binding	5	2	23.2	2.63×10^{-2}
photosynthesis	25	16	14.5	1.15×10^{-12}
photosystem II	11	8	12.8	1.10×10^{-4}
Human gut-associated				
mismatch repair	8	0	Inf	3.81×10^{-4}
mismatched DNA binding	6	0	Inf	8.88×10^{-3}
spore germination	5	0	Inf	3.95×10^{-2}
phosphoenolpyruvate-dependent sugar phosphotransferase system	11	4	14.3	8.53×10^{-4}
cobalamin biosynthetic process	9	4	11.7	1.17×10^{-2}
hydrolase activity, hydrolyzing O-glycosyl compounds	22	16	7.17	6.50×10^{-6}
carbohydrate metabolic process	39	35	5.81	9.65×10^{-10}

Lineage association

To score Pfam families based on their lineage-specificity, a metric derived from the F1 statistic, a common measure for assessing the performance of binary classifiers, was implemented. Application of this concept to a taxonomic lineage enables the F1 score to measure the ability of the lineage to predict the occurrence of a domain family within a taxonomic system or phylogenetic tree. The F1 score combines two terms called precision and sensitivity (also called recall) which together give a measure of accuracy.

$$2 \cdot \frac{\text{sensitivity} \cdot \text{precision}}{\text{sensitivity} + \text{precision}} \tag{4.1}$$

Here, precision was used to measure the degree to which a domain family is conserved (or retained) within members of a lineage, and sensitivity to measure the degree to which a domain family is unique to a lineage, respectively. At a given taxonomic level, these two terms for lineage L were computed as follows:

$$\text{precision} = \frac{\text{number of proteomes in lineage L containing the domain}}{\text{number of proteomes in lineage L}} \cdot 100\% \tag{4.2}$$

$$\text{sensitivity} = \frac{\text{number of proteomes in lineage L containing the domain}}{\text{total number of proteomes across all lineages containing the domain}} \cdot 100\% \tag{4.3}$$

The combined metric becomes the lineage-specificity score of lineage L. Each protein domain is assigned a lineage-specificity score at each taxonomic level. To facilitate interpretation of the data, the highest-scoring F1 score and taxonomic level it is calculated from were assigned to their respective domain families.

The distribution of sensitivity and precision scores for all Pfam families is shown in Figure 4.5a. Domain families with low sensitivity or precision scores are non-lineage-specific. For example, CRISPR_Cas6 (PF10040) which is scattered across the tree of life[189] has a high sensitivity score at the Superkingdom level for Bacteria since many of the proteomes that contain this domain are bacterial (90.19%), but has a low precision score at this level (6.80%) since only a small fraction of bacterial proteomes possess this domain. This produces a low F1 score of 12.65. Conversely, the domain CrgA (PF06781) is highly lineage-specific within the Class Actinobacteria as it has both a high sensitivity (99.65%)

and precision score (97.58%) at this level, resulting in a high F1 score of 98.61. Visualization of their phylogenomic distributions across the bacterial tree of life (Figure 4.5a) confirms this prediction.

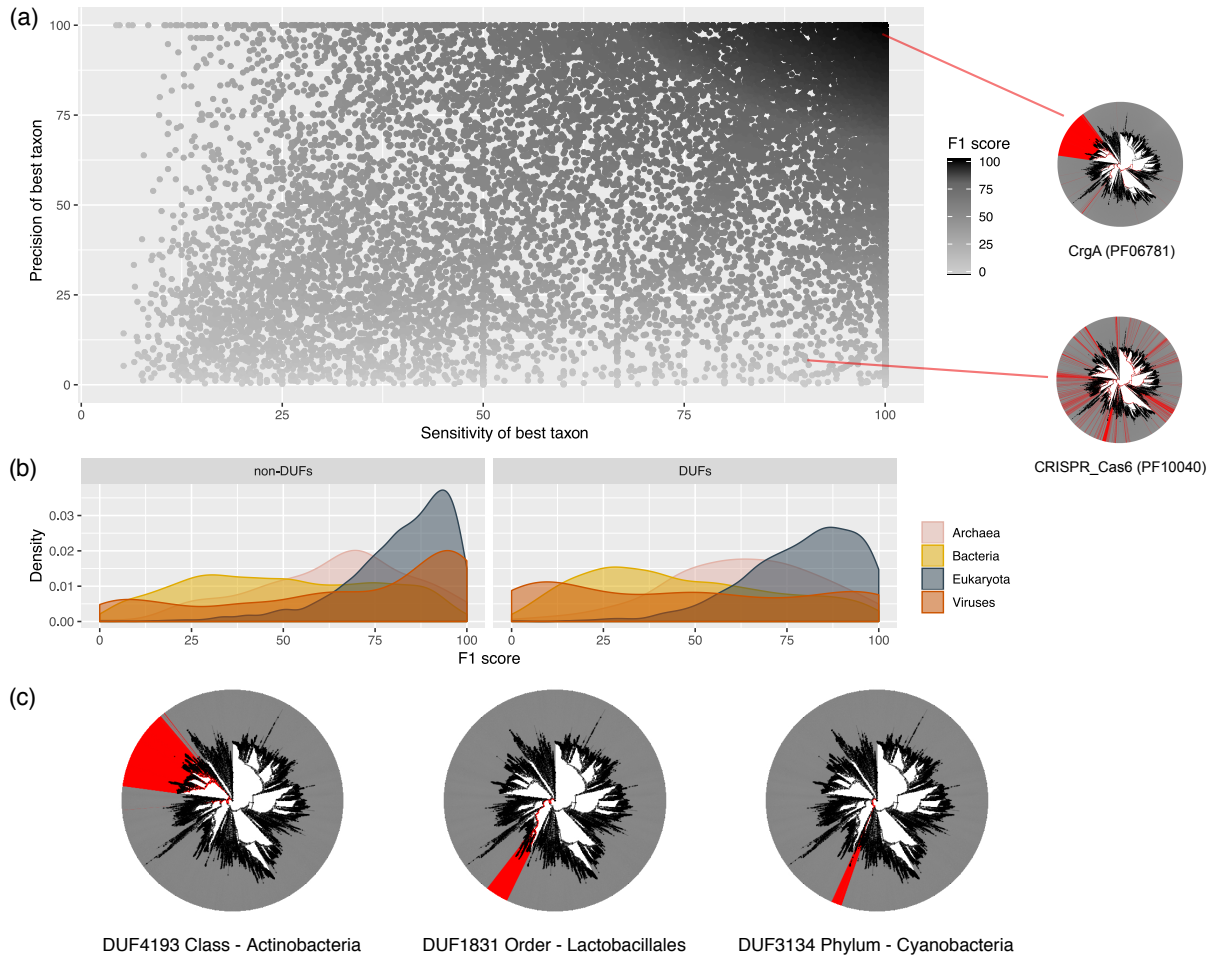

Figure 4.5: Measuring lineage-specificity of protein domain families. (a) Precision (percent of proteomes within a lineage containing a domain family) and sensitivity (percent of domain family members within a lineage) metrics are plotted for all 17,772 families in the Pfam proteome collection. For each family, the precision and sensitivity scores are plotted for the lineage that maximizes the best combination of the two scores according to the F1 metric. (b) Distributions of lineage-specificity scores (F1 statistic) for all DUF and non-DUF families partitioned by taxonomic group. The phylogenomic distributions of two families (CRISPR_Cas6 and CrgA) were generated using Annotree and highlight examples of a highly scattered and lineage-specific family, respectively. (c) Phylogenomic distributions of top-scoring lineage-specific DUF families in bacteria and archaea. Top DUF families with unique, non-redundant phylogenomic distributions are shown, with visualizations generated using AnnoTree. The bacterial trees from Annotree are shown here at a genome level. The taxonomic level and taxon that best describes the domain's lineage specificity is listed.

To identify straightforward examples of lineage-specific families for further analysis, domain families with extreme levels of precision and sensitivity (both with scores $\geq 95\%$) were selected. This resulted in the identification of 981 lineage-specific Pfam families including 178 DUFs, which was not significantly different ($P = 0.21$) from their background frequency in Pfam. Most of these lineage-specific families are of eukaryotic origin (649) compared to prokaryotic origin (120). The high frequency of lineage-specific eukaryotic domain families is also evident from the precision and sensitivity distributions (Figure 4.5b and Figure 6). Tables 2 and 4.5 list the top 20 Pfam and DUF families ranked by lineage-specificity, respectively. Visualization of phylogenomic distributions of Pfam domains with a high F1 score using a different taxonomic system (AnnoTree) confirmed that this metric is indicative of high levels of phylogenetic specificity across a wide variety of lineages (see examples in Figure 4.5c).

Table 4.5: Top five lineage specific DUFs in Eukaryota, Archaea, Bacteria, and Viruses. DUFs were ranked by the F1 score of their best taxonomic level and the number of proteomes in which they are present, excluding domains present in less than 20 species.

	Proteomes with domain	**Best taxonomic lineage**	**Best sensitivity**	**Best precision**	**F1 score**
Eukaryota					
DUF1191, DUF1639, DUF3444, DUF3475, DUF4370, DUF668	81	Streptophyta	100	100	100
DUF148, DUF2650	49	Chromadorea	100	100	100
DUF5380	21	Rhabditida	100	100	100
UPF0506	20	Platyhelminthes	100	100	100
DUF639	82	Streptophyta	98.78	100	99.39
Bacteria					
DUF3208, DUF3809	26	Deinococcus-Thermus	100	100	100
DUF4193	1153	Deinococcus-Thermus	99.57	99.05	99.31
DUF4191	1145	Actinobacteria	99.83	98.62	99.22
DUF3039	1151	Actinobacteria	99.30	98.62	98.96
DUF3043	1134	Actinobacteria	99.82	97.67	98.74
Archaea					
DUF2208	42	Thermoprotei	90.48	100	95.00
DUF2192	42	Thermoprotei	88.10	97.37	92.50
DUF655	377	Archaea	98.67	86.92	92.42
DUF1699	29	Methanosarcinales	89.66	92.86	91.23
DUF357	348	Archaea	99.14	80.61	88.92
Viruses					
DUF816	60	Baculoviridae	100	98.36	99.17
DUF682, DUF844, DUF884	59	Baculoviridae	100	96.72	98.33
DUF1477	38	Alphabaculovirus	94.74	100	97.30
DUF1247	37	Alphabaculovirus	94.59	97.22	95.89
DUF918	35	Alphabaculovirus	97.14	94.44	95.77

Pathogen association

Next, I sought to rank Pfam families based on their association with a phenotype of interest (bacterial pathogenicity). First, a dataset of 354 pathogen and 7897 non-pathogen bacterial proteomes was constructed based on the PATRIC database and metadata from Dhillon et al. The pathogens came from a wide-range of hosts including humans, animals and plants.[49] For each Pfam domain, its statistical overrepresentation in pathogen proteomes was calculated using a hypergeometric test (see Figure 4.6a). To account for proteome-specific duplications,

which could bias the enrichment statistic, only binary presence/absence of the domain in a proteome was assessed. 2007 significantly enriched ($P_{\text{adj}} < 0.05$) domains (including 517 DUFs) were identified in the pathogenic set out of 11,299 domains with hits in bacterial Pfam proteomes (Figure 4.6a). Among pathogen-associated domains, DUFs were slightly enriched (1.16-fold change, $P = 4.4\times10^{-4}$). As expected, pathogenic lineages such as the Enterobacteriaceae had the highest frequency of pathogen-associated domains per proteome (Figure 4.7).

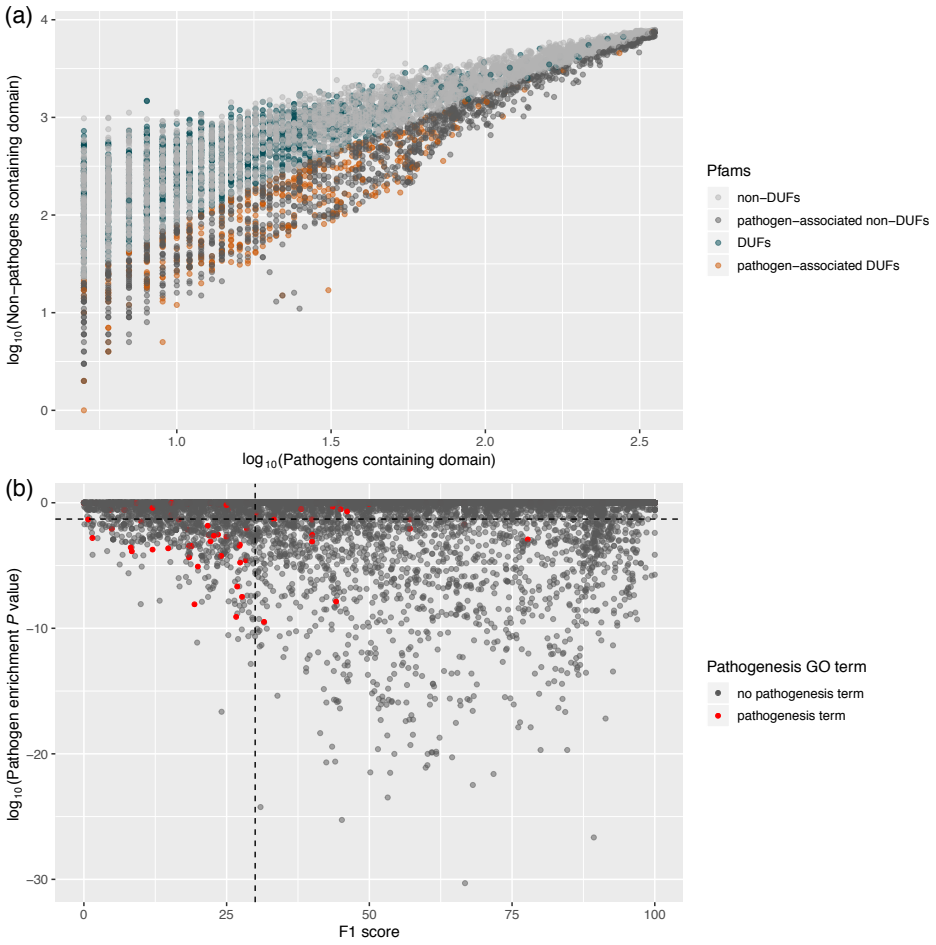

Figure 4.6: Scatterplots of Pfam domain pathogen-association. (a) Pfam domain presence in pathogen versus non-pathogen proteomes, with significant pathogen-associated patterns shown. Only domains present in more than four pathogens were included. (b) Trends in pathogenesis GO term annotation shown with respect to enrichment in pathogen proteomes and a measure of lineage specificity, the F1 score. The horizontal dotted line is at $\log_{10}(0.05)$, showing the pathogen-association threshold. The vertical dotted line is at an F1 score of 30.

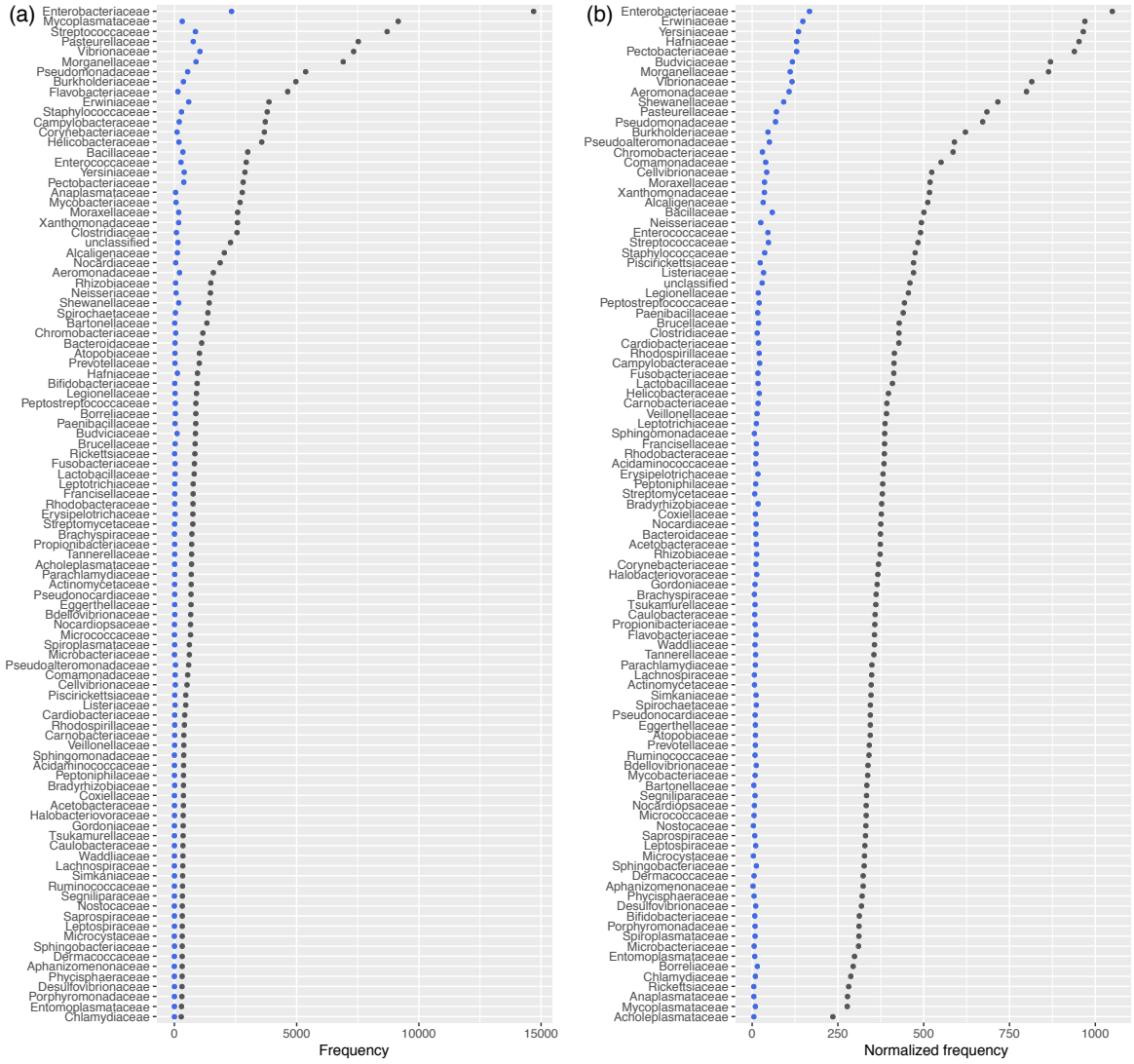

Figure 4.7: Distribution of family level taxonomic groups within the pathogen-enriched domain set. (a) Total instances and (b) instance rate normalized by number of pathogen proteomes in that taxonomic family. Each plot is ordered based on frequency. DUFs are in blue.

Also consistent with expectation, the GO term "pathogenesis" was significantly over-represented in this set of Pfam domains (2.67-fold above background frequency in Pfam

database, $P = 1.50 \times 10^{-6}$). Interestingly, when examining both pathogen-association and lineage-specificity (F1 score) together, a trend for domains with the GO term "pathogenesis" was observed. That is, domains with F1 scores <30 were 9-fold enriched in "pathogenesis" compared to domains with higher F1 scores (Figure 4.6b). This is consistent with the idea that many pathogen-associated protein families (i.e., virulence factors) tend to undergo horizontal gene transfer and therefore may be less likely to exhibit high lineage-specificity.[103] This also illustrates the utility of combining lineage information and pathogen-association for virulence factor discovery.

Among the top-scoring pathogen-associated Pfam families are numerous domains from known toxins and virulence factors (Table 4.6). For example, three of the four domains within the botulinum neurotoxin protein (Toxin_trans, Peptidase_M27, Toxin_R_Bind_N), a protein family previously thought to be restricted to *Clostridium* but recently demonstrated to be more broadly distributed,[190,191,356] occur in the list of top 20 pathogen-associated Pfam families (Table 4.6). As revealed by their lineage-specificity scores, these domains are more broadly distributed phylogenetically than other domain families in the top 20, which tend to have narrow lineage-specificity (e.g., *Mycoplasma* associated proteins). I propose that there are likely numerous novel virulence factors to be found within the 517 detected pathogen-associated DUFs.

Table 4.6: Top 20 pathogen-associated domains. Families were ranked by fold change and are present in at least five pathogens. Proteomes with domain include all proteomes in the Pfam proteome collection (not just bacteria) and so may be larger than the sum of the Pathogens and Background (bacterial) columns.

	Pathogens	**Background (bacterial)**	P_{adj}	**Fold change (bacterial)**	**Proteomes with domain**
DUF1410	5	1	0	111.54	6
DUF1600	5	2	0	55.77	7
DUF5378	5	2	0	55.77	7
Leader_Trp	5	2	0	55.77	7
MFS_Mycoplasma	25	11	0	50.70	36
DUF31	31	17	0	40.68	48
DUF5385	9	5	0	40.15	14
Lambda_Kil	5	3	0	37.18	49
Staphopain_pro	5	3	0	37.18	8
Toxin_trans	5	3	0	37.18	9
Lipoprotein_X	21	13	0	36.04	34
DUF2618	6	4	0	33.46	10
DUF2684	6	4	0	33.46	10
FinO_N	6	4	0	33.46	11
Mycoplasma_p37	24	16	0	33.46	40
DUF2714	22	15	0	32.72	37
Lipoprotein_10	22	15	0	32.72	37
Strep_SA_rep	7	5	0	31.23	15
Peptidase_M27	5	4	0	27.88	10
Toxin_R_bind_N	5	4	0	27.88	13

Eukaryotic-like domains in bacterial pathogens

Next, other ways in which the above metrics could be combined were considered to further narrow down lists of virulence candidates for experimental characterization. One biologically relevant combination is domains of eukaryotic origin that occur in bacteria and also appear pathogen-associated, since these represent potential "mimicry" proteins that facilitate modulation or disruption of host processes by microbial pathogens.[57,241,296] To identify candidate virulence factors with eukaryotic-like domains, the list of bacterial pathogen-associated domains was intersected with the list of domains that are most common in eukaryotes. A total of 49 domain families were identified by this analysis (Table 4.7). Among the identified proteins are known examples of molecular mimicry by bacterial pathogens including the RalF virulence factor of *Legionella* which mimics host Sec7 guanine exchange

factors (GEFs) (PF01369), Table 4.7). Additional *Legionella* secreted effectors, such as a protein family containing a eukaryotic RAS-GEF domain (PF00617), are also included in this list. Other interesting predictions including the Latrotoxin_C domain (PF15658) that is found in Spiders but is also present in *Wolbachia* species, which are insecticidal toxins. Each of these cases implies an ancestral horizontal gene transfer event from a eukaryotic species to bacteria.

Table 4.7: List of eukaryotic-like, pathogen-associated domains identified in bacterial genomes.

7TM_GPCR_Sri, BRICHOS, Choline_kinase, Cystatin, Cytadhesin_P30, DIT1_PvcA, DNA_pol_B, DNA_pol_B_exo1, DUF1479, DUF1726, DUF1729, DUF3827, DUF762, Dynein_heavy, Ecl1, Ehrlichia_rpt, Elongin_A, EMP24_GP25L, Erp_C, F-box, F-box-like, GDA1_CD39, GNAT_acetyltr_2, Helicase_RecD, His_Phos_2, HMG_CoA_synt_C, HMG_CoA_synt_N, IES5, Latrotoxin_C, LMP, Methyltransf_10, MRG, MyTH4, Octapeptide, P_C10, P16-Arc, PAM2, PBC, PC4, Peptidase_M16_M, PhoLip_ATPase_C, Proteasom_PSMB, PTPlike_phytase, Rad33, RasGEF, SAT, Sec7, YMF19, zf-Nse

Combining pathogen association with environment association

Identifying eukaryote-associated domains enriched in pathogens can identify general virulence factors with a broad array of possible eukaryotic hosts. To further focus predictions towards pathogen-associated protein families that are relevant to human disease (such as human enteropathogens), the list of pathogen-associated domain families was intersected with the list of families that were significantly more diverse in the human gut microbiome than the other environments. The top 20 of these are listed in Table 4.8. There is quite a striking enrichment of known virulence factors in these predictions with numerous DUF families interspersed within this list as well (Table 4.8). Families identified by this analysis include the LcrG family (PF07216), which encode a component of the *Yersinia* yop operon for secretion of virulence factors, BNR_3 (PF13859; bacterial neuraminidase), HrpB7 (PF09486; type III secretion effector), Glyco_transf_52 (PF07922) which produces lipooligosaccharide (a pathogenicity determinant), the toxin family Thiol_cytolysin (PF01289), and the virulence factor Pertactin (PF03212). DUFs within this list include DUF2492 (PF10678), DUF1430 (PF07242), and DUF3173 (DUF3173). Based on InterPro descriptions for entries IPR019620 and IPR006541, DUF2492 appears to be a metal binding sulfatase and may play a role in sulfated mucin metabolism. DUF1430 appears to be a transporter and occurs in numerous pathogens including *C. difficile*, *Enterococcus*, and *S. pneumoniae*. DUF3173 (PF11372) is largely restricted to Firmicutes including numerous pathogens, and appears to be conserved near phage integrase genes. DUF families identified by this analysis are of

particular relevance and should be prioritized for functional characterization in the context of human gut pathogenesis.

Table 4.8: Top 20 pathogen-associated Pfam families that are also enriched in the human gut microbiome. Families are ranked by fold change in pathogen proteomes. N = number of proteomes with domain, N_p = number of bacterial pathogen proteomes with domain, N_{np} = number of non-pathogen proteomes (bacterial) with domain.

	N	**N_p**	**N_{np}**	**P_{adj}**	**Fold change in pathogens**	**Human gut-association (P_{adj})**
LcrG	16	7	9	1.60×10^{-7}	17.35	1.37×10^{-21}
DUF4948	13	3	10	4.30×10^{-2}	6.69	7.23×10^{-27}
Mac-1	47	10	36	2.67×10^{-5}	6.20	4.18×10^{-16}
Gp58	58	6	24	3.62×10^{-3}	5.58	2.79×10^{-18}
BNR_3	47	7	30	2.05×10^{-3}	5.21	1.75×10^{-20}
zinc-ribbons_6	175	32	142	5.02×10^{-13}	5.03	4.02×10^{-21}
HrpB7	44	8	36	1.14×10^{-3}	4.96	1.68×10^{-23}
Glyco_transf_52	120	21	99	2.94×10^{-6}	4.73	1.23×10^{-21}
DUF2492	202	35	166	2.74×10^{-13}	4.70	2.27×10^{-26}
HDC	43	7	34	4.33×10^{-3}	4.59	7.11×10^{-27}
Thiol_cytolysin	187	30	153	1.20×10^{-10}	4.37	5.38×10^{-21}
Glyco_hydro_98C	32	5	27	3.25×10^{-2}	4.13	9.11×10^{-19}
HU-DNA_bdg	58	9	49	1.97×10^{-3}	4.10	2.27×10^{-23}
PagP	245	37	203	5.23×10^{-12}	4.07	1.22×10^{-16}
CBM32	40	6	34	2.01×10^{-2}	3.94	2.13×10^{-33}
Pertactin	316	45	266	2.79×10^{-13}	3.77	1.23×10^{-17}
DUF1430	98	14	84	1.70×10^{-4}	3.72	3.62×10^{-21}
DUF3173	112	16	96	4.76×10^{-5}	3.72	1.76×10^{-18}
Glyco_hydro_98M	42	6	36	2.59×10^{-2}	3.72	1.46×10^{-18}
MuF_C	121	15	91	1.02×10^{-4}	3.68	1.86×10^{-19}

Feasibility for structure determination

An additional perspective that must be considered in future efforts to characterize DUFs and other protein families is feasibility and novelty with respect to structural characterization. For structural feasibility, difficulties are associated with proteins that have multiple domains, transmembrane regions, and disordered regions.[213] All Pfam proteins were assessed based on these properties (collected from the Pfam database) to identify a subset that is likely more amenable to structure determination.

Structural Novelty: A feature that is important for structural prioritization is whether

a domain family already includes a structural representative within the PDB. 48.52% of Pfam entries (8700 of 17929) have no structural representative in the PDB, while 81.03% of DUFs (3281 of 4049) have no structural representative.

Domain architecture: 15.14% of Pfam entries have only single-domain architectures (2691 of the 17772 domains that have domain architectures in Pfam) while 28.76% (1158 of the 4026 DUFs with domain architectures in Pfam) of DUFs have only single-domain architectures (Figure 4.8). The almost two-fold increase in the frequency of single domain architectures for DUFs may be in part due to DUFs being shorter and more difficult to resolve in terms of domain boundaries.

Disorder: Predicted regions of disorder (IUPred) were collected from the Pfam database and analyzed for overlap with domain regions. The average percentage of predicted disordered residues across all domain family members along with the standard deviation are provided on `virfams.uwaterloo.ca`. Most Pfam families (71.76%), including DUFs, have very little ($< 10\%$) to no disordered residues (Figure 4.8).

Transmembrane regions: Predicted transmembrane regions (Phobius) in Uniprot sequences were collected from the Pfam database. The presence/absence of predicted transmembrane domains anywhere in a protein sequence or a protein domain was analyzed. Both of these metrics were expressed as the percentage of the number of domain family members possessing these features. As expected, predictions of transmembrane regions were more consistent for single domains than for whole proteins, as multidomain proteins may have transmembrane domains (Figure 4.8). 13.03% of all Pfam families and 17.83% of DUFs have a majority of family members with predicted transmembrane region-domain overlap.

By combining the structural representative, disordered residue and transmembrane domain metrics, a set of 1398 DUF families were identified that are predicted to be highly feasible for structure determination.

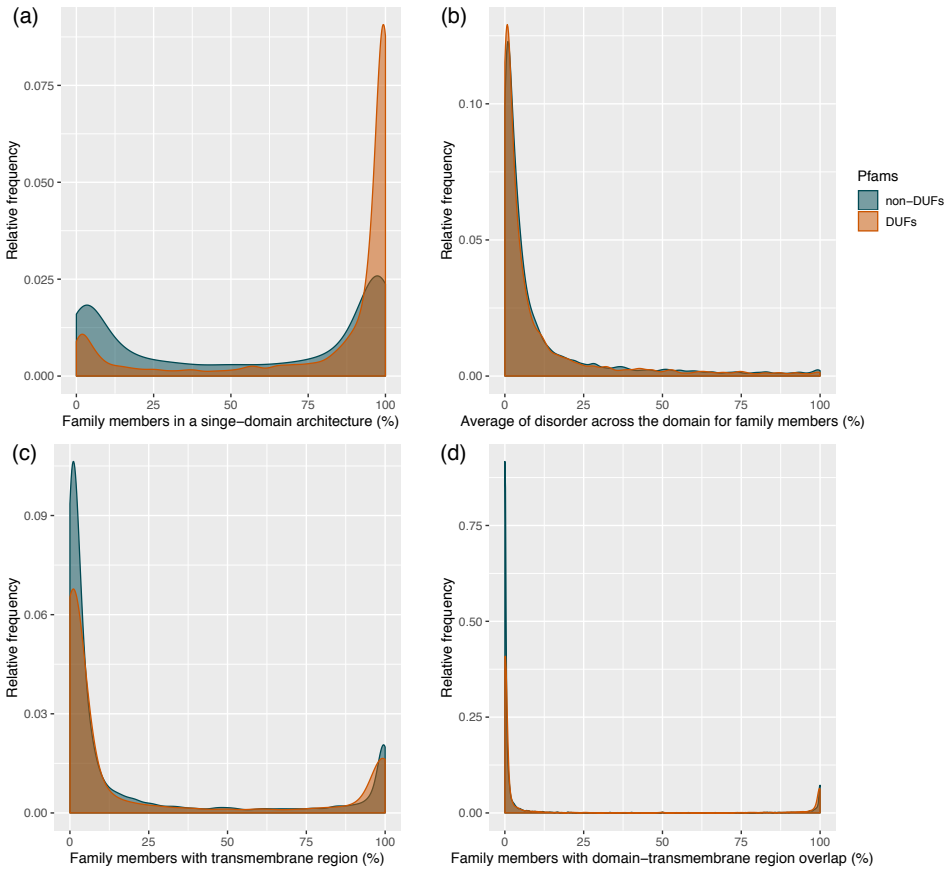

Figure 4.8: The distributions of additional filters for determining structural characterization feasibility in DUFs and other Pfam families. (a) Single-domain architecture frequency in families. (b) Number of residues across the domain with predicted disorder (IUPred) averaged across the family. (c) Frequency of one or more transmembrane regions predicted (Phobius) anywhere in the protein members of domain families. (d) Frequency of having any transmembrane region prediction (Phobius) that overlaps with the domain in Pfam families.

VirFams: an online database for statistical exploration of protein domain families

Finally, in order to provide these analyses to the community, an online database (virfams.uwaterloo.ca) was constructed[2] which facilitates interactive exploration of all Pfam domain families including DUFs. As an example demonstrating the use of our database, Figure 4.9a illustrates the VirFams page for Pfam family LcrG (PF07216) described earlier. A summary panel provides an overview of LcrG's scores according to its overall abundance, lineage-specificity, environmental association, and pathogen-association. This family is significantly enriched in the human gut metagenome, is significantly pathogen-associated, is non-lineage-specific and thus distributed across taxa, and is relatively low in abundance. VirFams also reports the top phylogenetically co-occurring Pfam domain families based on the PhyloCorrelate algorithm[3]. These include a variety of type III secretion system domains (the highest correlated domain is LcrV), which is consistent with the known role of LcrG as a type III secretion system component.[47] Also of note is Pfam domain DUF4765 (PF15962) which is detected in a putative cytotoxic necrotizing factor in *Moritella viscosa* but also shows up in a single-domain architecture in known pathogens like *Escherichia coli* O157:H7. In *E. coli*, this domain is in some predicted T3SS secreted effectors as well as other unannotated proteins. Like LcrG, it is overrepresented in human gut metagenomes, enriched in pathogenic organisms, distributed across three phyla, and is present in a low number of species. This DUF represents an intriguing target for experimental characterization, derived from the assembled association data and visualized using VirFams.

[2]Benjamin Tremblay designed and built the website from the domain metadata associations and the various rankings described in this chapter, with feature development from Dr. Andrew Doxey and I.

[3]Unpublished work by Benjamin Tremblay, Briallen Lobb, and Dr. Andrew Doxey.

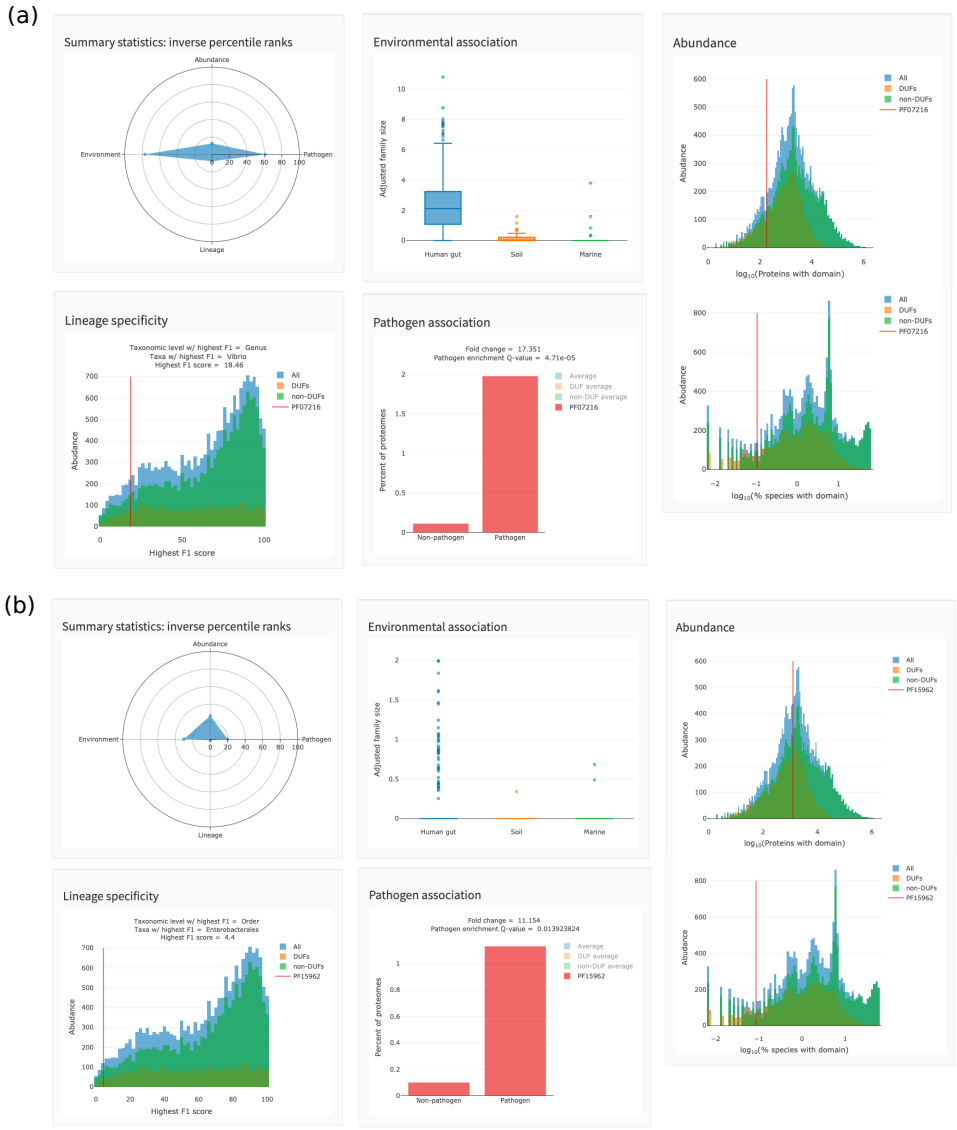

Figure 4.9: A compiled set of screenshots from the VirFams resource for protein domains (a) LcrG and (b) DUF4765.

4.4 Conclusion

In this work, all 17,929 protein domain families in the Pfam v32.0 database were analyzed in order to rank them based on several biological criteria. 1675 out of 4049 (41%) of all DUFs had significant lineage, pathogen, and/or environment associations. These non-homology based associations provide a biological context from which uncharacterized domain families (DUFs) can be prioritized for future studies. In addition, by combining different scores, it was possible to identify Pfam families with certain phenotypic or functional associations, such as candidate virulence factors in the human gut microbiome, as well as candidates predicted to be feasible for structure determination. While these associations are not predictions of specific molecular function, they form a framework to support other methods of functional inference. Here, alternative methods were successful in providing contextual clues for conserved protein families, assisting the prioritization and discovery of novel proteins of interest.

Chapter 5

Metagenomic ORFan annotation

Material in this chapter has been published as part of Lobb et al. (2015).[179] The published manuscript is available here:

> B. Lobb, D. A. Kurtz, G. Moreno-Hagelsieb, and A. C. Doxey. Remote homology and the functions of metagenomic dark matter. *Frontiers in Genetics*, 6:234, 2015.[179] https://doi.org/10.3389/fgene.2015.00234

Metagenomes have substantial problems with annotation coverage (Figures 1.1, 2.2, and 2.3), due to issues like short-fragmented reads and the presence of organisms with large taxonomic distances from well-characterized species. In genomes, the presence of conserved proteins of unknown function and confirmation of their expression in transcriptomics and proteomics can validate their existence as coding sequences within an organism. These proteins are key contributors to protein and domain families, such as the domain families discussed in Chapter 4. But some predicted coding regions do not have any similar sequences or other experimental data to lend weight to their existence outside of DNA. ORFans, predicted ORFs that lack detectable homology to any proteins currently in our databases, are highly prevalent in metagenomes. However, the extent to which ORFans encode real proteins, the degree to which they can be annotated, and their functional contributions, remain unclear. To gain insights into these questions, sensitive remote-homology detection methods were applied to functionally analyze ORFans from soil, marine, and human gut metagenome collections. I found that a considerable number of metagenomic ORFans exhibit significant remote homology to structurally characterized proteins, providing a means for ORFan functional profiling. The extent of detected remote homology far exceeds that

obtained for artificial protein families. As expected for real genes, the predicted functions of ORFans are significantly similar to the functions of their gene neighbors. Compared to the functional profiles predicted through standard homology searches, ORFans show biologically intriguing differences. Many ORFan-enriched functions are virus-related and tend to reflect biological processes associated with extreme sequence diversity. Each environment also possesses a large number of unique ORFan families and functions, including some known to play important community roles such as gut microbial polysaccharide digestion. Lastly, ORFans are a valuable resource for finding novel enzymes of interest, as I demonstrate through the identification of hundreds of novel ORFan metalloproteases that all possess a signature catalytic motif despite a lack of similarity to known proteins.

5.1 Introduction

Metagenomes are a rich resource of novel genes[90] from which the metabolic and physiological activities of entire microbial communities can potentially be inferred.[106] This difficult task relies largely on the accuracy of current methods for predicting function from sequence, which is challenging even for single microbial genomes.[338]

Standard homology-based annotation methods have become the most common strategy for metagenome annotation.[245] Here, metagenome-derived ORFs are searched using BLAST,[7] or related tools, against reference protein databases such as the NCBI non-redundant (nr) and Swissprot databases. Alternatively, reads can be scanned against databases of protein domain models such as the CDD[193] and Pfam,[73] where each protein family is represented by either PSSMs or HMMs. If functionally annotated hits in the databases are detected, functions are inherited from these hits.

Both frustrating and intriguing are the many predicted genes within metagenomes (and genomes) that cannot be readily annotated using standard homology-based methods. The most challenging among these genes are the ORFans, genes that lack detectable homologs in the database.[284] Initially identified in some of the first genomes,[61] ORFans have become a universal feature of newly sequenced genomes and metagenomes, despite an exponential increase in sequencing.[307] Estimates of ORFan content in metagenomes vary from 25 to 85% of total genes.[245] This proportion depends on numerous factors including read length, metagenome complexity, species novelty, homology detection methods and significance thresholds. In addition, a large fraction of metagenome-derived sequences come from microorganisms that resist current cultivation techniques,[88] which makes them dissimilar from database sequences and hard to annotate. Prakash and Taylor[245] showed that, of the genes in the human gut microbiome, 75% could be annotated, vs. only 50–55% of genes in

"complex metagenomes" from soil and ocean environments. Another recent study of a large prairie soil metagenome reported that only 30–38% of predicted proteins had detectable similarity (≥60% identity) to proteins in NCBI's M5nr database,[113] and this has dropped as low as 15% in some extreme cases (e.g., the cow rumen virome).

Several types of alternative, non-homology-based methods may be applicable to annotation of ORFan proteins. Genomic context methods, for instance, predict functions for uncharacterized ORFs based on functions of neighboring genes since gene neighborhoods in prokaryotes tend to possess a significant degree of functional consistency.[42,81,145,195,271,344] These "guilt by association" methods have previously been applied to metagenome annotation[108,328] but depend on assembled contigs, which can be difficult to obtain. Another popular class of prediction methods includes remote-homology detection approaches such as HMM profile-profile comparison. These methods are based on the principle that distant homologies may be apparent by comparison of conservation profiles between families, even if they are not apparent between single sequences.[270,274] The popular profile HMM-HMM comparison method, HHpred/HHsearch,[292] is among the most sensitive methods for homology detection and is consistently ranked among the top automatic structure prediction methods in recent CASP (Critical Assessment of protein Structure Prediction) competitions.

To my knowledge, no studies have applied remote homology to large-scale annotation of metagenomic ORFans, perhaps due to the considerable computation required. Thus, the functions and origins of ORFans, which can be abundant in environmental sequences, are unclear. Here, ORFans were identified and analyzed from three large metagenome collections: the Great Prairie Soil Metagenome Grand Challenge (hereby referred to as just soil), the Global Ocean Sampling (marine), and the Human Gut Microbiome (human gut), encompassing aquatic, host-associated, and terrestrial environments. Through an analysis of 35,307,707 total CDSs, thousands of novel ORFan protein families were identified, with ∼15% gaining an inferred functional annotation through remote homology to proteins of known structure. The structural predictions provide insights into the functions and evolutionary origins of ORFan proteins.

5.2 Methods

Datasets and identification of metagenomic ORFans

This ORFan identification pipeline was developed with Daniel Kurtz and he ran the initial steps of the pipeline on the soil and marine datasets.

Metagenomic sequence data was retrieved from three large metagenome collections: soil[113] [MGRAST IDs 4504797.3 and 4504798.3], marine[269] [http://camera.crbs.ucsd.edu/projects/details.php?id=CAM_PROJ_GOS], and human gut[248] [http://www.bork.embl.de/~arumugam/Qin_et_al_2010/].

For coding sequence prediction, FragGeneScan v1.18[262] was applied directly to the unassembled reads from the marine dataset. Due to the short read lengths from the soil and human gut datasets, FragGeneScan was applied to pre-assembled metagenomes from Howe et al.[113] and Qin et al.,[248] respectively. Segmasker from the BLAST v2.2.28+ package was used to identify repetitive regions in putative ORFs, and CDSs containing over 40% repetitive sequence were discarded. To annotate CDSs with domain family homologs, hmmsearch from HMMER v3.1b1 was used to scan the Pfam database (Pfam-A downloaded 15 May 2014), and remaining CDSs were scanned against the CDD (20 Feb. 2014 release from NCBI) using rpsblast from the BLAST v2.2.28+ package. An *E*-value cut-off of 10^{-3} was used for both methods. CDSs without identified domain family homologs, were clustered with CD-HIT v4.6.1 using a 60% identity threshold. Spurious coding sequence predictions were identified as singleton clusters (those containing one sequence), clusters whose representative (longest) sequence was shorter than 100 amino acids, and clusters comprised entirely of sequences with 99% or greater identity to the representative sequence. These spurious clusters were excluded from further analysis. Representative sequences of each remaining cluster were used for BLASTP database searches (downloaded 15 May 2014 from NCBI). Clusters with either no similarity to the nr database or with a top nr BLAST match exceeding the cutoff of $E = 10^{-3}$ (used previously by Kuchibhatla et al., 2013[149]) were defined as "ORFans". MSAs of the non-spurious clusters were generated with MUSCLE v3.8.31 (www.drive5.com/muscle), and these were further enlarged with sequences from the nr20 database (12 Aug. 2011 release from HH-suite[292]) using HHblits from the HH-suite v2.0.16 package with default settings.

Remote homology detection and FDR estimation

Profile-profile comparisons were performed using HHsearch from the HH-suite v2.0.16 package[292] with the PDB70 HMM database (17 May 2014 release from HH-suite) and default settings. For each prediction, an *E*-value and probability score were collected. To determine appropriate thresholds, remote homolog detection was repeated using random, reshuffled alignments as described below. Based on the results, a probability threshold of 80% was chosen with the *E*-value set at 1, equivalent to a ∼9% false discovery rate (see 5.3, Results). To obtain an FDR estimate, the pipeline was repeated using shuffled alignments which represent artificial sequence families that maintain compositional characteristics and

column-specific conservation.[102,196] One thousand ORFan clusters obtained by CD-HIT were randomly selected from each metagenome, and the columns of each cluster's MSA were shuffled. The shuffled alignments were run through the HHblits and HHsearch algorithms as described previously using the non-shuffled clusters.

Genomic context analysis

The coding sequence locations on contigs (for the soil and human gut datasets) and reads (for the marine dataset) were used to define genomic neighbors and perform genomic context analysis. The Pfam-GO mapping from InterPro[115] was used to assign GO terms to ORFs. For Pfam domain homologs, the GO terms of all significant ($E < 10^{-3}$) domain matches were included in its functional annotation. For the non-spurious CD-HIT clusters (ORFans and clusters with homologs from the NCBI nr database), a GO term collection was assigned to each cluster based on the top three significant remote homologs found by HHsearch, using the PDB-GO annotation table obtained from the EBI (`http://geneontology.org/gene-associations/gene_association.goa_pdb.gz`). GO terms were assigned to each coding sequence within the CD-HIT cluster.

For each metagenome, the list of GO terms for an ORFan were compared against the list of GO terms associated with its directly neighboring CDSs (one on either side, in the same orientation and within 1 kb) on the same contig, and calculated the number of shared terms (S) between both sets. This value was then summed for all ORFans within a metagenome (m) to obtain an overall statistic (S_m) reflecting the similarity between ORFans and their annotatable genomic neighbors. To estimate statistical significance, S_m was compared to a null distribution computed by swapping the ORFans amongst their original locations. The count was then calculated as above, shuffling ORFans only while maintaining the positions of all other CDSs. Shuffling followed by the shared GO terms summation was performed 1000 times.

Analysis of overrepresented functions

To determine the frequency of GO terms in each metagenome, 10,000 CDSs with Pfam domain hits were randomly selected from each metagenome and run through HHblits with only one iteration and a limit of 30 sequences in the output alignment followed by HHsearch with default settings (using the databases described previously). The functional information for ORFan sequence clusters and the subset of Pfam domain hits was gathered using the most confident GO term-associated HHsearch hit (using the PDB-GO map and

only assessing significant HHsearch hits). Similar to previous studies,[322,326] analyses were restricted to sixth level GO terms in the biological process or molecular function trees since this level was more informative (greater biological specificity) than other trimmed ontologies such as GO Slim terms. GO term levels were calculated using the "is a" relationship, with the starting terms (biological process and molecular function) being considered level one. Only the longest path from the root terms was considered. The frequency of each GO term in the Pfam and ORFan subsets and PDB70 were calculated, with zero counts converted to a pseudocount of 1 to avoid division errors. The fold change of each GO term in the ORFan sequence clusters over the Pfam domain hits subset was calculated and compared across metagenomes. P values were calculated in R using the binomial test with false discovery rate adjustment (p.adjust function) as described elsewhere.[53]

Analysis of environment-specific ORFan families

For each metagenome, the proportions of the total number of ORFans matching a PDB entry as the top remote homolog was computed. Three-dimensional scatterplots were generated with each axes representing this quantity. The binomial test was used to compute P values with background probabilities based on the total counts observed in the other two metagenomes. These P values were then corrected using the Bonferroni adjustment. The same procedure was repeated based on proportions of ORFans from each metagenome possessing GO terms (1769 total terms).

ORFan metalloprotease discovery

ORFan clusters were searched for those that: (i) possessed a top remote homolog match to a PDB entry possessing "protease" or "peptidase" terms in any functional description category; (ii) had a representative sequence with at least one match to a HExxH motif. ORFan CD-HIT clusters meeting both conditions were considered putative ORFan metalloproteases or metallopeptidases.

5.3 Results

Identification of ORFan sequences in three large metagenomes

With the goal of characterizing ORFans from diverse metagenomes, three large, publicly available datasets were retrieved and analyzed: the Great Prairie Soil Metagenome Grand

Challenge[113] (soil dataset), Global Ocean Sampling[269] (marine dataset), and Human Gut Microbiome[248] (human gut dataset). Metagenomes were selected from diverse biomes (terrestrial, aquatic, and host-associated) since observed differences in ORFan content and functions may be biologically relevant while commonalities may indicate general trends.

First, all genes within these metagenomes were predicted regardless of whether they could be verified through homology to known sequences. This initial set included a staggering number (35,307,707) of CDSs, equivalent to about 20% of the entries in the current NCBI GenBank database. Each coding sequence was processed using the computational pipeline described in Figure 5.1 (see Table 5.1 for statistics at each step), with the intention of separating the ORFans from the homology-annotatable sequences. Potential ORFans were identified as CDSs whose products lacked detectable homology to known protein domain families (Pfam and CDD) or proteins in the NCBI database (see section 5.2, Methods). Since these potential ORFans likely contain a mixture of real ORFan proteins and false positives,[87] additional steps were required to remove spurious ORFs. Therefore, CDSs were clustered and any singletons,[87,283] clusters with low sequence variation, and clusters composed exclusively of short fragments (see Methods 5.2) were removed. This left 85,422 (soil), 251,857 (marine), and 146,842 (human gut) putative ORFan proteins from each metagenome Table 5.1. By definition each ORFan within this final set is an apparent gene coding for a protein, is a member of a sequence cluster with at least one representative of 100 amino acids or longer, and yet has no detectable homology to any known protein or conserved domain family. All following analyses were performed on this set of ORFans.

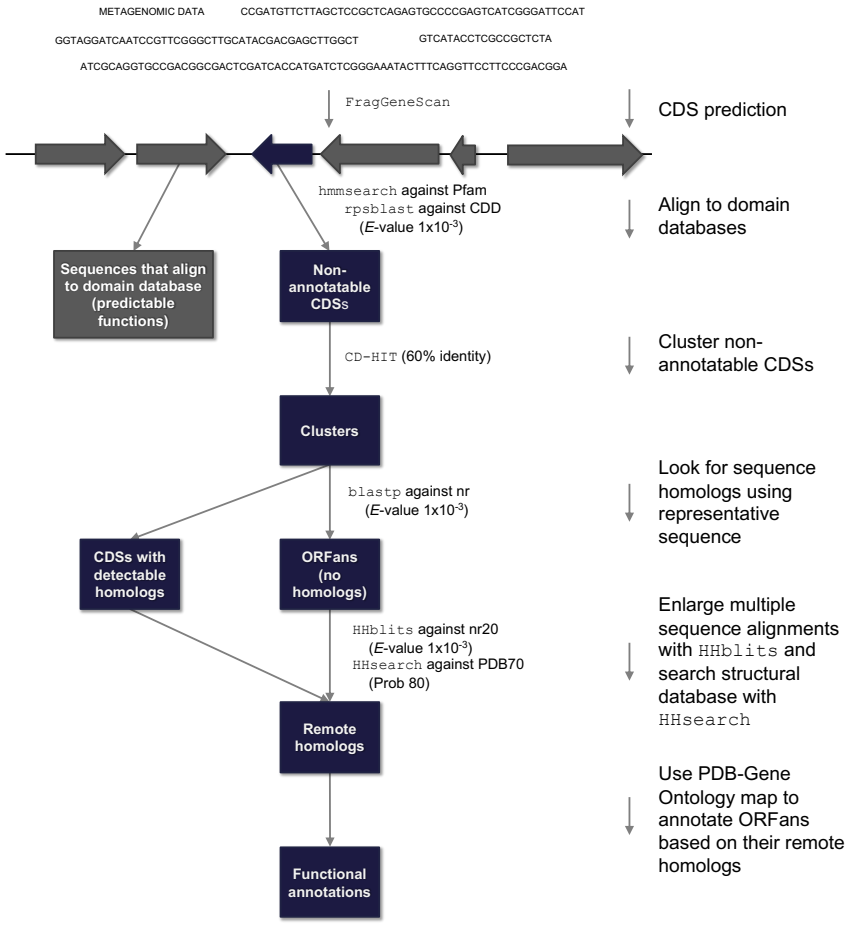

Figure 5.1: Pipeline for detection and functional annotation of metagenomic ORFan proteins. Protein CDSs were predicted from assembled metagenomic contigs, and searched against conserved domain databases. CDSs that could not be annotated by domain homology were further clustered, and representatives were BLASTed against the NCBI nr database. Remaining coding sequence clusters lacking detected homologs were considered ORFans, and these were subjected to remote homology detection using HHblits and HHsearch, which were used to perform profile-profile searches against the Protein Data Bank.

Table 5.1: Number of CDSs and ORFans at key stages of metagenomic ORFan identification.

	Soil	**Marine**	**Human Gut**
Predicted CDSs	5,606,711	17,204,095	12,496,901
CDSs removed containing conserved domain matches (Pfam and CDD)	2,480,274	4,542,071	4,674,912
Spurious (singleton, short and repetitive) CDSs removed	2,758,146	11,458,304	6,603,567
CDSs removed with BLAST matches to nr database	282,869	951,863	1,071,580
Candidate functional ORFans	85,422	251,857	146,842
ORFan CD-HIT clusters	33,013	73,428	32,078
Annotated (HHsuite) ORFan CDSs	21,358	38,900	13,638
Annotated (HHsuite) ORFan CD-HIT clusters	7848	10,973	3119

ORFans are shorter but compositionally similar to real proteins from their environments

Next we examined whether the detected ORFans share compositional characteristics with homology-annotatable CDSs (those with Pfam or CDD domain matches) from their environments. If so, this would suggest that predicted ORFans are under similar evolutionary pressures as real proteins and indicate potential functionality. We therefore investigated the distributions of coding sequence length and GC content (Table 5.2) for each coding sequence category. Biases have been observed previously for ORFans.[37,348,350] Consistent with previous studies, ORFans tend to be shorter in all datasets (Table 2), and the relative abundance of ORFans also decreases with increasing read length (Figure 5.2). Overall, the GC content distributions of the homology-annotatable CDSs and ORFans are highly similar within but vary considerably between metagenomes (Figure 5.3). Although the length distributions are also affected by sequencing method, this is not the case for GC content, suggesting that the predicted ORFans exhibit characteristics of the *real* (homology-annotatable) CDSs from their environments.

Table 5.2: Average GC content and length of domain-annotated vs. ORFan sequences from three metagenomes.

	Average GC content (%)	**Average CDS length (# of nucleotides, nt) excluding any sequences under 300 nt**
Soil Pfam and CDD hits	56.8	411.4
Soil ORFans	54.8	407.4
Marine Pfam and CDD hits	39.2	731.7
Marine ORFans	39.4	548.7
Human gut Pfam and CDD hits	46.6	781.7
Human gut ORFans	43.0	525.2

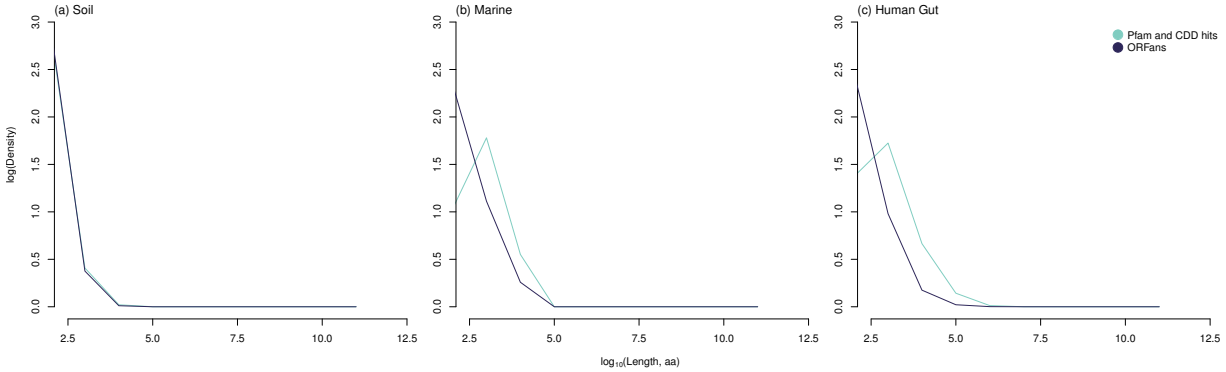

Figure 5.2: ORF length distributions for homology-annotatable versus ORFan sequences from three metagenomes. The relative abundance of ORFans decreases with increasing read length, which reflects the tendency for ORFans to be shorter than average proteins.

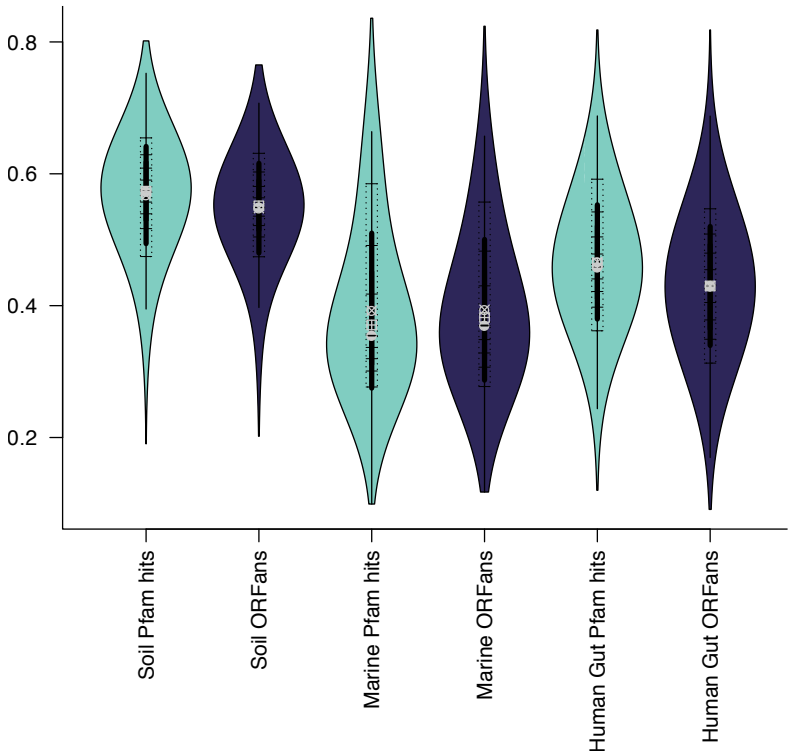

Figure 5.3: GC content distributions for homology-annotatable versus ORFan sequences from three metagenomes. Homology-annotatable and ORFan sequences display highly similar GC content distributions within the same environment, but these distributions differ significantly between environments.

Many ORFans exhibit remote homology to proteins of known structure

Although ORFans, by definition, do not possess detectable homology to existing protein families using standard database search techniques like BLAST or HMMER, we were interested whether remote homology detection techniques could prove effective. We applied profile-profile, remote homology detection using HHblits/HHsearch,[261,292] which compares the conservation profile derived from the MSA of the ORFans to those of known protein families. These methods can often identify remote relationships between protein families, even if individual members do not share detectable homology. To facilitate remote homology detection, we first generated initial MSAs for each ORFan cluster, and detected remote

homologs in the Protein Data Bank using HHblits/HHsearch. Since each ORFan cluster contained multiple non-redundant sequences, a non-trivial MSA and profile could be generated in each case. Thus, not only was the sequence clustering step useful in removing spurious ORFs, but it was also essential for generating the conservation profiles used in profile-profile comparison.

A considerable number of ORFans (73,896 sequences, 15.3%; 21,940 clusters, 15.8%) exhibited significant remote homology to proteins of known structure, with some metagenomes producing a greater fraction of annotated ORFans than others: 25.0% (soil), 15.4% (marine) and 9.3% (human gut) of ORFan clusters (Table 5.1; ORFan sequences and annotations available at http://doxey.uwaterloo.ca/ORFans/). This represents a new dataset of annotated, extremely divergent metagenome-derived proteins and provides a means to profile ORFan functions in general.

Despite thorough benchmarking of HHblits/HHsearch,[261] there remains a possibility that the predictions are false positives due to factors associated with the pipeline and dataset. Therefore, we empirically measured a false discovery rate by repeating the entire procedure on an artificial dataset composed of ORFan clusters with shuffled sequences (Figure 5.4a). Specifically, 3000 random ORFan clusters were selected (1000 from each metagenome), and their alignment columns were shuffled, thereby preserving conservation information and compositional characteristics, while destroying potential similarity to real proteins. Any detectable homology between these artificial protein families and the PDB database indicates a false positive prediction. The random dataset generally produced low HHsearch probability scores, whereas the real metagenomic ORFans resulted in a large abundance of high-scoring predictions (Figure 5.4a). At a probability score of 80% or higher, the HHsearch method was able to annotate 15.8% of the real ORFan clusters and only 1.4% of false sequence clusters, which is indicative of a low (∼9%) false discovery rate. This result provides support for the quality of the remote homology predictions, and suggests that many ORFans (15.3%) are divergent homologs of existing structural families.

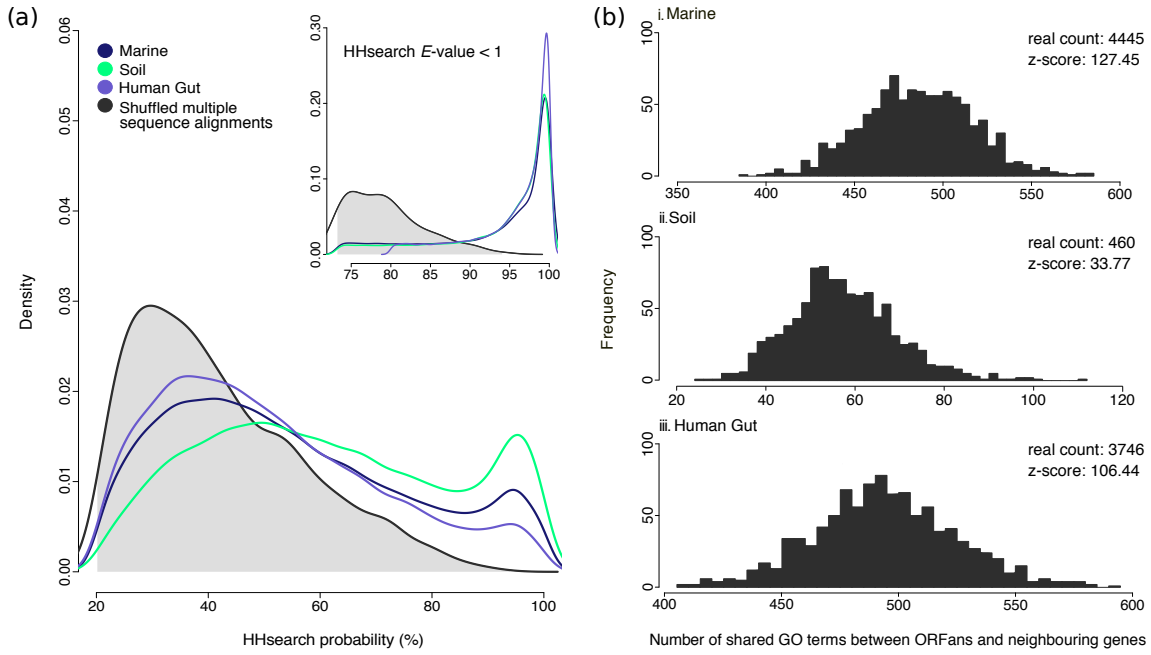

Figure 5.4: Estimated false discovery rate of ORFan remote homology detection and functional prediction. (a) Distributions of HHsearch probability scores for ORFans from three metagenomes, and shuffled sequences, searched against a PDB-derived HMM library. There is an abundance of high-scoring predictions (i.e., above 80% probability) for ORFan proteins compared to the expected (null) distribution. This separation becomes even greater when an HHsearch *E*-value threshold of 1 is applied (see inset). (b) The number of shared GO terms between functionally annotated ORFans (probability scores >80%) and their metagenomic neighbors (see Methods 5.2) is shown for three metagenomes. The null distributions, as estimated by randomly shuffling ORFan identities/positions, are shown along with the z-scores relative to these distributions. The mean values for the random distributions are: marine (486.3), soil (57.8), and human gut (494.4).

ORFan functions are consistent with those of their gene neighborhood

Given that a sizeable portion of metagenomic ORFans exhibit remote homology to protein structures, a key follow-up question concerns what functional information can be gained from

these detected relationships. For functional annotation, the same GO terms were assigned as those associated with their identified remote PDB homologs. To assess whether the predicted ORFan functions are accurate and thus biologically meaningful, their functional consistency with neighboring genes was measured, a well established phenomenon in prokaryotes.[42,81,145,195,271,344] If predicted ORFan functions are accurate, they should show significantly elevated functional consistency compared to a random distribution (see section 5.2, Methods). Functional consistency was calculated as the number of shared GO terms between an ORFan and its metagenomic neighbors, defined as one gene on either side of an ORFan, in the same orientation and within a 1 kb boundary. As a statistical test, the total number of shared GO terms for all annotated ORFans was computed and compared to an estimated random distribution in which the ORFans were shuffled amongst their original locations. ORFans from all three metagenomes exhibited extremely high, statistically significant levels of functional consistency with their neighbors (Figure 5.4b). This effect was abolished completely when the ORFans randomly swap their positions. Overall, the significant functional congruence between ORFans and their gene neighbors suggests that the predicted functions are of high quality and thus potentially meaningful for biological interpretation.

Enriched functions among ORFans

An important next question concerns the predicted ORFan functions themselves, how they compare to the homology-based functional profile inferred for the remaining metagenome, and what insights they may provide into hidden functions of their respective environments. To examine ORFan functions as a whole for each metagenome, ORFan functional profiles were computed as collections of GO terms and their frequencies, as based on previous studies.[317] Separate functional profiles were also calculated for 10,000 Pfam-annotated CDSs of each metagenome as a reference, to which ORFan functions could be compared.

These comparisons reveal that ORFans possess a distinct functional profile from that of homology-annotatable proteins. This is evident from a clustering analysis in which the ORFan functional profiles from the three environments group together (Figure 5.5). However, this is also somewhat expected since ORFans from different metagenomes will be inherently similar by virtue of *lacking* conserved functions present in the homology-annotated subset.

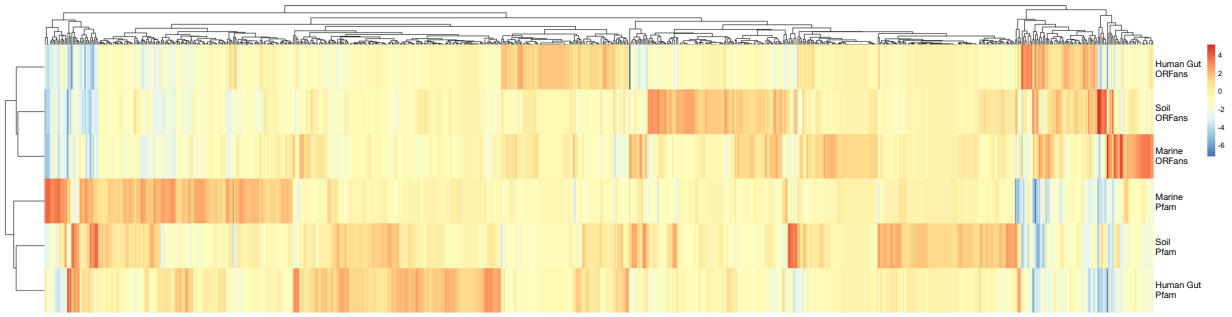

Figure 5.5: Heatmap of the log frequency of GO function terms in the Pfam-annotated subset and the ORFan subset. Only terms enriched (>1.25 fold) in at least one dataset are included in the heatmap to avoid display of invariant functions.

Consistent with the unique functional profile of ORFans, numerous functions were identified that were significantly overrepresented within the ORFans of each metagenome (Table 5.3, Table 3). These ORFan-enriched functions include terms relating to viral processes, carbohydrate metabolism, as well as several functions with particular relevance to their respective metagenomes (explored in following sections). Reported functions were also significantly enriched (all with adjusted $P < 0.05$) compared to the reference database (PDB) and are thus not simply due to random matches to PDB entries.

Table 5.3: Top five significantly enriched GO terms among ORFans in each metagenome relative to non-ORFans and the PDB. Only four significantly-enriched terms were found for the human gut metagenome samples.

GO term	ORFan clusters (individual sequences)	Proportion of ORFan clusters with GO term	Proportion of Pfam-annotated subset with GO term	Fold change	P_{adj} against Pfam-annotated subset	P_{adj} against PDB70
Soil-associated						
GDP-dissociation inhibitor activity	66 (157)	$1.1x10^{-2}$	$6.1x10^{-4}$	18.1	$7.5x10^{-55}$	$1.6x10^{-90}$
Dibenzothiophene catabolic process	35 (110)	$5.9x10^{-3}$	$4.9x10^{-4}$	12.0	$1.7x10^{-22}$	$3.7x10^{-55}$
Mitochondrial fission	28 (79)	$4.7x10^{-3}$	$3.7x10^{-4}$	12.8	$2.2x10^{-18}$	$7.1x10^{-40}$
Sequence-specific DNA binding	162 (415)	$2.7x10^{-2}$	$1.3x10^{-2}$	2.1	$6.4x10^{-14}$	$2.1x10^{-49}$
Viral release from host cell	14 (39)	$2.3x10^{-3}$	$1.2x10^{-4}$	19.2	$1.2x10^{-10}$	$5.1x10^{-2}$
Marine-associated						
Polysaccharide catabolic process	62 (210)	$7.2x10^{-3}$	$4.5x10^{-4}$	16.3	$1.2x10^{-48}$	$8.0x10^{-6}$
L-ascorbic acid binding	89 (306)	$1.0x10^{-2}$	$1.8x10^{-3}$	5.8	$4.7x10^{-35}$	$1.0x10^{-74}$
ADP-heptose-lipopolysaccharide heptosyltransferase activity	35 (136)	$4.1x10^{-3}$	$2.2x10^{-4}$	18.4	$1.6x10^{-28}$	$4.0x10^{-88}$
Phosphatidylinositol alpha-mannosyltransferase activity	26 (104)	$3.0x10^{-3}$	$1.1x10^{-4}$	27.3	$4.9x10^{-25}$	$5.6x10^{-42}$
Endonuclease activity	157 (576)	$1.8x10^{-2}$	$6.7x10^{-3}$	2.7	$1.6x10^{-24}$	$1.2x10^{-11}$
Human gut-associated						
Sequence-specific DNA binding	149 (617)	$6.5x10^{-2}$	$2.9x10^{-2}$	2.2	$3.2x10^{-15}$	$1.6x10^{-94}$
Polysaccharide catabolic process	49 (139)	$2.1x10^{-2}$	$4.6x10^{-3}$	4.6	$1.3x10^{-14}$	$1.2x10^{-21}$
Regulation of sporulation resulting in formation of a cellular spore	11 (74)	$4.8x10^{-3}$	$5.6x10^{-4}$	8.5	$2.3x10^{-4}$	$4.8x10^{-20}$
Ribonuclease activity	18 (88)	$7.8x10^{-3}$	$2.0x10^{-3}$	3.9	$3.7x10^{-3}$	$1.0x10^{-3}$

The detected enrichment of viral functions is consistent with previous suggestions that a large proportion of ORFans may be bacteriophage derived.[44] Since viruses undergo rapid rates of evolution and are relatively undersampled in genomic databases, their proteins may also appear significantly divergent from database sequences. The results provide strong support for this hypothesis since numerous virus-related functional terms are

significantly enriched (adjusted $P < 0.05$) among the annotated ORFans (Table 5.3, Table 3). For example, the term "viral release from host cell" was among top enriched ORFan functions in the marine ($P = 1.1\times10^{-16}$) and soil metagenomes ($P = 1.2\times10^{-10}$). Other enriched functional terms associated with viruses include "RNA ligase",[51] "lysozyme",[71] and "phospholipase"[353] (Table 5.3, Table 3).

Although enriched, we estimate that viral sequences may be a relatively small proportion of ORFans overall, similar to previous reports.[348] That is, only 4.1% (soil), 6.3% (marine) and 5.6% (human gut) of ORFans matched viral protein structures (Table 5.4), while the majority matched structures of bacterial origin. Interestingly, however, the proportions of viral PDB matches are roughly four-fold higher than that observed for the homology-annotatable proteins which ranges from 1.4 to 2.4%, which provides additional support for an enrichment of viral functions among metagenomic ORFans.

Table 5.4: Taxonomic composition of remote PDB matches to ORFans versus homology-annotatable CDSs from three large metagenomes.

	Soil (%)	**Marine (%)**	**Human Gut (%)**
Homology-annotatable CDSs			
Eukaryota	15.7	16.5	9.2
Bacteria	75.5	75.0	80.5
Archaea	7.4	7.2	7.8
Viruses	1.4	1.4	2.4
ORFans			
Eukaryota	17.6	18.1	14.6
Bacteria	72.4	69.7	73.6
Archaea	6.0	6.0	6.3
Viruses	4.1	6.3	5.6

Another common function overrepresented in the ORFans of all three metagenomes relates to carbohydrate degradation or transport. This finding is consistent with the considerable sequence and structural diversity of carbohydrate-active enzymes.[32] Enriched carbohydrate-related functions among ORFans include "polysaccharide catabolic process" in all three metagenomes (all with $P <1\times10^{-5}$), "cellulase activity" ($P = 6.1\times10^{-7}$) in the soil dataset and "phosphatidylinositol alpha-mannosyltransferase activity" in the marine dataset ($P = 4.9\times10^{-25}$) (Table 5.3, Table 3).

Ultimately, both the clustering and enrichment analyses demonstrate that ORFan functions do not merely mirror the functions expected from homology-annotatable proteins. Thus, the efforts of remote homology detection have uncovered a highly divergent sequence

space, including viral proteins and carbohydrate-active enzymes, which was not detectable in the annotatable subset of each metagenome.

Environment-specific ORFan families and functions

Potentially more interesting than the functions generally enriched among ORFans are the specific ORFan families and functions unique to each environment. Indeed, it has been hypothesized that ORFans may be unique in their potential to encode ecologically important functions.[336] One explanation for this is that environment-specific functions may be encoded in part by environment-specific genes that differ from characterized genes in the database.

To explore this in greater detail, metagenome-specific ORFan functions were visualized using 3D scatterplots (Figure 5.6), similar to previous three-way comparisons of metagenome functional profiles.[317] In these plots, ORFan functions that are of similar abundance in all three metagenomes will appear close to the origin, whereas ORFan functions that are relatively abundant in one metagenome will project outwards along that metagenome's axis. In addition to GO terms, the same analysis was also performed at the level of ORFan families, as represented by the top identified remote homolog in the PDB.

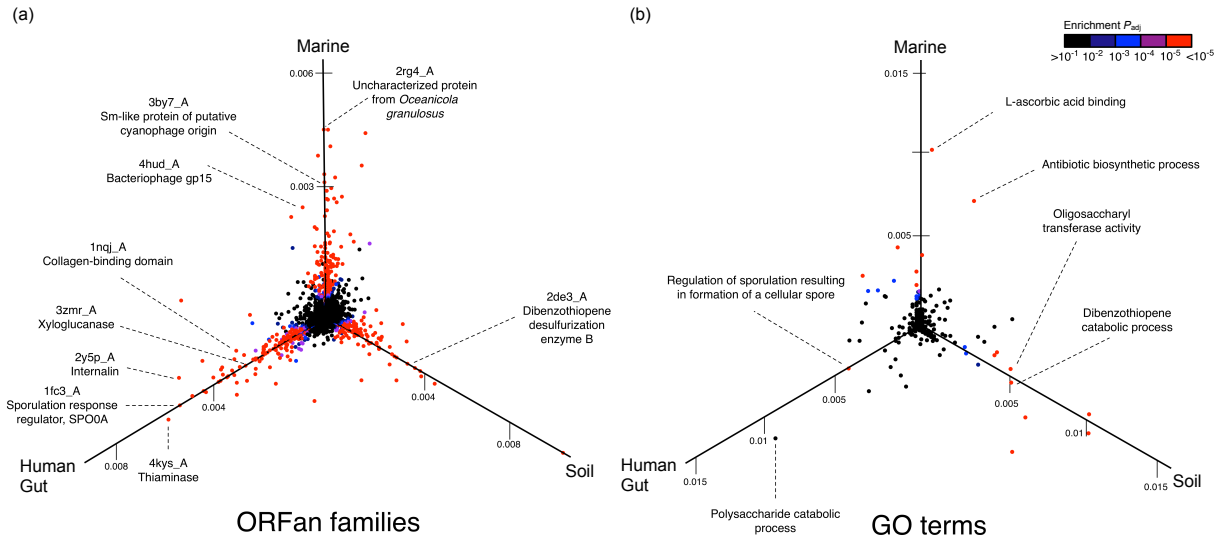

Figure 5.6: Metagenome-specific ORFan families and functions. Shown are projections of three-dimensional scatterplots in which each axis indicates the proportion of ORFans from a specific metagenome with a specific annotation: (a) families and (b) functions. ORFan families are defined based on their top remote homology match in the PDB database, and functions are defined by GO terms as described in section 5.2, Methods. Data points that project uniquely along one axis therefore indicate metagenome-specific ORFan families or functions, while those close to the origin indicate similar proportions among all three metagenomes. Cases described in the text have been labeled.

This three-way comparison reveals several broad functions (Figure 5.6, right) and a much larger number of families (Figure 5.6, left) that are significantly enriched in the ORFans from one metagenome. The following sections highlight some interesting examples from each environment.

Human gut-specific ORFans

Several of the most abundant human gut-specific ORFan families have predicted roles involved in gut metabolism and host interactions. These include human gut-specific ORFan homologs of thiaminase, an enzyme that breaks down vitamin B1, the virulence factor internalin, and the collagen-binding domain which could play roles in gut adherence or invasion (Figure 5.6).

Most intriguing are the ORFans with predicted functions in "polysaccharide catabolic process," a function that is significantly enriched ($P = 1.3\times10^{-14}$, Table 5.3) in the human gut dataset (Figure 5.6). This is of great interest in the context of the human gut microbiome because breakdown of indigestible dietary polysaccharides is one of the fundamental roles of intestinal bacteria.[77] Among the most abundant human gut-specific ORFan families is one with detected remote homology to PDB ID 3zmr, a crystal structure of xyloglucanase from the common human gut organism, Bacteroidetes.[158] This enzyme functions in the gut microbial digestion of the plant-cell wall derived polysaccharide, xyloglucan (XyG), and was only recently characterized as the first xyloglucanase enzyme in the gut microbial community.[158] The human gut-specific ORFans identified here exhibit remote homology to the Bacteriodetes-Associated Carbohydrate-binding Often N-terminal (BACON) domain within these enzymes, suggesting a function in gut carbohydrate metabolism.

Another human gut-specific ORFan family includes 74 ORFan proteins from 11 sequence clusters in the human gut metagenome collection with a predicted function in regulation of sporulation. This was the third most enriched function (by fold) among human gut ORFans ($P = 2.3\times10^{-4}$, Table 5.3) and yet was not enriched in the other two metagenomes as illustrated in Figure 5.6. These ORFans are primarily distant homologs of the DUF199/WHIA transcriptional regulator or the sporulation response regulator, SPO0A. While sporulation is a general function also observed elsewhere, numerous studies have demonstrated its particular enrichment within the human gut microbiome. This has been attributed to the relative abundance of gut Firmicutes species, which include many spore-forming members.[318] However, specific genes and sporulation pathways may be unique to the human gut microbiome. For instance, a recent analysis of Lachnospiraceae genomes revealed that key sporulation-related genes are exclusive to human gut-associated Lachnospiraceae and absent elsewhere.[205] It is therefore interesting that both ORFans and homology-annotatable proteins from the gut microbiome show this functional pattern. This data further implicates sporulation as a particularly important function within the human gut community, and provides motivation for further exploration of divergent gut sporulation proteins.

Marine-specific ORFans

Several abundant marine-specific ORFan families and functions are indicated in 5.6. Enriched functions include antibiotic biosynthesis and L-ascorbic acid (vitamin C) binding. Interestingly, the most abundant marine-specific ORFan families show patterns consistent with a marine environment. These include a family of ORFans with remote homology to a cyanophage (an abundant marine virus that infects oceanic cyanobacteria) protein,

and another family with remote homology to PDB ID 2rg4, an uncharacterized protein from the marine bacterium, *Oceanicola granulosus*. The identification of marine-specific ORFans matching viral structures (see Figure 5.6 for another example, bacteriophage gp15) is consistent with Yooseph et al.[351] who reported a viral origin for a significant number of divergent Global Ocean Sampling sequences.

Soil-specific ORFans

One of the most interesting soil-specific ORFan families has remote homology to dibenzothiophene (DBT) desulfurization enzyme B (PDB ID 2de3_A). This is also a significantly enriched ORFan function compared to non-ORFans from the same metagenome ($P = 1.7\times10^{-22}$, Table 5.3). DBT desulfurization genes have been identified in petroleum-polluted soils where they are implicated in DBT degradation, and are of interest to the oil industry to reduce the levels of sulfur in fuel.[60]

Targeted discovery of ORFan metalloproteases

Regardless of whether a particular function is overrepresented among ORFans and/or metagenome-specific, its detection within ORFans may be valuable for its own sake to expand its knowledge and sequence space. Indeed, metagenomes are a useful resource for the discovery of novel families of biotechnologically and scientifically important enzymes such as glycosyl hydrolases (Li et al., 2009) and proteases.[334]

To explore its potential as a resource for enzyme discovery, the annotated ORFans were mined for novel metalloproteases. Metalloproteases are of particular biological,[59,224] evolutionary[55,190,260] and biotechnological[3] interest. "Metallopeptidase activity" was also a significantly enriched function among ORFans from the marine dataset ($P = 1.6\times10^{-20}$, Table 3). Lastly, metalloproteases were also selected as a target function because these enzymes possess a convenient functional motif that provides additional evidence of predicted activity; namely, a conserved, zinc-binding, catalytic motif (HExxH). Remarkably, 257 ORFan sequence clusters possessed both this motif and significant remote homology to protease or peptidase structures (Table 5.5). One example is highlighted in Figure 5.7, in which a predicted ORFan family from the human gut displays significant remote homology to the zinc-metalloprotease domain of the anthrax toxin. Although the overall sequence similarity is quite weak, there are short regions of motif similarity and numerous residues within the catalytic site are conserved. The 257 ORFan subfamilies represent a rich resource of highly divergent metalloproteases that await future experimental characterization.

Table 5.5: Predicted ORFan clusters with the HExxH motif and remote homology to metalloprotease structures. The top three most abundant clusters by PDB match are listed.

	Number of clusters	**Remote homology match (PDB entry and description)**
Soil	96	Total
	10	3cqb_A Peptidase M48
	8	4jix_A Peptidase M56
	8	4in9_A Peptidase M10, Matrixin
Marine	132	Total
	24	3cqb_A Peptidase M48
	11	4jiu_A DUF45 metallopeptidase
	10	4jix_A Peptidase M56
Human Gut	29	Total
	5	3dte_A DUF955 peptidase-like domain
	3	3b4r_A Peptidase M50
	3	2y6d_A Peptidase M10

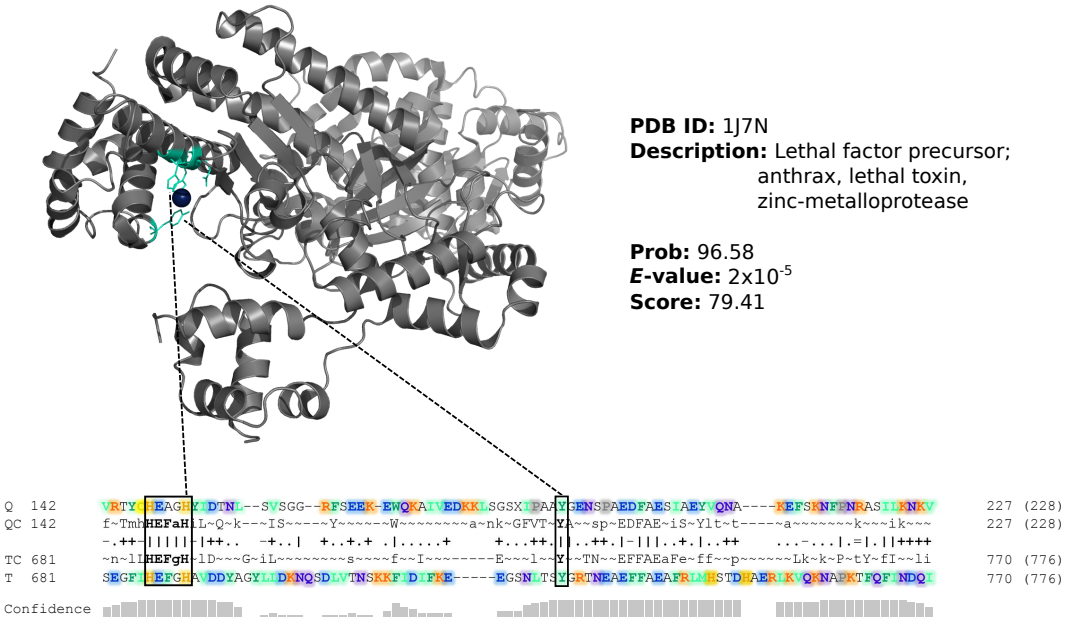

Figure 5.7: One example of 257 predicted metalloprotease ORFan sequence clusters. This example is a predicted metalloprotease ORFan from the human gut dataset with similarity to the protease domain of the anthrax toxin (shown). The catalytic zinc-metalloprotease (HExxH) catalytic motif is conserved between the query and template, however the remaining sequence similarity is weak. In general, ORFan metalloproteases were predicted based on detected remote homology to protein structures of known or putative proteases and peptidases, as well as presence of the HExxH motif.

5.4 Discussion

A pipeline was developed to identify and structurally annotate ORFans from three large and highly distinct metagenomes. A considerable fraction (15.3%) of metagenomic ORFans identified from this pipeline exhibit remote but significant homology to structurally characterized proteins. This is surprising since neither BLAST nor profile-based methods were able to annotate them. These findings are consistent with previous structural studies that have consistently revealed ORFans to be divergent members of existing protein families.[90] For instance, a previous analysis of 248 structures of DUF families selected from Pfam, determined that ∼2/3 are divergent members of known protein families.[123]

These structural studies, together with the 15.3% of annotated ORFans presented here, support a classic duplication-divergence model[231] in which ORFan genes might arise when one of two duplicated genes (paralogs) diverge rapidly to a point where homology becomes undetectable.

While initially attributed to an inadequate knowledge of sequence space, pseudogenes or prokaryotic "junk DNA",[9,212] or incorrectly annotated genes,[275] there is considerable evidence that many detected ORFans are functional.[114] A functional role for many ORFans is also supported by the many high quality functional annotations predicted by hhpred. These annotations are themselves supported by a low estimated false discovery rate based on non-homologous shuffled sequences, as well as the significant level of functional similarity detected between ORFans and their neighboring genes.

The overrepresented functions among ORFans are also consistent with previous but debated[348] claims that ORFans tend to be of viral and other mobilomic origins.[37,51] For instance, one study examined 119 prokaryotic genomes for gene clusters exhibiting atypical sequence composition and found that over 39% of ORFans were contained within these clusters, strongly suggesting that integrative elements are a major evolutionary source of ORFans.[37] Viral and mobilomic origins of ORFans make sense from a biological perspective given the rapid mutation rates observed in viral DNA as well as a technical one given the relative undersampling of viral sequences in the database.

Lastly, these results agree with previous suggestions that ORFans encode environment-specific roles,[130,307] specifically through the many metagenome-specific ORFan families and functions that were identified (Figure 5.6). Indeed, ORFans have been implicated in taxon-specific functions[336] and lineage-specific developmental or morphological adaptations.[23,130,307]

Although annotatable ORFans may represent a relatively minor component of a metagenome, they differ dramatically in their functional profiles from typical, homology-annotatable proteins. Their inclusion within metagenome annotation pipelines may not significantly alter overall estimates of metagenome functional profiles, but they are themselves interesting to pursue and expand our understanding of key protein functions of interest. Ultimately, ORFan characterization through remote homology provides a glimpse into the highly divergent, occasionally viral, and environmentally important functions they contribute to their respective microbial communities.

Chapter 6

Conclusion

> Data mining or knowledge discovery from data, in its most fundamental form, is to extract interesting, nontrivial, implicit, previously unknown and potentially useful information from data.
>
> *Introduction to Data Mining in Bioinformatics*
> Jason T.L. Wang et al.[333]

Large-scale sequence and structural data has lead to a wealth of novel proteins, but how can this data be effectively mined for new functions? Homology-based functional annotation is the most commonly used approach, found in sequence-sequence and sequence-model based tools. Annotations are transferred based on sequence similarity thresholds between matches. This is an efficient way to profile a new genome or metagenome, but is based on naive assumptions about function retention amongst similar proteins. Database matches may also be bereft of any experimental data, leaving conserved protein families with very little actual functional information. Even worse are the proteins that evade all homology-based annotation attempts, being so divergent from current database sequences as to end up in a "dark" fraction of unannotated protein sequences. While some of these unannotated sequences may be pseudogenes, proteomics and experimental characterization of previously unknown proteins have shown that this is not always the case. In this thesis, I have explored ways in which annotation can be used to find proteins of interest, when homology-based approaches succeed and when they fail, how often they fail, and what trends are associated with this lack of annotation.

Summary of findings

Various types of homology-based annotations were used for three case studies in Chapter 2. Targeting protein families was sometimes done with a specific database about a certain class of proteins if it existed, or a combination of more general methods if it did not. In order to boost confidence in general annotation predictions, four different methods were compared to find cellulases from two *Streptomyces* strains in Chapter 2.1. A more heavily curated database, focused purely on antibiotic resistance, was used to find antibiotic resistance proteins from a wastewater metagenome in Chapter 2.2. This focused database approach was also used again in Chapter 2.3, after identification of a potential human and fish pathogen from decomposing rainbow darter, to detect virulence factors within its genome. While combining annotation methods and using niche databases is effective when targeting certain protein families, when profiling new sequence data in a broad fashion, databases with pathway information can be useful to piece together annotations into a metabolic story. To this end, a metabolic pathway database was used to discover pathways of interest in Chapters 2.1 and 2.3. Using common annotation tools, with a combination of knowledge about the protein families in question and novel environments can lead to the discovery of proteins worth prioritizing for future experimental study. But, as has been discussed extensively throughout this thesis, homology-based annotation can only go so far. Protein families have not been uniformly characterized and even though extensive efforts have been made in model organisms like *E. coli*, there are taxa evolutionary quite distant from well-studied proteomes that lack substantial annotation coverage. Annotation completeness was thus investigated across the bacterial tree of life in Chapter 3 to see the extent of this phenomenon. Taxonomy was indeed a noticeable differentiator in annotation completeness, along with levels of study of the organism in question, and the organism's genome size. The percentage of genomic coding sequences receiving any kind of database match varied wildly, anywhere from 2 - 86%. This "annotated" fraction can include families of conserved proteins, well represented in databases but poorly or not at all experimentally characterized. Prioritizing these groups for further study can involve alternative methods of deducing functional context to incorporate the collected knowledge about these families and the organisms they are found in, as seen in Chapter 4. Here, environment, lineage, and pathogen-associations were used to distinguish uncharacterized protein families. The remaining "dark matter", left without any database matches after homology-based annotation is complete, was explored in Chapter 5. Named "ORFans", these sequences were shown to be associated with divergent protein families, such as viral RNA ligases. Remote homology techniques combined with genomic context and motif analyses revealed that some functional annotation is possible for even these most extreme cases. Novel protein discovery can be accomplished at all levels of functional annotation success, from easy-to-operate homology-based tools to

combinations of alternative methods, such as metadata associations and remote homology. Furthermore, tailoring the annotation to the situation based on available data and resources, makes novel protein discovery a more effective prospect.

Homology-based annotation

Case studies of homology-based genome and metagenome analysis

Homology detection remains essential for many function prediction methods. In Chapter 2, a main theme was looking at how homology-based annotation can be applied to datasets. One strategy is using large databases to analyze the broad functional profile of a genome or metagenome. The work in Chapter 2.1 used this approach to identify benzoate degredation and medically-relevant enediyene biosynthesis pathways in recently-sequenced *Streptomyces* strains, *Streptomyces* sp. NWU339 and *Streptomyces viridosporus* NWU49. A large database was also used on the rainbow darter necrobiome, revealing that the shifting functional profile reflected the microbial succession detected in the 16S rRNA data, as well as the binning analysis. Early stages of decomposition were dominated by Clostridiaceae and *Aeromonas* with associated pathways of pollutant degradation and biofilm formation also peaking at this time. Rikenellaceae was an especially major community member in the later stages of decomposition, the main contributors to the glycan metabolism and antibiotic synthesis pathways that peaked at the end of the time course.

The more delicate task of targeting specific protein families is better conducted using extensively curated niche databases or combining multiple methods to provide higher confidence results. In order to detect cellulases in the *Streptomyces* genomes, four different annotation methods were compared. Eleven sequences were predicted to be cellulases with all four methods, other predictions possibly being erroneous or other divergent glycosyl hydrolases. One area that some databases and thus annotation struggles with is linking experimental information or other predictive measures with annotations. When questioning an annotation's validity it is sometimes nigh impossible to trace back evidence in large databases like NCBI. This is dangerous for the propagation of the functional term originally assigned. Substitutions and indels can affect protein function and without comparisons to proteins with proven functions, small changes in sequence can accumulate unobserved with annotation transfer. Comparisons across different annotation methods and using profiles to find conserved residues can mitigate this effect, however, the only way to completely verify an annotation is testing.

Focused databases are a way to combine years worth of study on well characterized organisms or protein families. They have been used here in Chapters 2.2 and 2.3, to

contribute to the growing agricultural resistome research and to add to the still small body of work on necrobiome functional profiles. In Chapter 2.2, an antibiotic resistance database was used to profile the resistome in a research farm wastewater metagenome. A total of 31 antibiotic resistance genes (of the type where antibiotic resistance can be determined based on overall sequence similarity versus dependent on mutations) were uncovered in this sample. The most common resistance type was towards tetracycline and streptomycin, having ten and five genes associated with their resistance, respectively. Agricultural spread of antibiotic resistance is a key area of study as antibiotics are so fundamental to our current healthcare practices. Filling in the picture of wastewater resistomes adds to knowledge about the sequence complexity and relative abundance of the large suite of antibiotic resistance genes. In Chapter 2.3, a virulence factor database was used on a putatively pathogenic *Aeromonas veronii* MAG discovered in the fish necrobiome. Three hemolytic sequences were found in the bin, with another present in a smaller *Aeromonas* bin. Of note is that finding hemolytic toxins in dominant members of decomposer communities furthers theories of these toxins possibly playing a role in the decomposition process.[191] If shown to be true with further experimental study, this could broaden the known biological roles of toxins.

These, at the time, newly sequenced datasets, were also used to assess annotation coverage. 31 - 84% of the *Streptomyces* predicted CDSs were annotated by a range of six different annotation methods. The most optimistic annotation methods dropped to 56% for the wastewater metagenome and 58% for the fish necrobiome. Metagenomes are generally harder to annotate than genomes for reasons discussed later. One noticeable shift was that the Rikenellaceae bins had lower annotation completeness than the *Aeromonas* and Selenomonadaceae bins. These MAGs (not found to be clustered phylogenetically with the other *Alistipes* of the Rikenellaceae family) may contain divergent families that are not well studied, specific to life in their environmental habitats (swamp, zebrafish gut, and rainbow darter necrobiome).

Annotation completeness of bacterial genomes

As seen in Chapter 2, different genomes have different levels of annotation. Model organisms such as *Escherichia coli* and *Bacillus subtilis* formed the basis for early bacterial protein characterization and current sequence databases are biased towards human-centric organisms, especially medically-relevant species.[245,254] Now that the bacterial tree of life is rapidly expanding with faster and cheaper sequencing technologies, a thorough evaluation of annotation coverage is needed. Conducted in Chapter 3 is the largest analysis of genome annotation coverage across the bacterial phylogeny. Just over 27,000 genomes were annotated with Prokka, with previously calculated Pfam and KEGG annotations

compared. Annotation incompleteness ranged from 2 - 86%, 3 - 88%, and 1 - 81% with Prokka, KEGG, and Pfam, respectively. With a mean annotation coverage of 52±9% for Prokka, this is almost certainly affecting functional analyses and comparisons of genomes. So many predicted coding sequences are without easily accessible functional information. While it is possible to achieve better annotation coverage using multiple approaches,[98] one must rely on sometimes difficult-to-use pipelines and deal with the non-standardization of functional annotation terms. Investigating the extent to which popular annotation methods struggle with annotation coverage in certain areas of the bacterial phylogeny is important to spread awareness on how much "dark matter" is left behind in analyses of these organisms.

This work supported the observation that taxonomy plays a role in annotation coverage. Taxonomic classification has a significant effect on annotation coverage where phyla with model organisms like Proteobacteria and Firmicutes have a higher proportion of CDSs annotated on average. Recently classified phyla such as Patescibacteria[109,264] had the lowest average annotation completeness values. In a more fine grained look at research bias, genera such as *Escherichia*, *Staphylococcus*, and *Pseudomonas* (with over 75,000 mentions in Pubmed abstracts and titles) had higher average annotation coverage than other genera with fewer instances in Pubmed. This trend was noticeably twisted when obligate symbionts, like *Buchnera aphidicola*, with reduced genomes but close evolutionary relationships with model organisms had even higher annotation completeness. This leads into perhaps the most striking association with annotation completeness, genome size.

Annotation completeness is significantly affected by genome size. Even within taxonomic groups like phyla, orders, and genera, larger genomes trend towards lower annotation coverage. Ranea at al.,[258] found that protein families that did not scale with genome size were associated with translation, ribosomal structure and protein biosynthesis, whereas protein families that scaled either linearly or exponentially included proteins associated with amino acid transport, metabolism, gene regulation, signal transduction, and replication, recombination and repair. Moreover, many of the proteins that scaled linearly and exponentially with genome size, were poorly characterized, possibly being protein families that are not well represented in current databases and/or are undergoing higher rates of functional diversification. As accessory gene gain and loss has been linked to genome size,[20,254,316] these more divergent proteins represent a hard-to-annotate fraction that grows with genome size, allowing an organism to be more flexible and survive in varying conditions.

Within endosymbionts and other organisms undergoing genome reduction, a loss of accessory genes is accompanied by an increased rate of pseudogenization.[20,152,153,218] A higher proportion of pseudogenes, such as in stand-out example *Mycobacterium leprae*,[20] also impacts how annotation coverage scales with genome size. If some of the pseudogenes are treated as CDSs by gene prediction tools, this would potentially inflate the number of

unannotated CDSs. The flip side of this is that small, highly reduced genomes have fewer accessory proteins and therefore seem to be easier to annotate. It has been speculated that as genome reduction proceeds over time, pseudogenes are largely expelled from the genome through deletion bias, resulting in these small genomes.[20,217,218] While the detection of generally shorter pseudogene sequences can be stumbling block in annotation coverage, even translated short sequences can pose problems for annotation. Short proteins have poor database coverage and lower match scores during the annotation process, as well as being harder to identify within the genome. A recent study detected the synthesis of 36 proteins in *E. coli* less than 75 amino acids in length,[325] highlighting an important area in protein research moving forward.

There are several further analyses on this dataset that could help guide future research. Of the proteins that remained unannotated in this study, clustering them into families and identifying those with the largest taxonomic breadth would help prioritize families that would make a large impact on annotation coverage. A similar method looking at the SFams database found 6,668 protein families present in more than one taxonomic class and without any domain annotations which they labelled as "most wanted".[340] An additional approach is to associate higher-level functional terms to the annotated proteins and compare their proportion in genomes across the phylogeny to identify which divergent protein groups are potentially the worst represented in low annotation coverage clades, taking into consideration differences in functional profiles in different clades. This could point to classes of accessory proteins that have divergent, undetected members in understudied lineages. Mining the reservoirs of uncharacterized proteins could significantly improve annotation coverage amongst various taxonomic groups, possibly enabling us to learn about different methods of survival and about alternate opportunities for organisms to thrive.

Alternative approaches for analyzing and inferring protein function

Inferring biological associations for conserved domain families

With substantial proportions of genomes and metagenomes without homology-based annotations (discussed in Chapters 2 and 3), alternative methods are the way forward for uncovering functional clues. How might alternate methods be used for extracting functional information, and how much success can be gained from them? One interesting area to focus on is protein domain families. These are groups of proteins with regions of similar sequence conservation that are thought to fall under the same functional umbrella. Of the 17,929 Pfam v32.0 models, 22% (4049) are domains of unknown functions (DUFs).

Due to their presence in multiple organisms (and sometimes with proteomic data to prove their synthesis), these are most likely real proteins with cellular roles. In order to provide some biological context for these protein domain families, associations with environments, lineages, and pathogens were assessed in Chapter 4. Additional information including abundance and amenability to structural characterization was also analyzed to provide more ways to distinguish protein families.

This work revealed 4357 protein families strongly-associated with either soil, marine or human gut environments, 1056 of which were DUFs. Functions enriched in the environment-associated families included heme-binding (soil), photosynthesis (marine), and carbohydrate metabolic process (human gut). It follows that the DUFs may also have functions that are important for survival in their respective environments. An example discussed in Chapter 4 is DUF4906 (PF16249). This domain family is significantly associated with the human gut samples and is a part of a larger protein grouping that includes fimbriae components, indicating this domain family may be involved in cell adhesion to the human gut epithelium.

As for pathogen association, 2007 domains, including 517 DUFs, were found to be significantly overrepresented in a manually curated pathogen set (representative of a wide-range of hosts). The pathogen-enriched domains included many known virulence factors and toxins. The GO term "pathogenesis" was found to be enriched in the pathogen-associated domains, further supporting the analysis. An interesting finding was that, after determining the lineage-specificity of the domain families, the pathogen-enriched domains with a low lineage-association score showed an increased density of the "pathogenesis" term. Virulence factors are transferred via horizontal gene transfer (frequently found associated with genomic islands[24,103,228,300]), potentially indicating that domains that play a role in virulence are more broadly distributed as opposed to lineage-specific. However, as the GO terms are manually curated this may be due to a bias about which kinds of domains are labelled with this term.

Other ways to combine the association data include finding the intersection of bacterial pathogen-associated domains with domains that are most common in eukaryotes. This results in domain families that are potentially "mimicry" candidates. The concept of mimicry is about exploiting the host's processes with sequentially or structurally-similar proteins. These virulence factors are important in a wide-range of pathogen pathways[57,241,296] and here is reported a list of 49 putative mimicry protein families, including five DUFs. The association data here can also be incorporated with domain family abundance information based on occurrences in NCBI, select metagenomes, and their taxonomic breadth in Pfam proteomes. Additional information such as structural novelty, preference for single-domain architectures, and the frequency of disorded and transmembrane regions are also included for assessing the domain families for structural characterization.

Identifying and prioritizing proteins for experimental characterization can seem daunting with the vast expanse of protein sequences available. Even narrowing those down to a certain set of genomes can still feel like finding a needle in a haystack if there is no or limited accompanying functional information for a portion of the sequences. However, computational inquiries can help guide research, sometimes adding validation to wet-lab discoveries or justifying experimental approaches.[14,360] In order to easily visualize the data generated here, the VirFams (virfams.uwaterloo.ca) application was created. This website allows for the exploration of Pfam domains and the rankings discussed in Chapter 4. An example previously discussed in this thesis is DUF4765 (PF15962). VirFams makes it very straightforward to see that this domain family is overrepresented in pathogens, enriched in human gut metagenomes, present in three phyla and found in a low number of species, highlighting this domain as a priority for virulence factor testing. Also included is an implementation of hmmscan that identifies which of the detected domains may be pathogen-associated. This allows for the domain architecture as a whole to be evaluated for potential pathogen relevance.

Untapped functional novelty lies within protein "dark matter". The data generated and collected here has been made publicly available in order to facilitate research into these domain families and the attempt to identify proteins of interest, namely virulence factors. The alternative method, metadata associations, leveraged known information and database detection of these domain families to yield new biological contexts in which to view the uncharacterized proteins. 41% of all DUFs in Pfam were identified as having strong lineage, environment, and/or pathogen associations, providing more information for just under half of the uncharacterized fraction of this database. Combining different sources of data with expert knowledge (e.g. mimicry being an indicator of virulence) is a powerful way to find novel proteins.

A further point about improving current annotation practices is that more research is needed on the link between domain architecture combinations and function. As seen in Chapter 3, Pfam "annotates" many more proteins than other annotation methods. A large proportion of Pfam's models are DUFs, but even DUFs end up alongside non-DUF models. Domain shuffling is widespread and is a modular way that new proteins are generated.[22,163] How far can domain architecture go in providing information on overall function as well as functional diversification within a protein family? Recombination and shuffling of domains also are common in toxin families, an evolutionary tactic for function diversification.[56,355] It is intriguing to think of patterns in domain architecture being used to identify new toxin families. Nevertheless, the complexity of required models and the trouble in identifying a wide range of toxins versus non-toxins for training has been an obstacle so far.

Metagenomic ORFan annotation

While Chapter 4 looked at conserved protein families, another group of sequences that evade annotation are predicted CDSs which lack detectable homologs within current databases. These ORFans are predominantly found in metagenomes where community complexity, sequencing technology, coverage, final assembly, and coding sequence prediction software can impact the quality and length of the resulting CDSs.[245,254] As discussed above, shorter sequences are harder to annotate. Taxonomic diversity that does not substantially overlap with databases which are generally biased to Firmicutes, Proteobacteria and Actinobacteria also predisposes metagneomes to low annotation coverage.[245] Unlike families with database coverage (which have information about the organisms they are found in, and sometimes limited evidence for associations with other processes or proteins within the cell), there is much less extraneous data that can be used for metagenomic ORFans. Here in Chapter 5, a sensitive model-model sequence match strategy is employed. Briefly, this remote homology method uses conserved residues from known protein families and compares it to conserved residues in clusters of metagenomic ORFans to boost the chances of finding similar proteins. Discussed is how effective this strategy is and how reliable these predictions may be.

Of 35,307,707 predicted coding sequences from three large metagenomic datasets (from soil, marine, and human gut environments), 484,121 sequences were labelled "ORFans". These ORFans had similar GC content distributions to their respective metagenomes but tended to be a bit shorter in length (as was expected based on previous observations). This study revealed that remote homology was able to find significant matches for 73,896 sequences (15%) at a false discovery rate of ∼9%. To validate these predictions, a genomic context approach was used. A comparison of the functional terms associated with ORFans and their direct neighbors on their respective contigs found a high congruence in function, as has been established in bacterial genomes before.[42,81,145,195,271,344]

The ORFans in these soil, marine, and human gut biomes were found to be enriched with viral-associated terms like "viral release from host cell." This viral link to unannotated sequences has been shown previously, speculated to be due to the increased rate of evolution that viruses undergo, and the extreme lack of database coverage of viral sequences.[4,44] Other enriched ORFan functions include dibenzothiophene catabolic process (soil), L-ascorbic acid binding (marine), and polysaccharide catabolic process (human gut). These enriched ORFan functions are likely being underrepresented in metagenome function profiles, and with only 15% of the ORFans analyzed in this way, there are many more functions that are also being excluded.

Metalloproteases were discovered in the pool of ORFan sequences as well, specifically targeted because of their easily identifiable catalytic site motif. 257 clusters of ORFan

sequences had both remote homology to proteases (or peptidases) and contained the HExxH motif. These metalloproteases further reinforce the idea that this dataset is a resource for novel protein discovery. Building up the metalloprotease families with other metagenomic data could be a interesting way to explore and expand this protein family. Other deep dives on protein groups have gone on to analyze the sequential differences in each subtype, and combine knowledge about characterized versions of these sequences in order to inform new insights about novel branches of the family.[360]

This is the first large-scale analyses of how remote homology can be used to uncover functional information about ORFans. Here I've shown that this alternative technique does allow researchers to gain some ground in the "dark" corners of the protein universe, albeit in a computationally intensive way. There many functional associations reported here that are almost certainly being under reported in metagenomic analyses. This set of detected ORFans is a way to explore some of those functions, and a place to look for proteins of interest. Ultimately, as seen in Chapters 4 and 5, alternative annotation methods provide an avenue for expanding annotation coverage.

Final remarks

Modified from Lobb and Doxey (2016).[177]

The enormous diversity of protein sequences is both a challenge and an exciting resource for protein function discovery. There are numerous paths to finding functional novelty in sequence data. New functionality may be revealed not only by finding novel proteins ("dark matter"), but also by identifying homologs of conserved function in unexpected settings (such as in new species) and assessing associations of conserved uncharacterized families with known biological data. Thanks to increased coverage of structure and domain space, the accuracy of function prediction has improved and recent studies, including this thesis, are integrating data and methods in increasingly powerful ways and on larger scales. In addition, better annotations are making it easier to sort through genome-wide function predictions and discover new and unexpected biological phenomena. However, improved methods for predicting the impacts of substitutions and indels on protein function are critical to identify and interpret biological roles and functional differentiation. Furthermore, methods are needed to better predict the functional consequences of different domain combinations, which are so widespread in sequence data. Finally, even when annotation coverage of genomes and metagenomes can be lacking, advances in annotation such as remote homology and other alternate computational methods help widen the net of annotation coverage and lead to more protein discoveries. Developments in these areas will ultimately help to

more sensitively identify the adaptive variations that differentiate function within known and novel protein families, and extract functional novelty from ever-growing bioinformatic datasets.

Letter of Copyright Permission

The published articles included in this thesis are made with the permission of all publishers. At the time of writing (June 2020), the policies of the publishers Elsevier, American Society for Microbiology, and Frontiers do not require additional permissions for the reproduction of articles in theses. The relevant policies can be found at the following links:

1. Elsevier

   https://www.elsevier.com/about/policies/copyright

2. American Society for Microbiology

   https://journals.asm.org/content/statement-author-rights

   https://msystems.asm.org/content/editorial-policy

3. Frontiers

   https://www.frontiersin.org/about/policies-and-publication-ethics

For the article[180] published by Microbiology Society, the authors signed a Creative Commons licence. © [Briallen Lobb, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, and Andrew C. Doxey, 2020]. The definitive peer reviewed, edited version of this article is published in [Microbial Genomics, 6:3, 2020, doi:10.1099/mgen.0.000341].

References

[1] S. Abhiman. FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Research*, 33(Database issue):D197–D200, 2004.

[2] M. F. Adegboye, B. Lobb, O. O. Babalola, A. C. Doxey, and K. Ma. Draft genome sequences of two novel cellulolytic *Streptomyces* strains isolated from South African rhizosphere soil. *Genome Announcements*, 6(26):e00632–18, 2018.

[3] O. A. Adekoya and I. Sylte. The thermolysin family (M4) of enzymes: Therapeutic and biotechnological potential. *Chemical Biology & Drug Design*, 73(1):7–16, 2009.

[4] V. Aggarwala, G. Liang, and F. D. Bushman. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mobile DNA*, 8(1):12, 2017.

[5] E. Akiva, S. Brown, D. E. Almonacid, A. E. Barber, A. F. Custer, M. A. Hicks, C. C. Huang, F. Lauck, S. T. Mashiyama, E. C. Meng, et al. The structure-function linkage database. *Nucleic Acids Research*, 42(D1):D521–D530, 2013.

[6] J. Alneberg, B. S. Bjarnason, I. De Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, 2014.

[7] S. Altschul. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[9] J. O. Andersson and S. G. E. Andersson. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Molecular Biology and Evolution*, 18(5):829–839, 2001.

[10] A. Angulo, J. Nunnery, H. Bair, and W. Wint. Antimicrobial resistance in zoonotic enteric pathogens. *Revue Scientifique et Technique de l'OIE*, 23(2):485–496, 2004.

[11] K. Arakawa, Y. Nakayama, and M. Tomita. GPAC: Benchmarking the sensitivity of genome informatics analysis to genome annotation completeness. *In Silico Biology*, 6(1-2):49–60, 2006.

[12] A. Avanthi, S. Kumar, K. C. Sherpa, and R. Banerjee. Bioconversion of hemicelluloses of lignocellulosic biomass to ethanol: an attempt to utilize pentose sugars. *Biofuels*, 8(4):431–444, 2016.

[13] C. Balachandran, V. Duraipandiyan, K. Balakrishna, and S. Ignacimuthu. Petroleum and polycyclic aromatic hydrocarbons (PAHs) degradation and naphthalene metabolism in *Streptomyces* sp. (ERI-CPDA-1) isolated from oil contaminated soil. *Bioresource Technology*, 112:83–90, 2012.

[14] K. Bastard, A. A. T. Smith, C. Vergne-Vaxelaire, A. Perret, A. Zaparucha, R. D. Melo-Minardi, A. Mariage, M. Boutard, A. Debard, C. Lechaplais, et al. Revealing the hidden functional diversity of an enzyme family. *Nature Chemical Biology*, 10(1):42–49, 2013.

[15] C. Beaulieu, M. Khalil, S. Lerat, and N. Beaudoin. The plant pathogenic bacterium *Streptomyces scabies* degrades the aromatic components of potato periderm via the β-ketoadipate pathway. *Frontiers in Microbiology*, 10:2795, 2019.

[16] O. Beja. Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science*, 289(5486):1902–1906, 2000.

[17] L. M. Berent and J. B. Messick. Physical map and genome sequencing survey of *Mycoplasma haemofelis* (*Haemobartonella felis*). *Infection and Immunity*, 71(6):3657–3662, 2003.

[18] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[19] J. A. Blake, K. R. Christie, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, S. Burgess, T. Buza, C. Gresham, et al. Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.

[20] L. M. Bobay and H. Ochman. The evolution of bacterial genome architecture. *Frontiers in Genetics*, 8(72), 2017.

[21] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8):852–857, 2019.

[22] E. Bornberg-Bauer and M. M. Alba. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology*, 23(3):459–466, 2013.

[23] A. Böttger, A. C. Doxey, M. W. Hess, K. Pfaller, W. Salvenmoser, R. Deutzmann, A. Geissner, B. Pauly, J. Altstätter, S. Münder, et al. Horizontal gene transfer contributed to the evolution of extracellular surface structures: The freshwater polyp hydra is covered by a complex fibrous cuticle containing glycosaminoglycans and proteins of the PPOD and SWT (sweet tooth) families. *PLoS ONE*, 7(12):e52278, 2012.

[24] E. F. Boyd and H. Brüssow. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends in Microbiology*, 10(11):521–529, 2002.

[25] A. Broom, K. Trainor, D. W. MacKenzie, and E. M. Meiering. Using natural sequences and modularity to design common and novel protein topologies. *Current Opinion in Structural Biology*, 38:26–36, 2016.

[26] S. D. Brown and P. C. Babbitt. New insights about enzyme evolution from large scale studies of sequence and structure relationships. *Journal of Biological Chemistry*, 289(44):30221–30228, 2014.

[27] B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2014.

[28] Z. M. Burcham, J. L. Pechal, C. J. Schmidt, J. L. Bose, J. W. Rosch, M. E. Benbow, and H. R. Jordan. Bacterial community succession, transmigration, and differential gene transcription in a controlled vertebrate decomposition model. *Frontiers in Microbiology*, 10, 2019.

[29] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7):581–583, 2016.

[30] J. H. Campbell, P. O'Donoghue, A. G. Campbell, P. Schwientek, A. Sczyrba, T. Woyke, D. Soll, and M. Podar. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proceedings of the National Academy of Sciences*, 110(14):5540–5545, 2013.

[31] I. Can, G. T. Javan, A. E. Pozhitkov, and P. A. Noble. Distinctive thanatomicrobiome signatures found in the blood and internal organs of humans. *Journal of Microbiological Methods*, 106:1–7, 2014.

[32] B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research*, 37(Database):D233—-D238, 2009.

[33] D. O. Carter, D. Yellowlees, and M. Tibbett. Cadaver decomposition in terrestrial ecosystems. *Naturwissenschaften*, 94(1):12–24, 2006.

[34] Y.-C. Chang, Z. Hu, J. Rachlin, B. P. Anton, S. Kasif, R. J. Roberts, and M. Steffen. COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. *Nucleic Acids Research*, 44(D1):D330–D335, 2015.

[35] L. B. Chemes, G. de Prat-Gay, and I. E. Sánchez. Convergent evolution and mimicry of protein linear motifs in host-pathogen interactions. *Current Opinion in Structural Biology*, 32:91–101, 2015.

[36] K. L. Cobaugh, S. M. Schaeffer, and J. M. DeBruyn. Functional and structural succession of soil microbial communities below decomposing human cadavers. *PLoS ONE*, 10(6):e0130201, 2015.

[37] D. Cortez, P. Forterre, and S. Gribaldo. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biology*, 10(6):R65, 2009.

[38] M. Cygler, J. D. Schrag, J. L. Sussman, M. Harel, I. Silman, M. K. Gentry, and B. P. Doctor. Relationship between sequence conservation and three-dimensional structure in a large family of esterases, lipases, and related proteins. *Protein Science*, 2(3):366–382, 1993.

[39] R. D'Adamo, S. Pelosi, P. Trotta, and G. Sansone. Bioaccumulation and biomagnification of polycyclic aromatic hydrocarbons in aquatic organisms. *Marine Chemistry*, 56(1-2):45–49, 1997.

[40] H. Daims, E. V. Lebedeva, P. Pjevac, P. Han, C. Herbold, M. Albertsen, N. Jehmlich, M. Palatinszky, J. Vierheilig, A. Bulaev, et al. Complete nitrification by *Nitrospira* bacteria. *Nature*, 528(7583):504–509, 2015.

[41] A. Danchin and G. Fang. Unknown unknowns: essential genes in quest for function. *Microbial Biotechnology*, 9(5):530–540, 2016.

[42] T. Dandekar. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–328, 1998.

[43] S. Das, D. Lee, I. Sillitoe, N. L. Dawson, J. G. Lees, and C. A. Orengo. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics*, 31(21):3460–3467, 2015.

[44] V. Daubin. Bacterial genomes as new gene homes: The genealogy of ORFans in *E. coli*. *Genome Research*, 14(6):1036–1042, 2004.

[45] J. Davies and D. Davies. Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews*, 74(3):417–433, 2010.

[46] J. Davies, H. Wang, T. Taylor, K. Warabi, X.-H. Huang, and R. J. Andersen. Uncialamycin, a new enediyne antibiotic. *Organic Letters*, 7(23):5233–5236, 2005.

[47] K. L. DeBord, V. T. Lee, and O. Schneewind. Roles of LcrG and LcrV during type III targeting of effector yops by *Yersinia enterocolitica*. *Journal of Bacteriology*, 183(15):4588–4598, 2001.

[48] C. Desler, P. Suravajhala, M. Sanderhoff, M. Rasmussen, and L. J. Rasmussen. *In silico* screening for functional candidates amongst hypothetical proteins. *BMC Bioinformatics*, 10(1):289, 2009.

[49] B. K. Dhillon, M. R. Laird, J. A. Shay, G. L. Winsor, R. Lo, F. Nizam, S. K. Pereira, N. Waglechner, A. G. McArthur, M. G. I. Langille, and F. S. L. Brinkman. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Research*, 43(W1):W104–W108, 2015.

[50] H. Dinkel, K. V. Roey, S. Michael, N. E. Davey, R. J. Weatheritt, D. Born, T. Speck, D. Krüger, G. Grebnev, M. Kubań, et al. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Research*, 42(D1):D259–D266, 2013.

[51] A. J. Doherty, S. R. Ashford, H. S. Subramanya, and D. B. Wigley. Bacteriophage T7 DNA ligase. *Journal of Biological Chemistry*, 271(19):11083–11089, 1996.

[52] H. T. Dong, C. Techatanakitarnan, P. Jindakittikul, A. Thaiprayoon, S. Taengphu, W. Charoensapsri, P. Khunrae, T. Rattanarojpong, and S. Senapin. *Aeromonas jandaei* and *Aeromonas veronii* caused disease and mortality in nile tilapia, *Oreochromis niloticus* (l.). *Journal of Fish Diseases*, 40(10):1395–1403, 2017.

[53] A. C. Doxey, Z. Cheng, B. A. Moffatt, and B. J. McConkey. Structural motif screening reveals a novel, conserved carbohydrate-binding surface in the pathogenesis-related protein PR-5d. *BMC Structural Biology*, 10(1):23, 2010.

[54] A. C. Doxey, D. A. Kurtz, M. D. Lynch, L. A. Sauder, and J. D. Neufeld. Aquatic metagenomes implicate *Thaumarchaeota* in global cobalamin production. *The ISME Journal*, 9(2):461–471, 2014.

[55] A. C. Doxey, M. D. J. Lynch, K. M. Müller, E. M. Meiering, and B. J. McConkey. Insights into the evolutionary origins of clostridial neurotoxins from analysis of the *Clostridium botulinum* strain A neurotoxin gene cluster. *BMC Evolutionary Biology*, 8(1):316, 2008.

[56] A. C. Doxey, M. J. Mansfield, and B. Lobb. Exploring the evolution of virulence factors through bioinformatic data mining. *mSystems*, 4(3):e00162–19, 2019.

[57] A. C. Doxey and B. J. McConkey. Prediction of molecular mimicry candidates in human pathogenic bacteria. *Virulence*, 4(6):453–466, 2013.

[58] C.-J. Duan and J.-X. Feng. Mining metagenomes for novel cellulase genes. *Biotechnology Letters*, 32(12):1765–1775, 2010.

[59] A. S. Duarte, A. Correia, and A. C. Esteves. Bacterial collagenases - A review. *Critical Reviews in Microbiology*, 42(1):106–126, 2014.

[60] G. F. Duarte, A. S. Rosado, L. Seldin, W. de Araujo, and J. D. van Elsas. Analysis of bacterial community structure in sulfurous-oil-containing soils and detection of species carrying dibenzothiophene desulfurization (dsz) genes. *Applied and Environmental Microbiology*, 67(3):1052–1062, 2001.

[61] B. Dujon. The yeast genome project: what did we learn? *Trends in Genetics*, 12(7):263–270, 1996.

[62] B. E. Dutilh, L. Backus, R. A. Edwards, M. Wels, J. R. Bayjanov, and S. A. F. T. van Hijum. Explaining microbial phenotypes on a genomic scale: GWAS for microbes. *Briefings in Functional Genomics*, 12(4):366–380, 2013.

[63] U. Eckhard, H. Bandukwala, M. J. Mansfield, G. Marino, J. Cheng, I. Wallace, T. Holyoak, T. C. Charles, J. Austin, C. M. Overall, et al. Discovery of a proteolytic flagellin family in diverse bacterial phyla that assembles enzymatically active flagella. *Nature Communications*, 8(1):1–9, 2017.

[64] S. R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[65] S. R. Eddy. A new generation of homology search tools based on probabilistic inference. *Genome Informatics*, 23(1):205–211, 2009.

[66] R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

[67] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, et al. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, 2019.

[68] K. Ellrott, L. Jaroszewski, W. Li, J. C. Wooley, and A. Godzik. Expansion of the protein repertoire in newly explored environments: Human gut microbiome specific protein families. *PLoS Computational Biology*, 6(6):e1000798, 2010.

[69] A. M. Eren, Özcan C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, 2015.

[70] H. Fan, D. S. Hitchcock, R. D. Seidel, B. Hillerich, H. Lin, S. C. Almo, A. Sali, B. K. Shoichet, and F. M. Raushel. Assignment of pterin deaminase activity to an enzyme of unknown function guided by homology modeling and docking. *Journal of the American Chemical Society*, 135(2):795–803, 2013.

[71] J. Fastrez. Phage lysozymes. *Experientia Supplementum*, 75:35–64, 1996.

[72] N. Fierer, J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, J. A. Gilbert, D. H. Wall, and J. G. Caporaso. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, 109(52):21390–21395, 2012.

[73] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, et al. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222—-D230, 2013.

[74] R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server):W29–W37, 2011.

[75] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 2016.

[76] D. Fischer and D. Eisenberg. Finding families for genomic ORFans. *Bioinformatics*, 15(9):759–762, 1999.

[77] H. J. Flint, E. A. Bayer, M. T. Rincon, R. Lamed, and B. A. White. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Reviews Microbiology*, 6(2):121–131, 2008.

[78] D. E. Fouts, M. A. Matthias, H. Adhikarla, B. Adler, L. Amorim-Santos, D. E. Berg, D. Bulach, A. Buschiazzo, Y.-F. Chang, R. L. Galloway, et al. What makes a bacterial species pathogenic?: Comparative genomic analysis of the genus *Leptospira*. *PLOS Neglected Tropical Diseases*, 10(2):e0004403, 2016.

[79] N. Furnham, N. L. Dawson, S. A. Rahman, J. M. Thornton, and C. A. Orengo. Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. *Journal of Molecular Biology*, 428(2):253–267, 2016.

[80] M. Y. Galperin. Conserved 'hypothetical' proteins: New hints and new puzzles. *Comparative and Functional Genomics*, 2(1):14–18, 2001.

[81] M. Y. Galperin and E. V. Koonin. Who's your neighbor? New computational approaches for functional genomics. *Nature Biotechnology*, 18(6):609–613, 2000.

[82] M. Y. Galperin and E. V. Koonin. 'Conserved hypothetical' proteins: Prioritization of targets for experimental study. *Nucleic Acids Research*, 32(18):5452–5463, 2004.

[83] S. Ghosh and T. M. LaPara. The effects of subtherapeutic antibiotic use in farm animals on the proliferation and persistence of antibiotic resistance among soil bacteria. *The ISME Journal*, 1(3):191–203, 2007.

[84] J. Gibson and C. S. Harwood. Metabolic diversity in aromatic compound utilization by anaerobic microbes. *Annual Reviews in Microbiology*, 56(1):345–369, 2002.

[85] R. Gil, B. Sabater-Muñoz, A. Latorre, F. J. Silva, and A. Moya. Extreme genome reduction in *Buchnera* spp.: Toward the minimal genome needed for symbiotic life. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7):4454–4458, 2002.

[86] H. J. Gilbert and G. P. Hazlewood. Bacterial cellulases and xylanases. *Microbiology*, 139(2):187–194, 1993.

[87] J. A. Gilbert, D. Field, Y. Huang, R. Edwards, W. Li, P. Gilna, and I. Joint. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE*, 3(8):e3042, 2008.

[88] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–1359, 2006.

[89] V. Gligorijević, M. Barot, and R. Bonneau. deepNF: deep network fusion for protein function prediction. *Bioinformatics*, 34(22):3873–3881, 2018.

[90] A. Godzik. Metagenomics and the protein universe. *Current Opinion in Structural Biology*, 21(3):398–403, 2011.

[91] E. Goldschmidt-Clermont, T. Wahli, J. Frey, and S. E. Burr. Identification of bacteria from the normal flora of perch, *Perca fluviatilis* l., and evaluation of their inhibitory potential towards *Aeromonas* species. *Journal of Fish Diseases*, 31(5):353–359, 2008.

[92] J. M. González, L. Hernández, I. Manzano, and C. Pedrós-Alió. Functional annotation of orthologs in metagenomes: a case study of genes for the transformation of oceanic dimethylsulfoniopropionate. *The ISME Journal*, 13(5):1183–1197, 2019.

[93] C. Gonzalez-Serrano, J. Santos, M. Garcia-Lopez, and A. Otero. Virulence markers in *Aeromonas hydrophila* and *Aeromonas veronii* biovar sobria isolates from freshwater fish and from a diarrhoea case. *Journal of Applied Microbiology*, 93(3):414–419, 2002.

[94] N. F. Goodacre, D. L. Gerloff, and P. Uetz. Protein domains of unknown function are essential in bacteria. *mBio*, 5(1), 2013.

[95] A. Gorvitovskaia, S. P. Holmes, and S. M. Huse. Interpreting *Prevotella* and *Bacteroides* as biomarkers of diet and lifestyle. *Microbiome*, 4(1):15, 2016.

[96] M. Gouw, S. Michael, H. Sámano-Sánchez, M. Kumar, A. Zeke, B. Lang, B. Bely, L. B. Chemes, N. E. Davey, Z. Deng, et al. The eukaryotic linear motif resource–2018 update. *Nucleic Acids Research*, 46(D1):D428–D434, 2018.

[97] M. Grabowski, E. Niedzialkowska, M. D. Zimmerman, and W. Minor. The impact of structural genomics: the first quindecennial. *Journal of Structural and Functional Genomics*, 17(1):1–16, 2016.

[98] M. Griesemer, J. A. Kimbrel, C. E. Zhou, A. Navid, and P. D'Haeseleer. Combining multiple functional annotation tools increases coverage of metabolic annotation. *BMC Genomics*, 19(1), 2018.

[99] S. R. Grossman, K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, D. J. Park, D. Griesemer, E. K. Karlsson, S. H. Wong, et al. Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4):703–713, 2013.

[100] J. Guo, X. Fu, H. Liao, Z. Hu, L. Long, W. Yan, Y. Ding, L. Zha, Y. Guo, J. Yan, et al. Potential use of bacterial community succession for estimating post-mortem interval as revealed by high-throughput sequencing. *Scientific Reports*, 6(1), 2016.

[101] R. S. Gupta and D. W. Mathews. Signature proteins for the major clades of Cyanobacteria. *BMC Evolutionary Biology*, 10(1):24, 2010.

[102] H. Guturu, A. C. Doxey, A. M. Wenger, and G. Bejerano. Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632):20130029, 2013.

[103] C. Gyles and P. Boerlin. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary Pathology*, 51(2):328–340, 2013.

[104] D. H. Haft, M. DiCuccio, A. Badretdin, V. Brover, V. Chetvernin, K. O'Neill, W. Li, F. Chitsaz, M. K. Derbyshire, N. R. Gonzales, et al. RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, 46(D1):D851–D860, 2018.

[105] D. H. Haft, J. D. Selengut, and O. White. The TIGRFAMs database of protein families. *Nucleic Acids Research*, 31(1):371–373, 2003.

[106] J. Handelsman. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4):669–685, 2004.

[107] A. Haritash and C. Kaushik. Biodegradation aspects of polycyclic aromatic hydrocarbons (PAHs): A review. *Journal of Hazardous Materials*, 169(1-3):1–15, 2009.

[108] E. D. Harrington, A. H. Singh, T. Doerks, I. Letunic, C. von Mering, L. J. Jensen, J. Raes, and P. Bork. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proceedings of the National Academy of Sciences*, 104(35):13913–13918, 2007.

[109] B. P. Hedlund, J. A. Dodsworth, S. K. Murugapiran, C. Rinke, and T. Woyke. Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". *Extremophiles*, 18(5):865–875, 2014.

[110] J.-H. Hehemann, A. G. Kelly, N. A. Pudlo, E. C. Martens, and A. B. Boraston. Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. *Proceedings of the National Academy of Sciences*, 109(48):19786–19791, 2012.

[111] G. L. Holliday, S. A. Rahman, N. Furnham, and J. M. Thornton. Exploring the biological and chemical complexity of the ligases. *Journal of Molecular Biology*, 426(10):2098–2111, 2014.

[112] T. A. Hopf, C. P. I. Schärfe, J. P. G. L. M. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. J. J. Bonvin, and D. S. Marks. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3, 2014.

[113] A. C. Howe, J. K. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown. Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences*, 111(13):4904–4909, 2014.

[114] P. Hu, S. C. Janga, M. Babu, J. J. Diaz-Mejia, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, et al. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biology*, 7(4):e1000096, 2009.

[115] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, et al. InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37(Database):D211–D215, 2009.

[116] D. Hyatt, G. L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), 2010.

[117] E. R. Hyde, D. P. Haarmann, J. F. Petrosino, A. M. Lynne, and S. R. Bucheli. Initial insights into bacterial succession during human decomposition. *International Journal of Legal Medicine*, 129(3):661–671, 2014.

[118] J. Ijaq, M. Chandrasekharan, R. Poddar, N. Bethi, and V. S. Sundararajan. Annotation and curation of uncharacterized proteins- challenges. *Frontiers in Genetics*, 6(119), 2015.

[119] A. Jain and D. Kihara. Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences. *Bioinformatics*, 35(5):753–759, 2019.

[120] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru. High throughput ANI analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):1–8, 2018.

[121] E. Jami and I. Mizrahi. Composition and similarity of bovine rumen microbiota across individual animals. *PloS ONE*, 7(3), 2012.

[122] J. M. Janda, S. L. Abbott, and C. J. McIver. *Plesiomonas shigelloides* revisited. *Clinical Microbiology Reviews*, 29(2):349–374, 2016.

[123] L. Jaroszewski, Z. Li, S. S. Krishna, C. Bakolitsa, J. Wooley, A. M. Deacon, I. A. Wilson, and A. Godzik. Exploration of uncharted regions of the protein universe. *PLoS Biology*, 7(9):e1000205, 2009.

[124] G. T. Javan, S. J. Finley, I. Can, J. E. Wilkinson, J. D. Hanson, and A. M. Tarone. Human thanatomicrobiome succession and time since death. *Scientific Reports*, 6(1), 2016.

[125] B. Jia, A. R. Raphenya, B. Alcock, N. Waglechner, P. Guo, K. K. Tsang, B. A. Lago, B. M. Dave, S. Pereira, A. N. Sharma, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1):D566–D573, 2016.

[126] Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. D'Andrea, R. Lepore, C. S. Funk, I. Kahanda, K. M. Verspoor, A. Ben-Hur, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1):184, 2016.

[127] Y. Jiang, C. Xie, G. Yang, X. Gong, X. Chen, L. Xu, and B. Bao. Cellulase-producing bacteria of *Aeromonas* are dominant and indigenous in the gut of *Ctenopharyngodon idellus* (Valenciennes). *Aquaculture Research*, 42(4):499–505, 2011.

[128] S. W. Joseph, A. M. Carnahan, P. R. Brayton, G. R. Fanning, R. Almazan, C. Drabick, E. W. Trudo, and R. R. Colwell. *Aeromonas jandaei* and *Aeromonas veronii* dual infection of a human wound following aquatic exposure. *Journal of Clinical Microbiology*, 29(3):565–569, 1991.

[129] N. A. Joshi and J. N. Fass. Sickle: a sliding-window, adaptive, quality-based trimming tool for fastq files. https://github.com/najoshi/sickle, 2011.

[130] H. Kaessmann. Origins, evolution, and phenotypic impact of new genes. *Genome Research*, 20(10):1313–1326, 2010.

[131] M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[132] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database):D480–D484, 2008.

[133] M. Kanehisa, Y. Sato, and K. Morishima. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology*, 428(4):726–731, 2016.

[134] H. J. Kang, A. D. Wilkins, O. Lichtarge, and T. G. Wensel. Determinants of endogenous ligand specificity divergence among metabotropic glutamate receptors. *Journal of Biological Chemistry*, 290(5):2870–2878, 2014.

[135] A. E. Kaoutari, F. Armougom, J. I. Gordon, D. Raoult, and B. Henrissat. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, 11(7):497–504, 2013.

[136] L. Katz and R. H. Baltz. Natural product discovery: past, present, and future. *Journal of Industrial Microbiology & Biotechnology*, 43(2-3):155–176, 2016.

[137] K. A. Kazimierczak, H. J. Flint, and K. P. Scott. Comparative analysis of sequences flanking tet(W) resistance genes in multiple species of gut bacteria. *Antimicrobial Agents and Chemotherapy*, 50(8):2632–2639, 2006.

[138] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, and M. J. Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6):845–858, 2015.

[139] L. A. Kelley and M. J. E. Sternberg. Protein structure prediction on the web: a case study using the Phyre server. *Nature Protocols*, 4(3):363–371, 2009.

[140] K. Khafizov, C. Madrid-Aliste, S. C. Almo, and A. Fiser. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proceedings of the National Academy of Sciences*, 111(10):3733–3738, 2014.

[141] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12):1721–1729, 2016.

[142] A. Klotz, J. Georg, L. Bučinská, S. Watanabe, V. Reimann, W. Januszewski, R. Sobotka, D. Jendrossek, W. Hess, and K. Forchhammer. Awakening of a dormant cyanobacterium from nitrogen chlorosis reveals a genetically determined program. *Current Biology*, 26(21):2862–2872, 2016.

[143] E. V. Koonin. How many genes can make a cell: The minimal-gene-set concept. *Annual Review of Genomics and Human Genetics*, 1(1):99–116, 2000.

[144] E. V. Koonin and Y. I. Wolf. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719, 2008.

[145] J. O. Korbel, L. J. Jensen, C. von Mering, and P. Bork. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology*, 22(7):911–917, 2004.

[146] A. Kozińska. Dominant pathogenic species of mesophilic aeromonads isolated from diseased and healthy fish cultured in Poland. *Journal of Fish Diseases*, 30(5):293–301, 2007.

[147] B. B. Kragelund, K. Poulsen, K. V. Andersen, T. Baldursson, J. B. Krøll, T. B. Neergård, J. Jepsen, P. Roepstorff, K. Kristiansen, F. M. Poulsen, et al. Conserved residues and their role in the structure, function, and stability of acyl-coenzyme a binding protein. *Biochemistry*, 38(8):2386–2394, 1999.

[148] F. Krueger. Trim Galore!: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, 2015.

[149] D. B. Kuchibhatla, W. A. Sherman, B. Y. W. Chung, S. Cook, G. Schneider, B. Eisenhaber, and D. G. Karlin. Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently "orphan" viral proteins. *Journal of Virology*, 88(1):10–20, 2013.

[150] M. Kumar, M. Gouw, S. Michael, H. Sámano-Sánchez, R. Pancsa, J. Glavina, A. Diakogianni, J. A. Valverde, D. Bukirova, J. Čalyševa, et al. ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Research*, 48(D1):D296–D306, 2020.

[151] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081, 2009.

[152] C. H. Kuo, N. A. Moran, and H. Ochman. The consequences of genetic drift for bacterial genome complexity. *Genome Research*, 19(8):1450–1454, 2009.

[153] C. H. Kuo and H. Ochman. The extinction dynamics of bacterial pseudogenes. *PLoS Genetics*, 6(8):e1001050, 2010.

[154] K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Research*, 14(4):169–181, 2007.

[155] M. Könneke, A. E. Bernhard, J. R. de la Torre, C. B. Walker, J. B. Waterbury, and D. A. Stahl. Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, 437(7058):543–546, 2005.

[156] A. P. H. Y. M. G. L. J. Lu, Y. Xia. Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15(7):945–953, 2005.

[157] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[158] J. Larsbrink, T. E. Rogers, G. R. Hemsworth, L. S. McKee, A. S. Tauzin, O. Spadiut, S. Klinter, N. A. Pudlo, K. Urs, N. M. Koropatkin, et al. A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature*, 506(7489):498–502, 2014.

[159] J. M. Larsen. The immune response to *Prevotella* bacteria in chronic inflammatory disease. *Immunology*, 151(4):363–374, 2017.

[160] D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8(12):995–1005, 2007.

[161] G.-H. Lee, M.-S. Rhee, D.-H. Chang, J. Lee, S. Kim, M. H. Yoon, and B.-C. Kim. *Oscillibacter ruminantium* sp. nov., isolated from the rumen of korean native cattle. *International Journal of Systematic and Evolutionary Microbiology*, 63(6):1942–1946, 2013.

[162] Y. Lee, Y. Lee, and C. O. Jeon. Biodegradation of naphthalene, BTEX, and aliphatic hydrocarbons by *Paraburkholderia aromaticivorans* BN5 isolated from petroleum-contaminated soil. *Scientific Reports*, 9(1), 2019.

[163] J. G. Lees, N. L. Dawson, I. Sillitoe, and C. A. Orengo. Functional innovation from changes in protein domains and their combinations. *Current Opinion in Structural Biology*, 38:44–52, 2016.

[164] E. Lerat and H. Ochman. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Research*, 33(10):3125–3132, 2005.

[165] I. Letunic and P. Bork. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, 47(W1):W256–W259, 2019.

[166] M. Levitt. Nature of the protein universe. *Proceedings of the National Academy of Sciences*, 106(27):11079–11084, 2009.

[167] G. R. Lewin, C. Carlos, M. G. Chevrette, H. A. Horn, B. R. McDonald, R. J. Stankey, B. G. Fox, and C. R. Currie. Evolution and ecology of Actinobacteria and their bioenergy applications. *Annual Review of Microbiology*, 70(1):235–254, 2016.

[168] T. E. Lewis, I. Sillitoe, A. Andreeva, T. L. Blundell, D. W. Buchan, C. Chothia, D. Cozzetto, J. M. Dana, I. Filippis, J. Gough, et al. Genome3D: exploiting structure to help users understand their sequences. *Nucleic Acids Research*, 43(D1):D382–D386, 2014.

[169] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[170] H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[171] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[172] O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, 257(2):342–358, 1996.

[173] T. Lingner, S. Mühlhausen, T. Gabaldón, C. Notredame, and P. Meinicke. Predicting phenotypic traits of prokaryotes from protein domain frequencies. *BMC Bioinformatics*, 11(1), 2010.

[174] B. Liu, D. Zheng, Q. Jin, L. Chen, and J. Yang. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Research*, 47(D1):D687–D692, 2018.

[175] Y. Liu, P. M. Harrison, V. Kunin, and M. Gerstein. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biology*, 5(9):R64, 2004.

[176] B. Lobb, A. A. Adegoke, K. Ma, A. C. Doxey, and O. A. Aiyegoro. Metagenomic sequencing of wastewater from a South African research farm. *Microbiology Resource Announcements*, 7(16):e01323–18, 2018.

[177] B. Lobb and A. C. Doxey. Novel function discovery through sequence and structural data mining. *Current Opinion in Structural Biology*, 38:53–61, 2016.

[178] B. Lobb, R. Hodgson, M. D. Lynch, M. J. Mansfield, J. Cheng, T. C. Charles, J. D. Neufeld, P. M. Craig, and A. C. Doxey. Time series resolution of the fish necrobiome reveals a decomposer succession involving toxigenic bacterial pathogens. *mSystems*, 5(2), 2020.

[179] B. Lobb, D. A. Kurtz, G. Moreno-Hagelsieb, and A. C. Doxey. Remote homology and the functions of metagenomic dark matter. *Frontiers in Genetics*, 6:234, 2015.

[180] B. Lobb, B. J.-M. Tremblay, G. Moreno-Hagelsieb, and A. C. Doxey. An assessment of genome annotation coverage across the bacterial tree of life. *Microbial Genomics*, 6(3):e000341, 2020.

[181] Y. Loewenstein, D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, and A. Tramontano. Protein function annotation by homology-based inference. *Genome Biology*, 10(2):207, 2009.

[182] S. Louca, F. Mazel, M. Doebeli, and L. W. Parfrey. A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biology*, 17(2), 2019.

[183] S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research*, 48(D1):D265–D268, 2020.

[184] T. Lukk, A. Sakai, C. Kalyanaraman, S. D. Brown, H. J. Imker, L. Song, A. A. Fedorov, E. V. Fedorov, R. Toro, B. Hillerich, et al. Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proceedings of the National Academy of Sciences*, 109(11):4122–4127, 2012.

[185] S. Lutz. Beyond directed evolution - semi-rational protein engineering and design. *Current Opinion in Biotechnology*, 21(6):734–743, 2010.

[186] R. I. Mackie, R. I. Aminov, W. Hu, A. V. Klieve, D. Ouwerkerk, M. A. Sundset, and Y. Kamagata. Ecology of uncultivated *Oscillospira* species in the rumen of cattle, sheep, and reindeer as assessed by microscopy and molecular approaches. *Applied and Environmental Microbiology*, 69(11):6808–6815, 2003.

[187] K. S. Makarova. A DNA repair system specific for thermophilic archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Research*, 30(2):482–496, 2002.

[188] K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, and E. V. Koonin. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct*, 1(1):7, 2006.

[189] K. S. Makarova, D. H. Haft, R. Barrangou, S. J. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. M. Mojica, Y. I. Wolf, A. F. Yakunin, et al. Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology*, 9(6):467–477, 2011.

[190] M. J. Mansfield, J. B. Adams, and A. C. Doxey. Botulinum neurotoxin homologs in non-*Clostridium* species. *FEBS Letters*, 589(3):342–348, 2014.

[191] M. J. Mansfield and A. C. Doxey. Genomic insights into the evolution and ecology of botulinum neurotoxins. *Pathogens and Disease*, 76(4), 2018.

[192] M. J. Mansfield, T. G. Wentz, S. Zhang, E. J. Lee, M. Dong, S. K. Sharma, and A. C. Doxey. Bioinformatic discovery of a toxin family in *Chryseobacterium piperi* with sequence similarity to botulinum neurotoxins. *Scientific Reports*, 9(1):1–11, 2019.

[193] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, et al. CDD: NCBI's conserved domain database. *Nucleic Acids Research*, 43(D1):D222–D226, 2014.

[194] A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, et al. CDD: A conserved domain database for the functional annotation of proteins. *Nucleic Acids Research*, 39(Database):D225–D229, 2010.

[195] E. M. Marcotte. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[196] E. H. Margulies and E. Birney. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nature Reviews Genetics*, 9(4):303–313, 2008.

[197] D. S. Marks, T. A. Hopf, and C. Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, 2012.

[198] E. C. Martens, E. C. Lowe, H. Chiang, N. A. Pudlo, M. Wu, N. P. McNulty, D. W. Abbott, B. Henrissat, H. J. Gilbert, D. N. Bolam, and J. I. Gordon. Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biology*, 9(12):e1001221, 2011.

[199] L. M. Martins, R. F. Marquez, and T. Yano. Incidence of toxic *Aeromonas* isolated from food and human infection. *FEMS Immunology & Medical Microbiology*, 32(3):237–242, 2002.

[200] S. T. Mashiyama, M. M. Malabanan, E. Akiva, R. Bhosle, M. C. Branch, B. Hillerich, K. Jagessar, J. Kim, Y. Patskovsky, R. D. Seidel, et al. Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biology*, 12(4):e1001843, 2014.

[201] K. Mavromatis, N. N. Ivanova, I.-m. A. Chen, E. Szeto, V. M. Markowitz, and N. C. Kyrpides. The DOE-JGI standard operating procedure for the annotations of microbial genomes. *Standards in Genomic Sciences*, 1(1):63–67, 2009.

[202] W. W. McNab, M. J. Singleton, J. E. Moran, and B. K. Esser. Assessing the impact of animal waste lagoon seepage on the geochemistry of an underlying shallow aquifer. *Environmental Science & Technology*, 41(3):753–758, 2007.

[203] J. P. Meador, J. E. Stein, W. L. Reichert, and U. Varanasi. Bioaccumulation of polycyclic aromatic hydrocarbons by marine organisms. *Reviews of Environmental Contamination and Toxicology*, pages 79–165, 1995.

[204] T. Mechichi, E. Stackebrandt, N. Gad'on, and G. Fuchs. Phylogenetic and metabolic diversity of bacteria degrading aromatic compounds under denitrifying conditions, and description of *Thauera phenylacetica* sp. nov., *Thauera aminoaromatica* sp. nov., and *Azoarcus buckelii* sp. nov. *Archives of Microbiology*, 178(1):26–35, 2002.

[205] C. J. Meehan and R. G. Beiko. A phylogenomic view of ecological specialization in the Lachnospiraceae, a family of digestive tract-associated bacteria. *Genome Biology and Evolution*, 6(3):703–713, 2014.

[206] L. V. Mello, X. Chen, and D. J. Rigden. Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS Letters*, 584(11):2421–2426, 2010.

[207] K. Mendler, H. Chen, D. H. Parks, B. Lobb, L. A. Hug, and A. C. Doxey. Annotree: Visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Research*, 47(9):4442–4448, 2019.

[208] J. L. Metcalf, L. W. Parfrey, A. Gonzalez, C. L. Lauber, D. Knights, G. Ackermann, G. C. Humphrey, M. J. Gebert, W. V. Treuren, D. Berg-Lyons, et al. A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *eLife*, 2, 2013.

[209] J. L. Metcalf, Z. Z. Xu, S. Weiss, S. Lax, W. V. Treuren, E. R. Hyde, S. J. Song, A. Amir, P. Larsen, N. Sangwan, et al. Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science*, 351(6269):158–162, 2015.

[210] F. Meyer, R. Overbeek, and A. Rodriguez. FIGfams: Yet another set of protein families. *Nucleic Acids Research*, 37(20):6643–6654, 2009.

[211] G. W. Minshall, E. Hitchcock, and J. R. Barnes. Decomposition of rainbow trout (*Oncorhynchus mykiss*) carcasses in a forest stream ecosystem inhabited only by nonanadromous fish populations. *Canadian Journal of Fisheries and Aquatic Sciences*, 48(2):191–195, 1991.

[212] A. Mira. Microbial genome evolution: sources of variability. *Current Opinion in Microbiology*, 5(5):506–512, 2002.

[213] J. Mistry, E. Kloppmann, B. Rost, and M. Punta. An estimated 5% of new protein structures solved today represent a new Pfam family. *Acta Crystallographica Section D Biological Crystallography*, 69(11):2186–2193, 2013.

[214] A. L. Mitchell, T. K. Attwood, P. C. Babbitt, M. Blum, P. Bork, A. Bridge, S. D. Brown, H.-Y. Chang, S. El-Gebali, M. I. Fraser, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360, 2019.

[215] O. O. B. Mobolaji Felicia Adegboye. Phylogenetic characterization of culturable antibiotic producing *Streptomyces* from rhizospheric soils. *Molecular Biology*, 03(01), 2013.

[216] L. M. Molinari, D. d. O. Scoaris, R. B. Pedroso, N. d. L. R. Bittencourt, C. V. Nakamura, T. Nakamura, B. Abreu, and B. Dias. Bacterial microflora in the gastrointestinal tract of nile tilapia, *Oreochromis niloticus*, cultured in a semi-intensive system. *Acta Scientiarum Biological Sciences*, 25:267–271, 2003.

[217] N. A. Moran. Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108(5):583–586, 2002.

[218] N. A. Moran, H. J. McLaughlin, and R. Sorek. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science*, 323(5912):379–382, 2009.

[219] N. A. Moran and A. Mira. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biology*, 2(12):research0054.1, 2001.

[220] S. Moretti, B. Laurenczy, W. H. Gharib, B. Castella, A. Kuzniar, H. Schabauer, R. A. Studer, M. Valle, N. Salamin, H. Stockinger, and M. Robinson-Rechavi. Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Research*, 42(D1):D917–D921, 2013.

[221] R. Mudgal, S. Sandhya, N. Chandra, and N. Srinivasan. De-DUFing the DUFs: Deciphering distant evolutionary relationships of domains of unknown function using sensitive homology detection methods. *Biology Direct*, 10(1), 2015.

[222] K. D. Murrell and Y. Nawa. Animal waste: Risk of zoonotic parasite transmission. *Reviews on Environmental Health*, 13(4), 1998.

[223] A. Mushegian. The minimal genome concept. *Current Opinion in Genetics and Development*, 9(6):709–714, 1999.

[224] H. Nagase and J. F. Woessner. Matrix metalloproteinases. *Journal of Biological Chemistry*, 274(31):21491–21494, 1999.

[225] S. Nakjang, D. A. Ndeh, A. Wipat, D. N. Bolam, and R. P. Hirt. A novel extracellular metallopeptidase domain shared by animal host-associated mutualistic and pathogenic microbes. *PLoS ONE*, 7(1):e30287, 2012.

[226] V. Neduva and R. B. Russell. Linear motifs: Evolutionary interaction switches. *FEBS Letters*, 579(15):3342–3345, 2005.

[227] J.-Q. Ni, W. P. Robarge, C. Xiao, and A. J. Heber. Volatile organic compounds at swine facilities: A critical review. *Chemosphere*, 89(7):769–788, 2012.

[228] R. P. Novick, G. E. Christie, and J. R. Penadés. The phage-related chromosomal islands of gram-positive bacteria. *Nature Reviews Microbiology*, 8(8):541–551, 2010.

[229] N. R. Noyes, X. Yang, L. M. Linke, R. J. Magnuson, S. R. Cook, R. Zaheer, H. Yang, D. R. Woerner, I. Geornaras, J. A. McArt, et al. Characterization of the resistome in manure, soil and wastewater from dairy and beef production systems. *Scientific Reports*, 6(1), 2016.

[230] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, 2017.

[231] S. Ohno. *Evolution by gene duplication*. Springer-Verlag, 1970.

[232] M. R. Olm, A. Crits-Christoph, S. Diamond, A. Lavy, P. B. M. Carnevali, and J. F. Banfield. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems*, 5(1), 2020.

[233] A. C. Palmer and R. Kishony. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nature Reviews Genetics*, 14(4):243–248, 2013.

[234] H.-J. Park and E.-S. Kim. An inducible *Streptomyces* gene cluster involved in aromatic compound metabolism. *FEMS Microbiology Letters*, 226(1):151–157, 2003.

[235] D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. A. Chaumeil, and P. Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996, 2018.

[236] W. R. Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, 183(1):63–98, 1990.

[237] J. L. Pechal, T. L. Crippen, M. E. Benbow, A. M. Tarone, S. Dowd, and J. K. Tomberlin. The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. *International Journal of Legal Medicine*, 128(1):193–205, 2013.

[238] I. Pedruzzi, C. Rivoire, A. H. Auchincloss, E. Coudert, G. Keller, E. De Castro, D. Baratin, B. A. Cuche, L. Bougueleret, S. Poux, et al. HAMAP in 2015: Updates to the protein family classification and annotation system. *Nucleic Acids Research*, 43(D1):D1064–D1070, 2015.

[239] S. G. Peisajovich, J. E. Garbarino, P. Wei, and W. A. Lim. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science*, 328(5976):368–372, 2010.

[240] N. Perdigão, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, et al. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences of the United States of America*, 112(52):15898–15903, 2015.

[241] P. Petrenko and A. C. Doxey. mimicMe: a web server for prediction and analysis of host-like proteins in microbial pathogens. *Bioinformatics*, 31(4):590–592, 2014.

[242] P. Petrenko, B. Lobb, D. A. Kurtz, J. D. Neufeld, and A. C. Doxey. MetAnnotate: function-specific taxonomic profiling and comparison of metagenomes. *BMC Biology*, 13(1), 2015.

[243] D. Petrey, T. S. Chen, L. Deng, J. I. Garzon, H. Hwang, G. Lasso, H. Lee, A. Silkov, and B. Honig. Template-based prediction of protein function. *Current Opinion in Structural Biology*, 32:33–38, 2015.

[244] U. Pieper, B. M. Webb, G. Q. Dong, D. Schneidman-Duhovny, H. Fan, S. J. Kim, N. Khuri, Y. G. Spill, P. Weinkam, M. Hammel, et al. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, 42(D1):D336–D346, 2013.

[245] T. Prakash and T. D. Taylor. Functional assignment of metagenomic data: challenges and applications. *Briefings in Bioinformatics*, 13(6):711–727, 2012.

[246] K. Premke, P. Fischer, M. Hempel, and K.-O. Rothhaupt. Ecological studies on the decomposition rate of fish carcasses by benthic organisms in the littoral zone of Lake Constance, Germany. *Annales de Limnologie - International Journal of Limnology*, 46(3):157–168, 2010.

[247] M. N. Price, K. M. Wetmore, R. J. Waters, M. Callaghan, J. Ray, H. Liu, J. V. Kuehl, R. A. Melnyk, J. S. Lamson, Y. Suh, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, 557(7706):503–509, 2018.

[248] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[249] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[250] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. L. Chatelier, J. Yao, L. Wu, et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516):59–64, 2014.

[251] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2012.

[252] A. R. Quinlan. BEDTools: The Swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*, 47(1):11.12.1–11.12.34, 2014.

[253] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013.

[254] J. Raes, E. D. Harrington, A. H. Singh, and P. Bork. Protein function space: viewing the limits or limited by our view? *Current Opinion in Structural Biology*, 17(3):362–369, 2007.

[255] J. Raes, J. O. Korbel, M. J. Lercher, C. Von Mering, and P. Bork. Prediction of effective genome size in metagenomic samples. *Genome Biology*, 8(1):R10, 2007.

[256] M. Rahman, P. Colque-Navarro, I. Kühn, G. Huys, J. Swings, and R. Möllby. Identification and characterization of pathogenic *Aeromonas veronii* biovar sobria associated

with epizootic ulcerative syndrome in fish in Bangladesh. *Applied and Environmental Microbiology*, 68(2):650–655, 2002.

[257] C. Ran, C. Qin, M. Xie, J. Zhang, J. Li, Y. Xie, Y. Wang, S. Li, L. Liu, X. Fu, et al. *Aeromonas veronii* and aerolysin are important for the pathogenesis of motile aeromonad septicemia in cyprinid fish. *Environmental Microbiology*, 20(9):3442–3456, 2018.

[258] J. A. Ranea, D. W. Buchan, J. M. Thornton, and C. A. Orengo. Evolution of protein superfamilies and bacterial genome size. *Journal of Molecular Biology*, 336(4):871–887, 2004.

[259] M. E. Rateb, R. Ebel, and M. Jaspars. Natural product diversity of Actinobacteria in the Atacama Desert. *Antonie van Leeuwenhoek*, 111(8):1467–1477, 2018.

[260] N. D. Rawlings and A. J. Barrett. Evolutionary families of metallopeptidases. *Methods in Enzymology*, 248:183–228, 1995.

[261] M. Remmert, A. Biegert, A. Hauser, and J. Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175, 2011.

[262] M. Rho, H. Tang, and Y. Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, 38(20):e191–e191, 2010.

[263] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000.

[264] C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J. F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, 2013.

[265] D. B. Roche and T. Brüls. An assessment of the amount of untapped fold level novelty in under-sampled areas of the tree of life. *Scientific Reports*, 5(1), 2015.

[266] G. J. Rodriguez, R. Yao, O. Lichtarge, and T. G. Wensel. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proceedings of the National Academy of Sciences*, 107(17):7787–7792, 2010.

[267] R. Rosselló-Móra and R. Amann. Past and future species definitions for bacteria and archaea. *Systematic and Applied Microbiology*, 38(4):209–216, 2015.

[268] J. D. Rudolf, X. Yan, and B. Shen. Genome neighborhood network reveals insights into enediyne biosynthesis and facilitates prediction and prioritization for discovery. *Journal of Industrial Microbiology & Biotechnology*, 43(2-3):261–276, 2016.

[269] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, et al. The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, 5(3):e77, 2007.

[270] R. I. Sadreyev, D. Baker, and N. V. Grishin. Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Science*, 12(10):2262–2272, 2003.

[271] H. Salgado, G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proceedings of the National Academy of Sciences*, 97(12):6652–6657, 2000.

[272] R. Salwan and V. Sharma. The role of Actinobacteria in the production of industrial enzymes. *New and Future Developments in Microbial Biotechnology and Bioengineering*, pages 165–177, 2018.

[273] S. L. Salzberg. Next-generation genome annotation: We still struggle to get it right. *Genome Biology*, 20(1):92, 2019.

[274] A. Sánchez-Flores, E. Pérez-Rueda, and L. Segovia. Protein homology detection and fold inference through multiple alignment entropy profiles. *Proteins: Structure, Function, and Bioinformatics*, 70(1):248–256, 2007.

[275] Schmid K. J. and C. F. Aquadro. The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics*, 159:589–598, 2001.

[276] H. Schmitt, K. Stoob, G. Hamscher, E. Smit, and W. Seinen. Tetracyclines and tetracycline resistance in agricultural soils: microcosm and field studies. *Microbial Ecology*, 51(3):267–276, 2006.

[277] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12), 2009.

[278] E. Scholten, T. Lukow, G. Auling, R. M. Kroppenstedt, F. A. Rainey, and H. Diekmann. *Thauera mechernichensis* sp. nov., an aerobic denitrifier from a leachate treatment plant. *International Journal of Systematic and Evolutionary Microbiology*, 49(3):1045–1051, 1999.

[279] I. Schröder, S. Rech, T. Krafft, and J. M. Macy. Purification and characterization of the selenate reductase from *Thauera selenatis*. *Journal of Biological Chemistry*, 272(38):23765–23768, 1997.

[280] T. Seemann. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[281] B. Shen, X. Yan, T. Huang, H. Ge, D. Yang, Q. Teng, J. D. Rudolf, J. R. Lohman, et al. Enediynes: exploration of microbial genomics to discover new anticancer drug leads. *Bioorganic & Medicinal Chemistry Letters*, 25(1):9–15, 2015.

[282] A. Sheydina, R. Y. Eberhardt, D. J. Rigden, Y. Chang, Z. Li, C. C. Zmasek, H. L. Axelrod, and A. Godzik. Structural genomics analysis of uncharacterized protein families overrepresented in human gut bacteria identifies a novel glycoside hydrolase. *BMC Bioinformatics*, 15(1):112, 2014.

[283] N. Siew. The ORFanage: an ORFan database. *Nucleic Acids Research*, 32(90001):281D—-283, 2004.

[284] N. Siew and D. Fischer. Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins: Structure, Function and Genetics*, 53(2):241–251, 2003.

[285] C. J. Sigrist, E. De Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios. New and continuing developments at PROSITE. *Nucleic Acids Research*, 41(D1):D344–D347, 2012.

[286] I. Sillitoe, N. Dawson, T. E. Lewis, S. Das, J. G. Lees, P. Ashford, A. Tolulope, H. M. Scholes, I. Senatorov, A. Bujan, et al. Cath: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Research*, 47(D1):D280–D284, 2019.

[287] I. Sillitoe, T. E. Lewis, A. Cuff, S. Das, P. Ashford, N. L. Dawson, N. Furnham, R. A. Laskowski, D. Lee, J. G. Lees, et al. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(D1):D376–D381, 2014.

[288] A. H. Singh, T. Doerks, I. Letunic, J. Raes, and P. Bork. Discovering functional novelty in metagenomes: Examples from light-mediated processes. *Journal of Bacteriology*, 191(1):32–41, 2008.

[289] G. Singh, A. Verma, and V. Kumar. Catalytic properties, functional attributes and industrial applications of β-glucosidases. *3 Biotech*, 6(1):3, 2016.

[290] J. Skolnick and J. S. Fetrow. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends in Biotechnology*, 18(1):34–39, 2000.

[291] T. F. Smith, M. S. Waterman, et al. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[292] J. Söding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960, 2005.

[293] B. Song, N. J. Palleroni, L. J. Kerkhof, and M. M. Häggblom. Characterization of halobenzoate-degrading, denitrifying *Azoarcus* and *Thauera* isolates and description of *Thauera chlorobenzoica* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 51(2):589–602, 2001.

[294] E. Stackebrandt. Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today*, 33:152–155, 2006.

[295] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[296] C. E. Stebbins and J. E. Galán. Structural mimicry in bacterial virulence. *Nature*, 412(6848):701–705, 2001.

[297] R. A. Studer, B. H. Dessailly, and C. A. Orengo. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochemical Journal*, 449(3):581–594, 2013.

[298] X.-L. Su, Q. Tian, J. Zhang, X.-Z. Yuan, X.-S. Shi, R.-B. Guo, and Y.-L. Qiu. *Acetobacteroides hydrogenigenes* gen. nov., sp. nov., an anaerobic hydrogen-producing bacterium in the family Rikenellaceae isolated from a reed swamp. *International Journal of Systematic and Evolutionary Microbiology*, 64(Pt_9):2986–2991, 2014.

[299] H. Sugita, K. Tanaka, M. Yoshinami, and Y. Deguchi. Distribution of *Aeromonas* species in the intestinal tracts of river fish. *Applied and Environmental Microbiology*, 61(11):4128–4130, 1995.

[300] S. J. H. Sui, A. Fedynak, W. W. Hsiao, M. G. Langille, and F. S. Brinkman. The association of virulence factors with genomic islands. *PloS ONE*, 4(12), 2009.

[301] Y.-M. Sung, A. D. Wilkins, G. J. Rodriguez, T. G. Wensel, and O. Lichtarge. Intramolecular allosteric communication in dopamine d2 receptor revealed by evolutionary amino acid covariation. *Proceedings of the National Academy of Sciences*, 113(13):3539–3544, 2016.

[302] T. V. Sydenham, M. Arpi, K. Klein, and U. S. Justesen. Four cases of bacteremia caused by *Oscillibacter ruminantium*, a newly described species. *Journal of Clinical Microbiology*, 52(4):1304–1307, 2014.

[303] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2014.

[304] J. Takahashi, Y. Ichikawa, H. Sagae, I. Komura, H. Kanou, and K. Yamada. Isolation and identification of n-butane-assimilating bacterium. *Agricultural and Biological Chemistry*, 44(8):1835–1840, 1980.

[305] S. Takahashi, J. Tomita, K. Nishioka, T. Hisada, and M. Nishijima. Development of a prokaryotic universal primer for simultaneous analysis of bacteria and archaea using next-generation sequencing. *PLoS ONE*, 9(8):e105592, 2014.

[306] R. L. Tatusov. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, 2000.

[307] D. Tautz and T. Domazet-Lošo. The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702, 2011.

[308] N. Terrapon, V. Lombard, H. J. Gilbert, and B. Henrissat. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics*, 31(5):647–655, 2014.

[309] J. M. Thornton, C. A. Orengo, A. E. Todd, and F. M. G. Pearl. Protein folds, functions and evolution. *Journal of Molecular Biology*, 293(2):333–342, 1999.

[310] B. Tian, N. H. Fadhil, J. E. Powell, W. K. Kwong, and N. A. Moran. Long-term exposure to antibiotics has caused accumulation of resistance determinants in the gut microbiota of honeybees. *MBio*, 3(6):e00377–12, 2012.

[311] W. Tian and J. Skolnick. How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, 333(4):863–882, 2003.

[312] V. H. Tierrafría, C. Mejía-Almonte, J. M. Camacho-Zaragoza, H. Salgado, K. Alquicira, C. Ishida, S. Gama-Castro, and J. Collado-Vides. MCO: Towards an ontology and unified vocabulary for a framework-based annotation of microbial growth conditions. *Bioinformatics*, 35(5):856–864, 2019.

[313] L. Tomás-Gallardo, H. Gómez-Álvarez, E. Santero, and B. Floriano. Combination of degradation pathways for naphthalene utilization in *Rhodococcus* sp. strain tfb. *Microbial Biotechnology*, 7(2):100–113, 2013.

[314] P. Tompa, N. E. Davey, T. J. Gibson, and M. M. Babu. A million peptide motifs for the molecular biologist. *Molecular Cell*, 55(2):161–169, 2014.

[315] E. Toprak, A. Veres, J.-B. Michel, R. Chait, D. L. Hartl, and R. Kishony. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics*, 44(1):101–105, 2011.

[316] M. Touchon, C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics*, 5(1):e1000344, 2009.

[317] S. G. Tringe. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, 2005.

[318] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[319] L. A. Turner and C. Bucking. The role of intestinal bacteria in the ammonia detoxification ability of teleost fish. *The Journal of Experimental Biology*, 222(24):jeb209882, 2019.

[320] L. UfartÃ, G. Potocki-Veronese, and Ã. Laville. Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology. *Frontiers in Microbiology*, 6, 2015.

[321] O. Uhlik, J. Wald, M. Strejcek, L. Musilova, J. Ridl, M. Hroudova, C. Vlcek, E. Cardenas, M. Mackova, and T. Macek. Identification of bacteria utilizing biphenyl, benzoate, and naphthalene in long-term contaminated soil. *PLoS ONE*, 7(7):e40653, 2012.

[322] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen. A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14(5):535–542, 2006.

[323] R. C. van Ham, J. Kamerbeek, C. Palacios, C. Rausell, F. Abascal, U. Bastolla, J. M. Fernández, L. Jiménez, M. Postigo, F. J. Silva, et al. Reductive genome evolution in *Buchnera aphidicola*. *Proceedings of the National Academy of Sciences of the United States of America*, 100(2):581–586, 2003.

[324] M. A. H. J. van Kessel, D. R. Speth, M. Albertsen, P. H. Nielsen, H. J. M. O. den Camp, B. Kartal, M. S. M. Jetten, and S. Lücker. Complete nitrification by a single microorganism. *Nature*, 528(7583):555–559, 2015.

[325] C. E. VanOrsdel, J. P. Kelly, B. N. Burke, C. D. Lein, C. E. Oufiero, J. F. Sanchez, L. E. Wimmers, D. J. Hearn, F. J. Abuikhdair, K. R. Barnhart, et al. Identifying new small proteins in *Escherichia coli*. *Proteomics*, 18(10):1700064, 2018.

[326] T. Vazin, K. G. Becker, J. Chen, C. E. Spivak, C. R. Lupica, Y. Zhang, L. Worden, and W. J. Freed. A novel combination of factors, termed SPIE, which promotes dopaminergic neuron differentiation from human embryonic stem cells. *PLoS ONE*, 4(8):e6606, 2009.

[327] J. C. Venter. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74, 2004.

[328] G. Vey and G. Moreno-Hagelsieb. Beyond the bounds of orthology: functional inference from metagenomic context. *Molecular BioSystems*, 6(7):1247, 2010.

[329] A. Via, B. Uyar, C. Brun, and A. Zanzoni. How pathogens use linear motifs to perturb host cell networks. *Trends in Biochemical Sciences*, 40(1):36–48, 2015.

[330] J. J. Vitti, S. R. Grossman, and P. C. Sabeti. Detecting natural selection in genomic data. *Annual Review of Genetics*, 47(1):97–120, 2013.

[331] J. Vlasblom, K. Zuberi, H. Rodriguez, R. Arnold, A. Gagarinova, V. Deineko, A. Kumar, E. Leung, K. Rizzolo, B. Samanfar, et al. Novel function discovery with

GeneMANIA: a new integrated resource for gene function prediction in *Escherichia coli*. *Bioinformatics*, 31(3):306–310, 2014.

[332] F. H. Wallrapp, J.-J. Pan, G. Ramamoorthy, D. E. Almonacid, B. S. Hillerich, R. Seidel, Y. Patskovsky, P. C. Babbitt, S. C. Almo, M. P. Jacobson, and C. D. Poulter. Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. *Proceedings of the National Academy of Sciences*, 110(13):E1196–E1202, 2013.

[333] J. T. Wang, M. J. Zaki, H. T. Toivonen, and D. Shasha. Introduction to data mining in bioinformatics. In *Data mining in bioinformatics*, pages 3–8. Springer, 2005.

[334] T. Waschkowitz, S. Rockstroh, and R. Daniel. Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. *Applied and Environmental Microbiology*, 75(8):2506–2516, 2009.

[335] A. R. Wattam, J. J. Davis, R. Assaf, S. Boisvert, T. Brettin, C. Bun, N. Conrad, E. M. Dietrich, T. Disz, J. L. Gabbard, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Research*, 45(D1):D535–D542, 2016.

[336] G. A. Wilson, N. Bertrand, Y. Patel, J. B. Hughes, E. J. Feil, and D. Field. Orphans as taxonomically restricted and ecologically important genes. *Microbiology*, 151(8):2499–2501, 2005.

[337] S. Wischgoll, D. Heintz, F. Peters, A. Erxleben, E. Sarnighausen, R. Reski, A. Van Dorsselaer, and M. Boll. Gene clusters involved in anaerobic benzoate degradation of *Geobacter metallireducens*. *Molecular Microbiology*, 58(5):1238–1252, 2005.

[338] J. C. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLoS Computational Biology*, 6(2):e1000667, 2010.

[339] S. Wu and Y. Zhang. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35(10):3375–3382, 2007.

[340] S. K. Wyman, A. Avila-Herrera, S. Nayfach, and K. S. Pollard. A most wanted list of conserved microbial protein families with no known domains. *PLoS ONE*, 13(10):e0205749, 2018.

[341] W. Xiong, Y. Sun, X. Ding, M. Wang, and Z. Zeng. Selective pressure of antibiotics on args and bacterial communities in manure-polluted freshwater-sediment microcosms. *Frontiers in Microbiology*, 6, 2015.

[342] Q. Xu, M. Shoji, S. Shibata, M. Naito, K. Sato, M.-A. Elsliger, J. Grant, H. Axelrod, H.-J. Chiu, C. Farr, et al. A distinct type of pilus from the human microbiome. *Cell*, 165(3):690–703, 2016.

[343] X. Yan, H. Ge, T. Huang, D. Yang, Q. Teng, I. Crnovčić, X. Li, J. D. Rudolf, J. R. Lohman, Y. Gansemans, et al. Strain prioritization and genome mining for enediyne natural products. *MBio*, 7(6):e02104–16, 2016.

[344] I. Yanai, J. C. Mellor, and C. DeLisi. Identifying functional links between genes using conserved chromosomal proximity. *Trends in Genetics*, 18(4):176–179, 2002.

[345] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang. The I-TASSER suite: protein structure and function prediction. *Nature Methods*, 12(1):7–8, 2014.

[346] J. Yang and Y. Zhang. Protein structure and function prediction using I-TASSER. *Current Protocols in Bioinformatics*, 52(1):5–8, 2015.

[347] L. Yang, J.-B. Muhadesi, M.-M. Wang, B.-J. Wang, S.-J. Liu, and C.-Y. Jiang. *Thauera hydrothermalis* sp. nov., a thermophilic bacterium isolated from hot spring. *International Journal of Systematic and Evolutionary Microbiology*, 68(10):3163–3168, 2018.

[348] Y. Yin and D. Fischer. On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. *BMC Evolutionary Biology*, 6(1):63, 2006.

[349] Y. Yin, X. Mao, J. Yang, X. Chen, F. Mao, and Y. Xu. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, 40(W1):W445–W451, 2012.

[350] I. Yomtovian, N. Teerakulkittipong, B. Lee, J. Moult, and R. Unger. Composition bias and the origin of ORFan genes. *Bioinformatics*, 26(8):996–999, 2010.

[351] S. Yooseph, G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, et al. The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biology*, 5(3):e16, 2007.

[352] R. You, Z. Zhang, Y. Xiong, F. Sun, H. Mamitsuka, and S. Zhu. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, 34(14):2465–2473, 2018.

[353] Z. Zádori, J. Szelei, M.-C. Lacoste, Y. Li, S. Gariépy, P. Raymond, M. Allaire, I. R. Nabi, and P. Tijssen. A viral phospholipase A2 is required for parvovirus infectivity. *Developmental Cell*, 1(2):291–302, 2001.

[354] J. Załuga, P. Stragier, S. Baeyen, A. Haegeman, J. V. Vaerenbergh, M. Maes, and P. D. Vos. Comparative genome analysis of pathogenic and non-pathogenic *Clavibacter* strains reveals adaptations to their lifestyle. *BMC Genomics*, 15(1):392, 2014.

[355] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–560, 2012.

[356] S. Zhang, F. Lebreton, M. J. Mansfield, S.-I. Miyashita, J. Zhang, J. A. Schwartzman, L. Tao, G. Masuyer, M. Martinez-Carranza, P. Stenmark, et al. Identification of a botulinum neurotoxin-like toxin in a commensal strain of *Enterococcus faecium*. *Cell Host & Microbe*, 23(2):169–176.e6, 2018.

[357] W. Zhang, W. Ding, Y.-X. Li, C. Tam, S. Bougouffa, R. Wang, B. Pei, H. Chiang, P. Leung, Y. Lu, et al. Marine biofilms constitute a bank of hidden microbial diversity and functional potential. *Nature Communications*, 10(1), 2019.

[358] Y. Zhang, C. Zhang, D. B. Parker, D. D. Snow, Z. Zhou, and X. Li. Occurrence of antimicrobials and antimicrobial resistance genes in beef cattle storage ponds and swine treatment lagoons. *Science of The Total Environment*, 463-464:631–638, 2013.

[359] Y.-H. P. Zhang, M. E. Himmel, and J. R. Mielenz. Outlook for cellulase improvement: screening and selection strategies. *Biotechnology Advances*, 24(5):452–481, 2006.

[360] S. Zhao, R. Kumar, A. Sakai, M. W. Vetting, B. M. Wood, S. Brown, J. B. Bonanno, B. S. Hillerich, R. D. Seidel, P. C. Babbitt, et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature*, 502(7473):698–702, 2013.

[361] S. Zhao, A. Sakai, X. Zhang, M. W. Vetting, R. Kumar, B. Hillerich, B. S. Francisco, J. Solbiati, A. Steves, S. Brown, et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife*, 3, 2014.

[362] N. Zhou, Y. Jiang, T. R. Bergquist, A. J. Lee, B. Z. Kacsoh, A. W. Crocker, K. A. Lewis, G. Georghiou, H. N. Nguyen, M. N. Hamid, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1):1–23, 2019.

[363] Y.-G. Zhu, T. A. Johnson, J.-Q. Su, M. Qiao, G.-X. Guo, R. D. Stedtfeld, S. A. Hashsham, and J. M. Tiedje. Diverse and abundant antibiotic resistance genes in chinese swine farms. *Proceedings of the National Academy of Sciences*, 110(9):3435–3440, 2013.

[364] C. M. Zmasek and A. Godzik. This déjà vu feeling - analysis of multidomain protein evolution in eukaryotic genomes. *PLoS Computational Biology*, 8(11):e1002701, 2012.

[365] V. V. Zverlov, N. Schantz, and W. H. Schwarz. A major new component in the cellulosome of *Clostridium thermocellum* is a processive endo-β-1, 4-glucanase producing cellotetraose. *FEMS Microbiology Letters*, 249(2):353–358, 2005.

APPENDICES

Supplementary Material: Chapter 2

Supplementary Figure 1

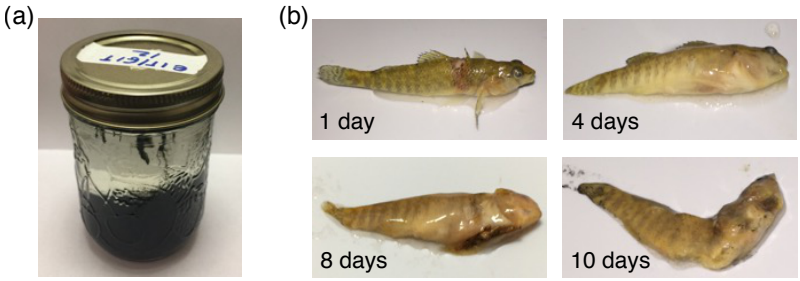

Figure 1: Decomposition setup and images of fish decay. (a) Example of the mason jars used for decay of female rainbow darter (*Etheostoma caeruleum*) in a microcosm of the Grand River. (b) Representative images of decay of female rainbow darter for enrichment of the necrobiome at 1, 4, 8, and 10 days. *Pictures taken by Dr. Paul Craig.*

Supplementary Figure 2

Figure 2: Bubble-plot depicting the relative frequency (as a percentage) of ASVs in the fish necrobiome at four time points. Light gray boxes indicate shared family level taxonomic affiliation. Bubbles are displayed only if the ASV taxonomic affiliation was $\geq$2%. For other ASVs, see Data Set S1A in the supplemental material of the original paper.[178] Bubbles are colored by decomposition time (days).

Supplementary Figure 3

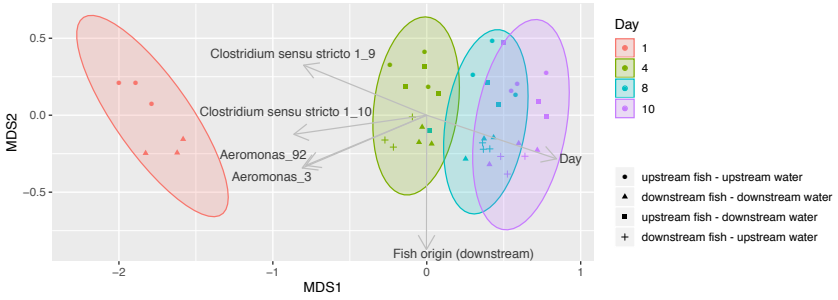

Figure 3: A nonmetric multidimensional scaling (NMDS) ordination of necrobiomes based on microbial community composition, using Bray-Curtis distances generated from ASV frequency profiles. Stress is 0.098. Together, 99% of the variance is represented based on the R^2 value between distance in ordination space and distance in the original matrix. Vectors with R^2 values greater than 0.7 were shown on the plot. Ellipses are colored by decomposition time (days).

Supplementary Table 1

Table 1: High-confidence cellulase annotations for *Streptomyces* sp. NWU339 and *Streptomyces viridosporus* NWU49. The full list of cellulase predictions is File ??a.

	RAST	**PROKKA**	**KEGG**	**dbCAN**
NWU339				
Contig_146_ 25180_27099	Endoglucanase (EC 3.2.1.4) - FIG00516442	Endoglucanase E1 [EC 3.2.1.4]	endoglucanase [EC:3.2.1.4] - K01179	GH5_51
Contig_190_ 182065_183603	Endoglucanase E1 precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase E1) (Cellulase E1) (Endocellulase E1) - FIG00817790	Endoglucanase E1 [EC 3.2.1.4]	endoglucanase [EC:3.2.1.4] - K01179	GH5_1; CBM2
Contig_7_ 97558_95177	Beta-glucosidase (EC 3.2.1.21) - FIG00001469	Beta-glucosidase BoGH3B [EC 3.2.1.21]	beta-glucosidase [EC:3.2.1.21] - K05349	GH3
Contig_7_ 103396_101957	Beta-glucosidase (EC 3.2.1.21) - FIG00001469	Bifunctional beta-D-glucosidase/beta-D-fucosidase [EC 3.2.1.21]	beta-glucosidase [EC:3.2.1.21] - K05350	GH1
Contig_187_ 63312_61873	Beta-glucosidase (EC 3.2.1.21) - FIG00001469	Bifunctional beta-D-glucosidase/beta-D-fucosidase [EC 3.2.1.21]	beta-glucosidase [EC:3.2.1.21] - K05350	GH1
NWU49				
Contig_177_ 236137_238065	Endoglucanase (EC 3.2.1.4) - FIG00516442	Endoglucanase E1 [EC 3.2.1.4]	endoglucanase [EC:3.2.1.4] - K01179	GH5_51
Contig_162_ 138192_136807	Endoglucanase celA precursor (EC 3.2.1.4) (Endo-14-beta-glucanase) (Cellulase) - FIG01126837	Endoglucanase CelA [EC 3.2.1.4]	endoglucanase [EC:3.2.1.4] - K01179	GH5_2; CBM2
Contig_126 _205924_204485	Beta-glucosidase (EC 3.2.1.21) - FIG00001469	Bifunctional beta-D-glucosidase/beta-D-fucosidase [EC 3.2.1.21]	beta-glucosidase [EC:3.2.1.21] - K05350	GH1

continued on the next page

Table A.2 - (*continued from the previous page*)

	RAST	**PROKKA**	**KEGG**	**dbCAN**
NWU49				
Contig_73_88157_89596	Beta-glucosidase (EC 3.2.1.21) - FIG00001469	Bifunctional beta-D-glucosidase/beta-D-fucosidase [EC 3.2.1.21]	beta-glucosidase [EC:3.2.1.21] - K05350	GH1
Contig_144_274763_272382	Beta-glucosidase (EC 3.2.1.21) - FIG00001469	Beta-glucosidase BoGH3B [EC 3.2.1.21]	beta-glucosidase [EC:3.2.1.21] - K05349	GH3
Contig_90_29739_27268	Beta-glucosidase (EC 3.2.1.21) - FIG00001469	Thermostable beta-glucosidase B [EC 3.2.1.21]	beta-glucosidase [EC:3.2.1.21] - K05349	GH3

Supplementary Material: Chapter 3

Supplementary Figure 4

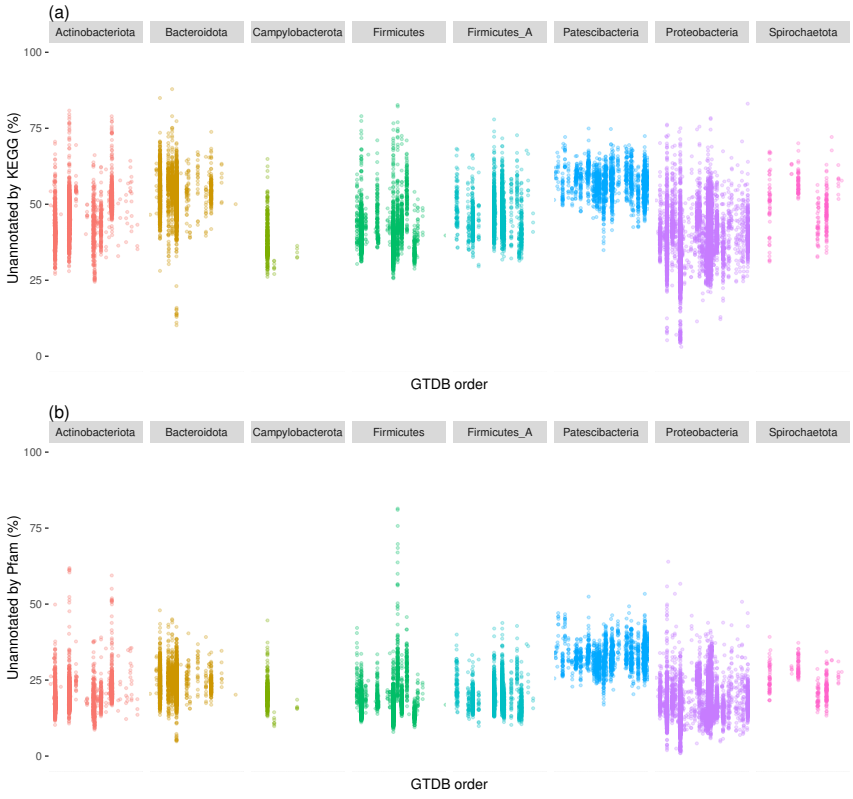

Figure 4: Taxonomic separation of genome annotation coverage by order using

the taxonomic nomenclature from the Genome Taxonomy Database (GTDB). Only the most common GTDB phyla are shown. (a) KEGG genome annotation. (b) Pfam genome annotation.

Supplementary Figure 5

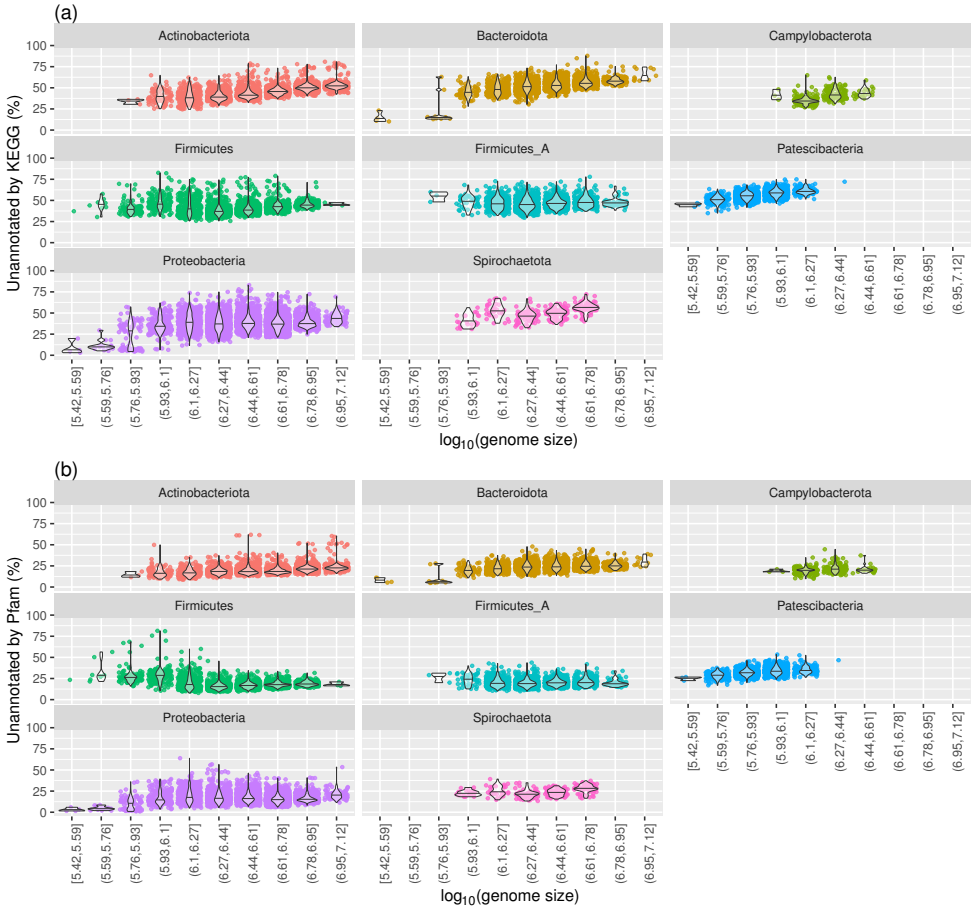

Figure 5: Effect of genome size (bp) on genome annotation coverage. $\log_{10}$(genome size) is binned into 10 distinct bins to better display the trend. The most common GTDB phyla are displayed separately. (a) KEGG genome annotation. (b) Pfam genome annotation.

Supplementary Material: Chapter 4

Supplementary Figure 6

Figure 6: Lineage-specificity distributions for Pfam families. (a) Lineage-specificity scores

using the F1 statistic combining sensitivity and precision. (b) Number of lineage-specific domains identified in various taxonomic groups with sensitivity and precision scores $\geq 95\%$. Bacteria, archaea, and viruses have no lineage-specific domains above the 95% threshold for certain taxonomic levels because of fewer proteomes present in Pfam leading to a weaker signal (in the case of archaea and viruses) and because of unclassified taxonomic levels within the NCBI taxonomy system. An overrepresentation of eukaryotic lineage-specific domains occur at the phylum level, due primarily to the Streptophyta and Chordata clades. *Chlamydia* and *Borreliella* account for just under half the bacterial domains with genus-level lineage specificity, while Cyanobacteria and Deinococcus-Thermus account for around two-thirds of the phylum-level lineage specificity of the bacterial domains. The class-level lineage specificity of the bacterial domains is primarily due to Actinobacteria. Archaeal domain families have high class-level lineage specificity, driven primarily by the Thermococci. The most common family-level lineage specific viral domains are found in Baculoviridae and Poxviridae, while the genus-level domains have a more diverse origin.

Supplementary Table 2

Table 2: Top five lineage specific domains in Eukaryota, Archaea, Bacteria, and Viruses. Families were ranked by the F1 score of their best taxonomic level and the number of proteomes in which they are present, excluding domains present in less than 20 species.

	Proteomes with domain	**Best taxonomic lineage**	**Best sensitivity**	**Best precision**	**F1 score**
Eukaryota					
MH1	315	Metazoa	100	100	100
PDDEXK_6	95	Viridiplantae	100	100	100
DM4_12	91	Arthropoda	100	100	100
APO_RNA-bind, BES1_N, BRX_N, COBRA, DUF1191, DUF1639, DUF3444, DUF3475, DUF4370, DUF668, LOB, NAM,PH_2, PHD_Oberon, tify,VQ, WCOR413, XS, Zein-binding,zf-UDP	81	Streptophyta	100	100	100
Moulting_cycle	65	Nematoda	100	100	100
Bacteria					
HP0268	93	Epsilonproteobacteria	100	100	100
LidA_Long_CC	28	Legionellaceae	100	100	100
DUF3208, DUF3809, PilN_bio_d, Taq-exonuc	26	Deinococcus-Thermus	100	100	100
RbpA	1154	Actinobacteria	99.10	99.48	99.70
HP1451_C	92	Epsilonproteobacteria	100	98.92	99.46
Archaea					
BAT	100	Halobacteria	96	100	97.96
S-layer	26	Methanosarcinales	100	92.86	96.30
DUF2208	42	Thermoprotei	90.48	100	95
FpoO	25	Methanosarcinales	100	89.29	94.34
DUF2192	42	Thermoprotei	88.10	97.37	92.5
Viruses					
Baculo_helicase, Baculo_p33, Baculo_VP1054, Baculo_VP39, LEF-9	61	Baculoviridae	100	100	100

continued on the next page

Table C.5 - (*continued from the previous page*)

	Proteomes with domain	**Best taxonomic lineage**	**Best sensitivity**	**Best precision**	**F1 score**
DNA_pol_B_3, Pox_A21, Pox_E10, Pox_G5, Pox_L3_FP4, Pox_LP_H2, Pox_Rap94, Pox_Rif, Pox_VERT_large, VirDNA-topo-I_N	39	Poxviridae	100	100	100
Orbi_VP4, Orbi_VP5	29	Orbivirus	100	100	100
Late_protein_L1	222	Papillomaviridae	99.55	100	99.77
Baculo_VP91_N, LEF-4, LEF-8	62	Baculoviridae	98.39	100	99.19

Supplementary Material: Chapter 5

Supplementary Table 3

Table 3: Enriched GO terms among ORFans from each metagenome. Two backgrounds were used: 1) a random subset of Pfam-annotated CDSs from the same metagenome; 2) the PDB70 database. Some GO terms that were enriched in the random Pfam-annotated subset were not significantly enriched relative to the background database (PDB).

GO term	Fold change	ORFan hits	PFAM hits	P_{adj} against Pfam	P_{adj} against PDB70
Soil					
GDP-dissociation_inhibitor_activity	18.11	66	5	7.47×10^{-55}	1.58×10^{-90}
dibenzothiophene_catabolic_process	12.01	35	4	1.69×10^{-22}	3.66×10^{-55}
mitochondrial_fission	12.81	28	3	2.16×10^{-18}	7.14×10^{-40}
serine-type_endopeptidase_inhibitor_activity	15.78	23	2	9.43×10^{-17}	1.00
calcium_ion_binding	2.63	111	58	1.44×10^{-15}	1.00
sequence-specific_DNA_binding	2.08	162	107	6.39×10^{-14}	2.14×10^{-49}
viral_release_from_host_cell	19.21	14	1	1.20×10^{-10}	5.14×10^{-2}
pectate_lyase_activity	4.49	36	11	6.51×10^{-10}	1.35×10^{-38}
cellulase_activity	6.52	19	4	6.04×10^{-7}	3.62×10^{-4}
polysaccharide_catabolic_process	2.56	54	29	2.63×10^{-6}	1.47×10^{-8}
1-alkyl-2-acetylglycerophosphocholine_esterase_activity	8.23	12	2	8.76×10^{-5}	1.47×10^{-8}
peptidylglycine_monooxygenase_activity	3.06	29	13	4.47×10^{-4}	1.22×10^{-37}
peptide_metabolic_process	2.94	30	14	6.40×10^{-4}	8.51×10^{-27}
oligogalacturonide_lyase_activity	3.20	21	9	9.62×10^{-3}	2.25×10^{-39}
oligosaccharyl_transferase_activity	2.18	35	22	4.75×10^{-2}	5.15×10^{-48}
transition_metal_ion_binding	2.16	33	21	9.92×10^{-2}	1.00

continued on the next page

Table D.1 - (*continued from the previous page*)

GO term	**Fold change**	**ORFan hits**	**PFAM hits**	P_{adj} **against Pfam**	P_{adj} **against PDB70**
Soil					
ribonuclease_activity	1.93	38	27	2.69×10^{-1}	1.35×10^{-4}
metallopeptidase_activity	1.52	84	76	3.34×10^{-1}	2.21×10^{-10}
lysozyme_activity	3.77	11	4	3.99×10^{-1}	1.00
cysteine-type_peptidase_activity	2.25	23	14	6.67×10^{-1}	1.00
Marine					
calcium_ion_binding	6.97	133	20	6.51×10^{-62}	1.00
polysaccharide_catabolic_process	16.26	62	4	1.20×10^{-48}	8.05×10^{-6}
lysozyme_activity	40.90	39	1	4.90×10^{-45}	1.00
L-ascorbic_acid_binding	5.83	89	16	4.69×10^{-35}	1.01×10^{-74}
protein_kinase_activity	9.79	56	6	2.04×10^{-32}	1.00
ADP-heptose-lipopolysaccharide_ heptosyltransferase_activity	18.35	35	2	1.63×10^{-28}	4.01×10^{-88}
dephosphorylation	4.45	89	21	1.43×10^{-26}	1.00
phosphatidylinositol_alpha-mannosyltransferase_activity	27.27	26	1	4.88×10^{-25}	5.60×10^{-42}
endonuclease_activity	2.74	157	60	1.55×10^{-24}	1.25×10^{-11}
CDP-glycerol_glycerophosphotransferase_ activity	25.17	24	1	3.51×10^{-22}	1.24×10^{-25}
metallopeptidase_activity	3.44	95	29	1.64×10^{-20}	3.53×10^{-6}
sequence-specific_DNA_binding	2.27	167	77	6.21×10^{-18}	6.35×10^{-33}
rRNA_binding	6.03	46	8	6.62×10^{-18}	1.00
viral_release_from_host_cell	20.98	20	1	1.10×10^{-16}	4.96×10^{-2}
RNA_ligase_(ATP)_activity	20.98	20	1	1.10×10^{-16}	1.45×10^{-26}
phosphatase_activity	4.82	46	10	2.98×10^{-14}	1.00
pectate_lyase_activity	18.88	18	1	4.63×10^{-14}	2.25×10^{-11}
proline_3-hydroxylase_activity	9.09	26	3	1.98×10^{-13}	1.63×10^{-46}
serine-type_endopeptidase_inhibitor_activity	16.78	16	1	1.57×10^{-11}	1.00
spermatogenesis	16.78	16	1	1.57×10^{-11}	1.00
N-methyltransferase_activity	16.78	16	1	1.57×10^{-11}	5.31×10^{-9}
mitochondrial_fission	16.78	16	1	1.57×10^{-11}	2.65×10^{-16}
sulfate_assimilation,_phosphoadenylyl_ sulfate_reduction_by_phosphoadenylyl-sulfate_ reductase_(thioredoxin)	15.73	15	1	2.65×10^{-10}	6.61×10^{-11}
phospholipase_activity	15.73	15	1	2.65×10^{-10}	1.58×10^{-11}
phosphoadenylyl-sulfate_reductase_(thioredoxin)_activity	15.73	15	1	2.65×10^{-10}	1.72×10^{-14}
peptidoglycan_beta-N-acetylmuramidase_activity	14.68	14	1	4.18×10^{-9}	6.93×10^{-19}
phosphorelay_signal_transduction_system	3.55	44	13	4.46×10^{-9}	1.00
cysteine-type_peptidase_activity	13.63	13	1	6.17×10^{-8}	1.00
O-methyltransferase_activity	13.63	13	1	6.17×10^{-8}	8.54×10^{-1}

continued on the next page

Table D.1 - (*continued from the previous page*)

GO term	**Fold change**	**ORFan hits**	**PFAM hits**	P_{adj} **against Pfam**	P_{adj} **against PDB70**
Marine					
cell_redox_homeostasis	2.82	51	19	3.16×10^{-7}	1.00
cholesterol_binding	12.59	12	1	8.46×10^{-7}	1.00
blood_coagulation	5.51	21	4	1.44×10^{-6}	1.00
phosphoric_diester_hydrolase_activity	5.51	21	4	1.44×10^{-6}	1.00
protein_complex_assembly	5.51	21	4	1.44×10^{-6}	6.63×10^{-1}
peptide_metabolic_process	11.54	11	1	1.07×10^{-5}	6.74×10^{-4}
lactone_biosynthetic_process	11.54	11	1	1.07×10^{-5}	1.32×10^{-15}
viral_life_cycle	10.49	10	1	1.24×10^{-4}	1.00
triglyceride_lipase_activity	10.49	10	1	1.24×10^{-4}	1.00
1-alkyl-2-acetylglycerophosphocholine_esterase_activity	4.72	18	4	2.14×10^{-4}	2.20×10^{-12}
apoptotic_process	3.15	27	9	6.72×10^{-4}	1.00
glycerol_ether_metabolic_process	3.67	21	6	1.19×10^{-3}	1.00
scyllo-inosamine-4-phosphate_amidinotransferase_activity	6.29	12	2	1.43×10^{-3}	4.97×10^{-21}
endopeptidase_activity	5.77	11	2	9.16×10^{-3}	1.00
intracellular_protein_transport	3.02	23	8	9.16×10^{-3}	1.00
cytokinesis_by_binary_fission	5.77	11	2	9.16×10^{-3}	6.23×10^{-3}
response_to_mercury_ion	4.20	12	3	7.79×10^{-2}	8.54×10^{-1}
glycine_amidinotransferase_activity	3.85	11	3	3.33×10^{-1}	5.84×10^{-11}
Human Gut					
sequence-specific_DNA_binding	2.21	149	260	3.25×10^{-15}	1.63×10^{-94}
polysaccharide_catabolic_process	4.61	49	41	1.35×10^{-14}	1.16×10^{-21}
calcium_ion_binding	4.27	31	28	8.64×10^{-8}	1.00
regulation_of_sporulation_resulting_in_formation_of_a_cellular_spore	8.48	11	5	2.33×10^{-4}	4.79×10^{-20}
ribonuclease_activity	3.85	18	18	3.68×10^{-3}	1.02×10^{-3}
lysozyme_activity	5.50	10	7	3.64×10^{-2}	1.00
transition_metal_ion_binding	2.65	22	32	9.67×10^{-2}	1.00

Glossary

CDD Conserved Domain Database. This database is a collection of PSSM models including their own NCBI-curated domains and other domain databases like Pfam, COG, and TIGRFAMS.[193] 4, 101, 103, 106, 108

CDSs Coding sequences. A coding sequence is a DNA or RNA sequence that codes for a protein (including the stop codon). 6, 7, 31, 46, 50, 52–57, 61, 63, 68, 73, 102–104, 106–108, 113, 116, 127–129, 132, 183

COG Clusters of Orthologous Groups. This is a database with functional categories applied to "phylogenetically-classified" protein groups based on all-by-all sequence comparison.[306] 20, 23, 25, 28, 30, 52

DUF Domain of unknown function. A domain family from the Pfam database[75] that has no or limited functional information assigned to its members. 6, 8, 10, 14, 15, 70–72, 74–79, 81, 85–88, 90, 91, 93–97, 99, 122, 129–131

GO Gene Ontology. An ontology for gene product functions that is divided into molecular function, biological process, and cellular location classes.[19] GO terms are mapped onto many databases (e.g. Pfam, InterPro, and the PDB). 2, 52, 69, 74, 75, 81, 82, 89–91, 104, 105, 112–115, 117, 118, 130, 183

HMM Hidden Markov model. This model is a way to incorporate different types of information (e.g. conservation of residues, insertions and deletions) in order to probabilistically label or score a sequence. One use in bioinformatics is to represent a protein family, to sensitively find protein sequences that should match. 4, 9, 55, 101–103, 112

KEGG Kyoto Encyclopedia of Genes and Genomes. Established in 1995 this database contains many different facets and organizations of function, including: pathway maps,

functional orthologs, biochemical reactions, and disease states.[132] 20, 22, 25, 28–31, 33, 36, 40–42, 44, 45, 48, 53, 55–58, 60–65, 127, 128, 178

MAG Metagenome-assembled genome. This is an assembly from binned metagenomes that is ideally made up of one taxa. Bins are created by comparing the differential coverage (across multiple samples) and sequence composition of the contigs.[6] 17, 36, 38, 43, 44, 48, 49, 127

MSA Multiple sequence alignment. An alignment of more than two sequences with which conservation of residues can be analyzed. A MSA enables the study of evolutionary, structural, and functional relationships. 3, 103, 104, 110, 111

necrobiome Microbial decomposition community. This term is from nekrós (the Greek word for dead body) and includes microorganisms that may have come from the surrounding environment or the host. 18, 32, 33, 36–41, 44, 48–50, 126, 127, 172–174

ORF Open reading frame. An open reading frame extends from the beginning of the start codon to just before the stop codon in a DNA sequence, namely everything that is ultimately translated into a protein. 68, 100–104, 106, 109, 111

ORFan A coding sequence without detectable homologs in current databases, named for a condensed form of "orphan ORFs". First defined in 1999 by Daniel Fischer and David Eisenberg.[76] 8, 15, 71, 100–123, 125, 132, 133, 183

PDB Protein Data Bank. Established in 1971, the PDB is a repository for 3D structures of protein, DNA, and RNA.[18] 75, 95, 103–105, 111–121, 183

PSSM Position-specific scoring matrix. The matrix consists of amino acid substitution scores that change based on sequence position, taking into consideration conservation of residues. The scoring matrix is created from a conservation profile of a multiple sequence alignment. 4, 101

resistome Collection of antibiotic resistance genes, specifically called the antibiotic resistome. 18, 26, 31, 50, 127